A novel method for probabilistic coverage estimation of sensor networks based on 3D vector representation in complex urban environments

A. Afghantoloee[1], F. Karimipour[2], M. A. Mostafavi[1]

[1]Center for Research in Geomatics, Department of Geomatics, Université Laval, Quebec, Canada
Email: ali.afghantoloee.1@ulaval.ca;mir-abolfazl.mostafavi@scg.ulaval.ca

[2] Dept. of Geomatic Engineering, University of Tehran, Amir-abad Street, Tehran, Iran
Email: fkarimipr@ut.ac.ir

Abstract

Wireless Sensor Networks (WSNs) are widely used for monitoring and observation of dynamic phenomena. A sensor in WSNs covers only a limited region, depending on its sensing and communicating ranges, as well as the environment configuration. For efficient deployment of sensors in a WSN the coverage estimation is a critical issue. Probabilistic methods are among the most accurate models proposed for sensor coverage estimation. However, most of these methods are based on raster representation of the environment for coverage estimation which limits their quality. In this paper, we propose a probabilistic method for estimation of the coverage of a sensor network based on 3D vector representation of the environment.

1. Introduction

Nowadays, WSNs have found various applications in industry, security, agriculture, military and disaster management. Efficient monitoring and management of dynamic phenomena in the real world necessitates its efficient and accurate coverage. The efficiency of the coverage of a sensor network depends on optimal position of each sensor node within the network. An individual sensor covers only a limited area, which depends on its sensing capacity, range of communication as well as the environment's complexity. The total area covered by a WSN is obtained from the union of the regions covered by individual sensors. Therefore, efficient deployment of sensors in a WSN is a critical issue that affects the coverage as well as communication between sensors.

Several optimization methods (i.e., global or local, deterministic or stochastic, etc.) have been proposed to detect and eliminate coverage holes and hence increase the coverage of sensor networks (Argany et al. 2015). One of the key issues of all deployment optimization algorithms is accurate estimation of the coverage of an individual sensor. Sensing model of individual sensors ─which could be binary or probabilistic, omnidirectional or directional─ has significant impact on the precise coverage estimation of a sensor network using diverse optimization algorithms. Most of the sensor coverage estimation methods use a raster representation of the environment (Akbarzadeh et al. 2013) for optimization purposes that limits their precision and efficiency. This is because raster representations are constrained by their spatial resolution, and their regular shapes result in redundant data for unoccupied areas. Few vector-based optimization algorithms are proposed in the literature, which are mostly based on 2D vector-based representation of the environment and do not adequately consider the presence of manmade and natural obstacles in the sensing areas (Wang and Cao 2011).

To overcome these limitations, in this paper we propose a probabilistic sensor coverage estimation method based on precise 3D vector-base representation of the environment and we present some results of an ongoing research project that aims at better optimization of a sensor network in a 3D complex urban area.

2. Probabilistic sensing models

Akbarzadeh et al. (2013) presented an improvement to optimization models by proposing a probabilistic sensing model for individual sensors that considers not only the impact of distance on sensing capacity, but also the impact of angle between the sensor direction and the line connecting the sensor to a given target (Figure 1). However, this model is still limited in accurately optimizing and estimating the coverage of a sensor network as it is based on a raster representation of the environment.

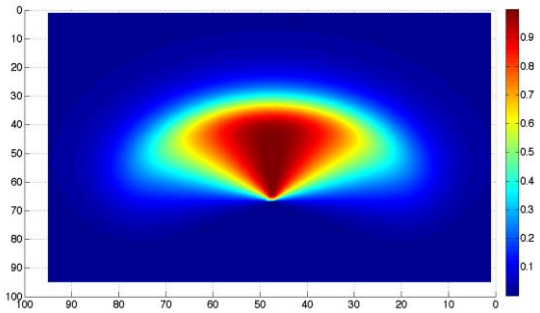


Figure 1: Probabilistic sensing model of a sensor with limited distance and angle range (Akbarzadeh 2013)

Advances in geospatial methods and technologies provide precise and timely collection of 3D spatial data allowing the creation of multi-resolution and multi-purpose 3D vector model of the environment that can significantly improve the efficiency and accuracy of optimization and coverage estimation of a sensor network in a 3D urban environment.

3. 3D vector-based Probabilistic sensor coverage estimation

Consider a representation of the environment as a set of polygons that form buildings, terrain and obstacles, and a sensor that has a limited range in distance and field of view. Then, in a 3D probabilistic sensor coverage model, quality of detection of a polygon depends on its distance to the sensor as well as the angle between the polygon and the sensor. In Figure 2, for example, polygon **A** is further from sensor **S** compared to polygon **B**, thus polygon **A** is detected with a lower quality than polygon **B**. On the other hand, although polygons **B** and **C** are, in average, located at the same distance to the sensor **S**, polygon **C** has a lower detection quality than polygon **B** as it has a more oblique direction respect to the sensor **S**. Even, as the distance and direction differ from point to point on an individual polygon, each point on a polygon may have a different detection quality.

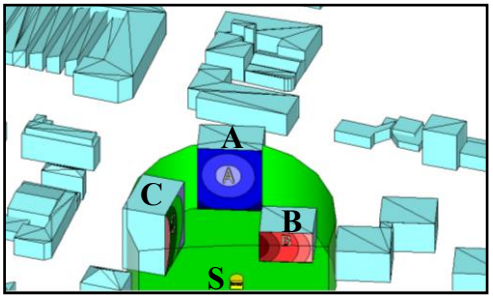


Figure 2: Detection quality of target based on distance and direction.

To practically implement the proposed methodology, the polygons, or the fraction of polygons, that are visible by the sensor are determined (Afghantoloee *et al.* 2014). Then, these visible polygons are discretized to a grid (Figure 3) and the detection quality of cells is estimated and finally categorized to certain classes.

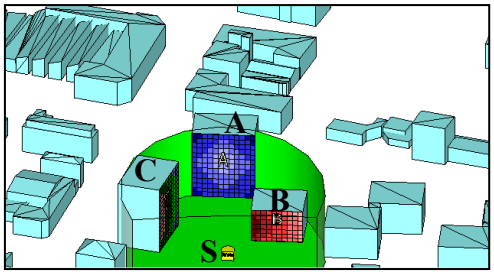


Figure 3: Surface rasterizing in 3D space.

The detection quality of a cell is determined using the area of each cell multiplied by the probability of coverage respect to its distance and direction from the sensor. This is obtained using the following equations (Akbarzadeh et al. 2013):

$$C = Area(q) * P(\|q - p_s\|) * P(\theta - \angle(q, p_s)) \quad (1)$$

$$P(\|q - p_s\|) = \begin{cases} \dfrac{1}{1 + \exp\left[-\dfrac{\alpha}{\|q - p_s\|} - \beta\right]} & \|q - p_s\| \le d_s \\ 0 & otherwise \end{cases} \quad (2)$$

$$P(\|q - p_s\|) = \begin{cases} \left(\dfrac{\cos(\theta - \angle(q, p_s)) + 1}{2}\right)^{\omega} & (\theta - \angle(q, p_s)) \in [-\alpha_s, \alpha_s] \\ 0 & otherwise \end{cases} \quad (3)$$

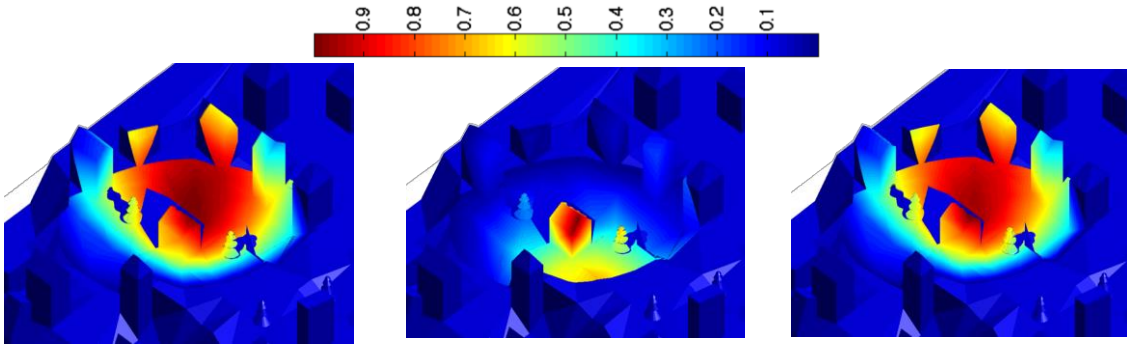
where $Area(q)$ is the area of the pixel q; d_s *and* α_s are respectively the distance and angle range of the sensor; $\angle(q,p_s)$ is the pan angle of the sensor relative to the pixel q; $\|q\text{-}p_s\|$ is the distance between the sensor and pixel q; θ is pan angle of the sensor; and α, β, and ω are the parameters for configuring the probability function which can be estimated from the observation behavior of the sensor.

4. Case study

In order to evaluate the proposed strategy, a directional sensor with a 3 and 30 meters height and radius, pan angle of 45° and direction of [1,1,-1] was considered in 3D environment. The parameters α, β, and ω are respectively considered as 350, 10, and 3. All features in the 3D vector model are constructed by polygons with 1 cm accuracy, which have a counterclockwise structure. Using perspective based visibility estimation methodology and 3D rasterizing of the polygons, the probabilistic coverage of all cells of the 3D model was calculated. Figure 4 illustrates the probabilistic coverage considering (a) distance, (b) direction and (c) both together. Figure 5 shows the results for the same case in a 2.5 raster dimension model with 20 cm resolution. Comparing the results (Table 1) indicates that the proposed strategy allows a more precise probabilistic coverage estimation compared to a raster model.

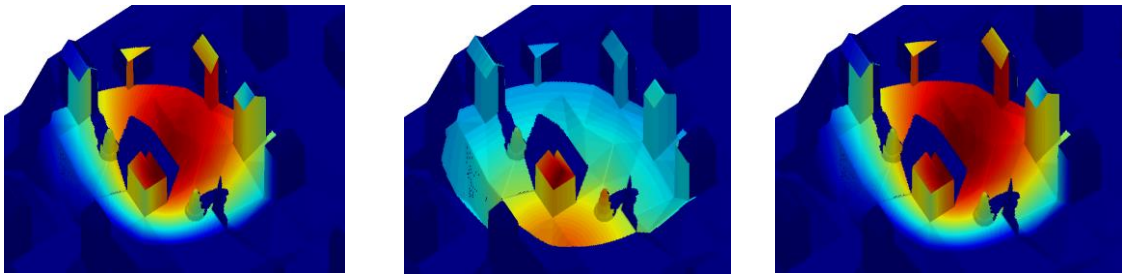
4. Conclusions

This paper proposed a novel method based on 3D urban vector models for estimation of the coverage of wireless sensor networks. This method has more advantages compared to raster-based DSM/DEM models as it considers the coverage of all facets of features like buildings and walls, and even under the features such as bridges and balconies. The proposed method for a probabilistic sensing estimation model led to a more realistic coverage estimation for individual sensors in a sensor network in a 3D complex environment.



(a) (b) (c)

Figure 4: Probabilistic coverage estimation in a 3D vector model: (a) direction (b) distance, and (c) both together.



(a) (b) (c)

Figure 5: Probabilistic coverage estimation in a raster model: (a) direction (b) distance, and (c) both together.

Table 1. The comparison coverage probabilistic estimation in Raster and Vector model.

Coverage probability	Raster_20cm (m^2)	Vector_without wall (m^2)	Vector_with wall (m^2)
Distance	438.66	446.5768	691.6791
Direction	1329.570	1349.364	2073.376
Distance & Direction	364.928	370.363	568.989

References

Afghantoloee A, Doodman S, Karimipour F and Mostafavi MA, 2014. Coverage Estimation of Geosensor in 3d Vector Environments, *GIResearch 2014*.

Akbarzadeh V., Gagne C., Parizeau M, Argany M and Mostafavi MA, 2013, Probabilistic sensing model for sensor placement optimization based on line-of-sight coverage. *IEEE Transactions on Instrumentation and Measurement*. 62:293-303.

Argany M, Mostafavi MA, Gagné C, 2015. Context-Aware Local Optimization of Sensor Network Deployment. *Journal of Sensor and Actuator Networks (JSAN)*. 4(3): 160-188

Wang, Y., Cao, G., 2006. Movement-assisted sensor deployment, IEEE Infocom (INFOCOM‘04), pp. 640-652.

Multi-frequency segmentation of movement trajectories

Sean C. Ahearn[1]

[1]*Center for Advanced Research of Spatial Information (CARSI), Hunter College - CUNY* (sahearn@hunter.cuny.edu)

Abstract

This paper presents a new multi-frequency segmentation method for movement trajectories. The method is presented using a case study of a Turkey Vulture data set. The approach show promising results to automatically extract behavioral modes from movement trajectories at multiple temporal frequencies.

1. Introduction

Computational Movement Analysis focuses on the characterization of the trajectory of individuals across space and time. The end goal is often to understand the behavioral state of the animal by analyzing this complex signal. A level of complexity that is often ignored is the fact that different behaviors occur at different spatial and temporal scales (Ahearn et al. 2001; Ahearn and Smith, 2006; de Weerd et al. 2015; Soleymani et al., 2014). This research proposes a multi-frequency analysis technique that decomposes a movement parameter (i.e. speed) time series, derived from movement trajectories, into a range of spatiotemporal frequencies and then re-combines them into the *Multi-frequency Laplacian Series (MFLS)*.

2. Methods

Long-term GPS data of a Turkey Vulture obtained from the Movebank Data Repository was used for our analysis. The data set includes 32,000 fixes with a temporal resolution of 1 hour, tracked over 3.5 years. The data was manually classified by domain experts into four different behavior modes: *breeding grounds, fall migration, non-breeding grounds,* and *spring migration* (Dodge et al. 2014). This resulted in 18 distinct segments. For this analysis, the trajectory was converted to a *time series* of the speed at each GPS fix as calculated from two sequential fixes.

The approach used to create the multi-frequency decomposition of a trajectory is after Burt (1983). The method generates a new data series, from the original time series, that is effectively, what would result from the convolution of two different sized Gaussian filters (DoG). To generate the series, a number of steps are necessary. First, a Gaussian shaped filter

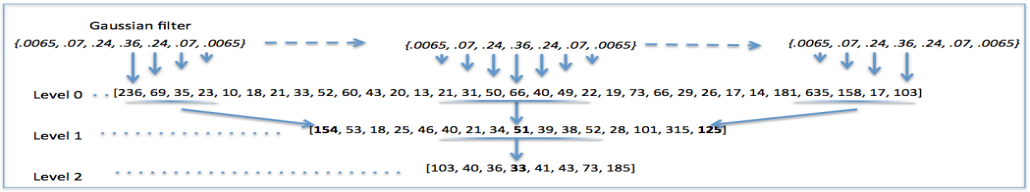


Figure 1: Gaussian Pyramid

. is convolved with the series by centering it at every other value in the series (Figure 1, Level 0). The result of each convolution is the next value in Level ***x+1***. Recursively applying the method

results in ***n*** levels; with each level being ½ the size of the previous Level ***x***. Figure 1 shows 3 levels of the Gaussian series (original plus 2 levels) created using this approach.

The next step is to expand each of the levels of the Gaussian series, obtained from the previous step, to the original size of the series at Level 0, with the same kernel used for reduction. This process is also carried out with a recursive algorithm. Figure 2a is an example of 6 expansion levels plus the original (Level 0, in red), starting with expansion level 6 and going to expansion level 11 (shown in blue).

The third step is to produce the Difference of Gaussian Filters (DoG). This is accomplished by subtracting each expanded level from the previous level (i.e. expanded Level 1 – Level 0, Level 2 – Level 1, … Level n – Level (n-1)). This results in the Laplacian series (i.e. DoG) shown in green in Figure 2a. Where the Laplacian series (a second derivative function) crosses zero, it represents an inflection point called a ***zero-crossing***.

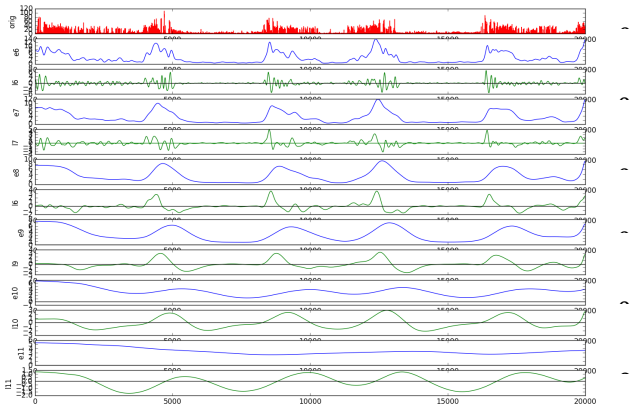


Figure 2a: Original data (red), five level expanded series (i.e. Exp 6-11, in blue) and Laplacian series (Lap 6-11, in green).

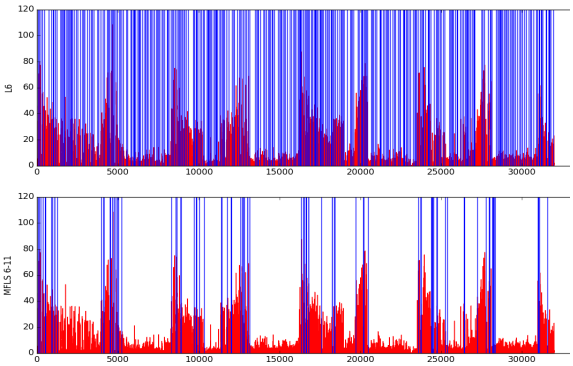


Figure 2b: Laplacian series 6 (top) and Multifrequency Laplacian series 6-11 (bottom) overlayed on original time series.

How to combine this range of frequencies into the *Multi-frequency Laplacian Series (MFLS)* is after Ahearn (1988). The methodology uses a procedure that captures broader scale trends using the lower spatial frequencies, while preserving the higher frequencies that define the beginning and ending of these trends (or behaviors). Combining the levels is accomplished by taking the maximum of the absolute value of the Laplacian value among the different levels being combined (Ahearn, 1988). An additional step taken in this research is to use the penultimate frequency band in the Laplacian series to define the number and approximate location of the transition points (i.e. zero-crossing indexes) between behaviors, and the indexes derived from the MFLS to refine the locations of the transition points. That is, the algorithm finds the closest index in the MFLS to each of the indexes in Level (*max*-1), where *max* is 11 in this case, and "substitutes" it in its place to create what we call the *MFLS nearest*. Level (*max*-1) was chosen because the analysis showed that one level less than the maximum level often corresponds to the temporal frequency of the longest temporal feature, in this case the seasonal behavioral patterns of the turkey vulture.

3. Results

Creation of the MFLS requires the selection of the starting level in the frequency range to be used in the process. Selection of the ending level (i.e. the lowest frequency band) is done automatically at the level (in this case Level 11) where the model reaches a minimum number of zero-crossings and remains unchanged with the addition of the next level. For the dataset used in this study, Level 6 (L6) is selected for the beginning frequency band. Once determined, the algorithm combines levels 6 through 11 of the Laplacian series.

Figure 2b above, shows a comparison of segments extracted at the zero-crossings (represented as vertical lines) for level 6 (L6 on top) and MFLS6-11 (bottom). Note the lack of definition of the segments related to the seasonal behavior (e.g. migration) of the Turkey Vulture in L6 (top of figure 2b) and the clear definition of separate segments in the MFLS 6-11 (bottom of figure 2b).

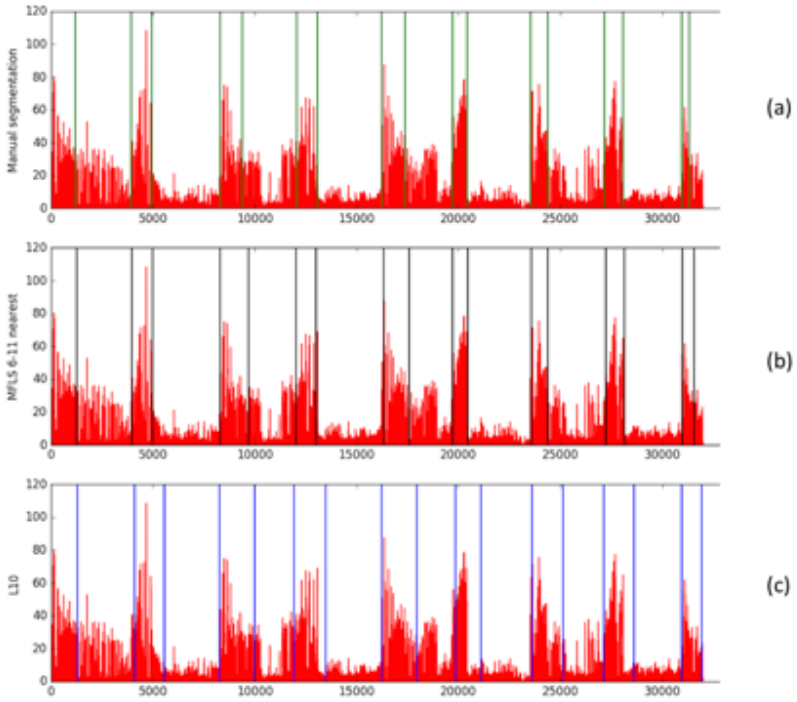


Figure 3: Manual segmentation (a); MFLS 6-11 nearest (b); L10 (c).

Comparison of the results with the analysis conducted by Dodge et al. (2014) was done to get a better understanding of the relationship between the multi-frequency segmentation and a biologic interpretation of the trajectory (Figure 3). In order to do this, the manually segmented trajectory (Figure 3a) is compared with the *MFLS 6-11 nearest* (Figure 3b) and Level L10 of the Laplacian Pyramid (Figure 3c). The manual classification by domain experts resulted in four different behavior modes: *breeding grounds, fall migration, non-breeding grounds,* and *spring migration* (Dodge et al. 2014) and 18 distinct segments. The L10 frequency band found the same number of segments, 18 as the manual classification. The *MFLS6-11 nearest,* which uses the global scale frequencies (i.e. L10) for defining segments therefor had 18 segments. The difference was

in the timing of the transitions between behavioral modes. The average difference between the transition times of the manual classification and the *MFLS6-11 nearest* was 76 hours. The average difference between the transition times of the manual classification and Level L10 was 326 hours. Given that the time series occurs over a 3.5-year period the precision of the definition of the temporal transition points is quite high (i.e. within 3 days) for the *MFLS6-11 nearest.* It is less so for the L10 due to the smoothing that occurs at this low frequency.

4. Conclusion

The results of this analysis show promise for analyzing a time series at multiple temporal scales. The strength of this methodology is that it captures the low frequency phenomena, in this case the different behavioral modes of the Turkey Vulture, while preserving the high frequency transitions from one behavior to the other. Using just a single frequency level for segmentation will not yield segments that are closely aligned (temporally) with the manually defined segments, thus the need for the MFLS. Additionally, with the exception of the selection of the first high frequency band in the MFLS, the process is totally automated and requires no training, parameterization or thresh-holding.

References

Ahearn, S. C., Smith, J. L. D., Joshi, A. R., & Ding, J. (2001). TIGMOD: an individual-based spatially explicit model for simulating tiger/human interaction in multiple use forests. *Ecological Modelling*, *140*(1), 81-97.

Ahearn, S. C., and J. L. D. Smith. "Modeling the interaction between humans and animals in multiple-use forests: a case study of Panthera tigris." *GIS, Spatial Analysis, Modeling. ESRI Press, Redlands, California, USA* (2005): 358-387.

Ahearn, Sean C. "Combining Laplacian images of different spatial frequencies (scales): implications for remote sensing image analysis." *Geoscience and Remote Sensing, IEEE Transactions on* 26.6 (1988): 826-831.

Burt, Peter J., and Edward H. Adelson. "The Laplacian pyramid as a compact image code." *Communications, IEEE Transactions on* 31.4 (1983): 532-540.

Dodge, Somayeh, Gil Bohrer, Keith Bildstein, Sarah C. Davidson, Rolf Weinzierl, Marc J. Bechard, David Barber et al. "Environmental drivers of variability in the movement ecology of turkey vultures (Cathartes aura) in North and South America." *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 369, no. 1643 (2014): 20130195.

de Weerd, Nelleke, Frank van Langevelde, Herman van Oeveren, Bart a. Nolet, Andrea Kölzsch, Herbert H. T. Prins, and W. Fred de Boer. 2015. “Deriving Animal Behaviour from High-Frequency GPS: Tracking Cows in Open and Forested Habitat.” *Plos One* 10 (6): e0129030. doi:10.1371/journal.pone.0129030.

Soleymani, Ali, Jonathan Cachat, Kyle Robinson, Somayeh Dodge, Allan V Kalueff, and Robert Weibel. 2014. “Integrating Cross-Scale Analysis in the Spatial and Temporal Domains for Classification of Behavioral Movement.” *Journal of Spatial Information Science* 8 (8): 1–25. doi:10.5311/JOSIS.2014.8.

The Life Cycle of Volunteered Geographic Information (VGI) Contributors: the OpenStreetMap Example

D. Bégin[1], R. Devillers[1,2], S. Roche[2]

[1] Department of Geography, Memorial University, St. John's (NL), Canada
Email:{d.begin; rdeville}@mun.ca

[2] Centre de recherche en géomatique, Université Laval, Québec (QC), Canada
Email: stephane.roche@scg.ulaval.ca

1. Introduction

The Web 2.0 changed the way Internet users interact with knowledge (Gore 1998; Goodchild 2007) by allowing knowledge sharing through various online systems (e.g. Wikipedia). In GIScience, Volunteered Geographic Information (VGI) has attracted the attention of scholars due to its ability to crowdsource geographic information potentially useful in many contexts (Haklay 2014; Arsanjani *et al.* 2015).

Classifications of VGI contributors have been proposed, based on users' motivation (Coleman *et al*. 2009) or on the volume of their contributions (Panciera *et al.* 2010; Neis and Zipf 2012). Existing studies show that the nature of the contributions broadens with the time spent in a project (Kim 2000; Panciera *et al.* 2009) but none clearly linked them to the timespans of the different stages in the life cycle of contributors. This paper presents the first detailed analysis of the time over which contributors participate to a VGI project by using OpenStreetMap (OSM) data, identifying sets of contributors that share similar temporal patterns of contributions, and discussing the potential impacts on contributions.

2. Contributors' Timespan Distribution

While OSM data can be accessed by anyone, only registered users can edit the database. Once registered, no mechanism identifies users that stop contributing to the project. We define a 'registered user' as someone that created an OSM account, while a 'contributor' is a registered user that started at least one editing session (i.e. a changeset). 'Contributors' timespan' refers to the timespan between a contributor's first and last edit.

All the transactions made in OSM until September 1, 2014, were extracted and loaded into a PostgreSQL 9.3 database. Statistical analyses and visualizations were performed using the R 3.2.1 software.

A first analysis compared cumulative OSM registered users with actual contributors, creating daily Contributors/Registered Users ratios (Figure 1). Ratios reveal wide variations over time, ranging from 6% to 47%, for an average of 30.9%. Results support Neis and Zipf (2012) findings that only a third of registered users eventually become contributors.

A complementary cumulative distribution function (CCDF) of contributors' timespan was also generated (Figure 2). It represents the proportion of contributors who edited the database for a similar period of time or longer. Five pivotal points were identified based on this figure and on additional analyses.

A first pivotal point is found at about one hour of contributions, where 15% of participants stopped contributing in a matter of seconds. This abrupt break in the curve represents new contributors that made only a few edits, or even none, before the OSM API automatically closes their one and only editing session left idle for an hour.

The proportion of OSM users who contributed data keeps decreasing rapidly for about an hour then it slows down until it reaches our second pivotal point after 24 hours (one day). Analyses show that 60% of contributors did not edit data beyond this point, a proportion

similar to what was found in Wikipedia (Panciera *et al.* 2009).

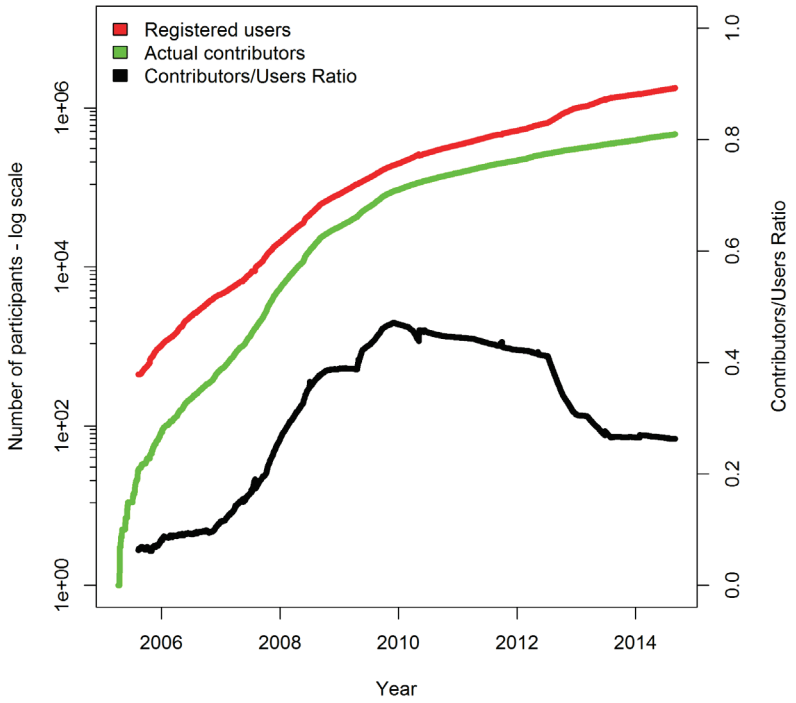


Figure 1: Participation in the OSM project over time.

The proportion of contributors then follows a regular (log based) decline for about a year after which it slows down significantly. Fewer than 20% of participants contributed beyond a first year, our third pivotal point. The fourth point is about five years later and concerns less than 1% of participants.

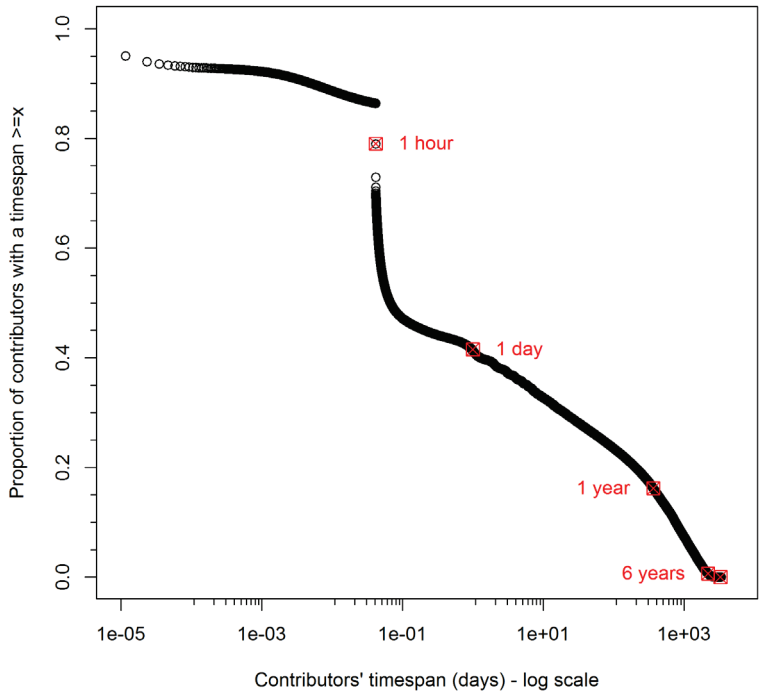


Figure 2: Complementary cumulative distribution function (CCDF) of contributors' timespan showing the proportion of contributors by timespan (in days).

The fifth and last point is located at nine years, a time at which only a few dozen accounts were active. Many of OSM developers and innovators (Beal and Bohlen 1957) have then kept editing the database over the years.

3. Evolution of Edits Over the Life Cycle of Contributors

The contributions made at each stage of the life cycle are described in Table 1. The table shows the proportion of participants who reached each stage among the members of their cohort ('Users'), the volume of edits made for the 10^{th}, 50^{th} and 90^{th} percentiles (P*), the average number of days actually involved in contributing ('Days'), and the average number of days spent between contributions ('Idle').

Table 1: Life cycle stages of OSM registered users and contribution metrics.

Name	Starts after	Lasts for	Users (%)	P10	P50	P90	Days	Idle
Lurkers	NA	NA	69.1	NA	NA	NA	NA	NA
Contributors	*See breakdown below*		30.9	0	16	1.7e3	11	17
Contributors' stages	…	…	…	…	…	…	…	…
Visitors	0 second	1 hour	21.0	0	2	42	1	NA
Novices	1 hour	23 hours	37.8	0	6	93	1	NA
Amateurs	1 day	364 days	25.1	4	107	2.6e3	6	13
Adherents	1 year	5 years	20.0	15	716	3.7e4	46	21
Veterans	6 years	3 years	13.2	373	2.3e4	4.4e5	253	8
Elders	9 years	NA	33.3	3.8e3	4.6e4	3.8e5	320	9

The *lurkers* represent the registered users that have not contributed to OSM yet (69.1%). The remaining 30.9% of registered users that have at least started an editing session are *contributors*. Contributors' life cycle stages are described above.

Visitors and *Novices* contribute little to OSM, with a median contribution (P50) around five edits. Further analyses of the data indicate that the few most productive visitors and novices accounts were dedicated to large data imports, with tens of thousands of edits done within this short period of time.

Amateurs and *Adherents* have similar contribution profiles. The volume of edits from the lower half (P50) of contributors increases proportionally to the time they spent in the project but this proportion doubles for the most active ones (P90).

Veterans and *Elders* show different contribution profiles, with larger and more evenly distributed contributions, a behaviors also observed in Wikipedia (Geiger and Halfaker 2013). Furthermore, these participants contribute on average twice more often than those from all other stages, with an average of one day out of nine (idle).

Overall, the proportion of participants who keep contributing within a cohort is slightly higher than what was found within Wikipedia (Zhang et al. 2012).

4. Conclusion

This paper reported on an analysis of the time over which OSM users contribute to the project. Pivotal points found on the cumulative complementary distribution function show six different phases in the life cycle of contributors.

By assessing each phase, we found that contributions are more frequent, larger, and evenly distributed against participants that spent more than five years in the project. We also found that for each phase, about 20% of participants keep contributing and move to the next phase. Since these core contributors produce most of the data over a long period of time, we

expect that the data quality either improves or becomes increasingly uniform over the years.

Further analysis should assess if core contributors can be identified from their early behaviors in order to keep them engaged by designing crowdsourcing activities adequately.

Acknowledgements

The authors would like to acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) for their financial support through the Discovery Grant of the second author.

References

Arsanjani, J.J., Zipf, A., Mooney, P., Helbich, M., 2015. OpenStreetMap in GIScience, In: Lecture Notes in Geoinformation and Cartography, Cartwright, W., Gartner, G., Meng, L. and Peterson, M.P. (Eds.), Springer, Switzerland, pp. 324.

Beal, G.M., Bohlen, J.M., 1957. The diffusion process, Iowa State College, Ames (USA).

Coleman, D.J., Georgiadou, Y., Labonté, J., 2009. Volunteered geographic information: The nature and motivation of producers. International Journal of Spatial Data Infrastructures Research, 4, pp. 332-358.

Geiger, R.S., Halfaker, A., 2013. Using edit sessions to measure participation in Wikipedia, In: Proceedings of the 2013 conference on Computer supported cooperative work, pp. 861-870.

Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4), pp. 211-221.

Gore, A., 1998. The Digital Earth: Understanding our planet in the 21st Century. Australian surveyor, 43(2), pp. 89-91.

Haklay, M., 2014. OpenStreetMap studies (and why VGI not equal OSM). Po Ve Sham – Muki Haklay's personal blog (August 14), pp.1-2.

Kim, A.J., 2000. Community building on the web: Secret strategies for successful online communities, Addison-Wesley Longman Publishing,.

Neis, P., Zipf, A., 2012. Analyzing the Contributor Activity of a Volunteered Geographic Information Project—The Case of OpenStreetMap. ISPRS International Journal of Geo-Information, 1(2), pp. 146-165.

Panciera, K., Halfaker, A., Terveen, L., 2009. Wikipedians are born, not made: a study of power editors on Wikipedia, In: Proceedings of the ACM 2009 international conference on Supporting group work, pp. 51-60.

Panciera, K., Priedhorsky, R., Erickson, T., Terveen, L., 2010. Lurking? cyclopaths?: a quantitative lifecycle analysis of user behavior in a geowiki, In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1917-1926.

Zhang, D., Prior, K., Levene, M., 2012. How long do Wikipedia editors keep active? In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, pp. 4.

Spatiotemporal enabled Content-based Image Retrieval

Mariana Belgiu[1], Martin Sudmanns[1], Dirk Tiede[1], Andrea Baraldi[1,2], Stefan Lang[1]

[1]Department of Geoinformatics – Z_GIS, University of Salzburg, Schillerstrasse 30, 5020 Salzburg, Austria
Email: {mariana.belgiu; martin.sudmanns;dirk.tiede;stefan.lang}@sbg.ac.at

[2] Department of Agricultural and Food Sciences, University of Naples Federico II, Italy
Email: andrea6311@gmail.com

Abstract

Remote sensing has emerged as a powerful tool in a number of applications, but many challenges such as the development of "smart' (knowledgeable), effective and efficient Earth observation (EO) image-content extraction and content-based image retrieval systems in response to big sensory data acquisitions by ever-increasing spaceborne EO imaging sensors remain unsolved. In this paper, we discuss the need to explicitly specify a priori 4D spatiotemporal scene domain knowledge to be mapped onto the image domain in terms of 2D image features and spatial constraints. This 4D-to-2D mapping capability implies the solution of the vision problems, where the semantic gap from sensory data to high-level information products must be filled in.

1. Introduction

In the last decades, the quantity and quality of Earth Observation (EO) images have increased tremendously. There is therefore an increasing demand for innovative and intelligent solutions, suitable for coping with the dual problems of (automatic) EO image-content extraction and content-based EO image retrieval. Two main image retrieval approaches are available including annotation-based image retrieval (query by text) and content-based image retrieval (CBIR). The first approach is limited to querying image archives using keywords or metadata facets (Quartulli and Olaizola 2013) such as geographic coverage, time of acquisition or sensor parameters. The second approach makes use of visual context including color, texture or shape to index and retrieve images from archives. CBIR is challenged by the semantic gap (Smeulders et al. 2000) between low-level features extracted from satellite optical images (coded as a finite, rectangular grid of digital numbers as multi-spectral reflectance values) and the high-level semantic concepts used by users to describe a 4D real-world scene in user-speak.

Despite important advances in the domain of CBIR (Datcu et al. 2003; Durbha and King 2005), an operational spatiotemporal semantic querying systems for EO image time-series analysis and retrieval remains a visionary goal of the remote sensing community (Wang et al. 2013). This paper expands upon the successful implementation of a CBIR called ImageQuerying system and discusses the need to explicitly specify a priori 4D spatiotemporal scene domain knowledge to be mapped onto the image domain in terms of 2D image features and spatial constraints.

2. ImageQuerying System

ImageQuerying system is a web-based CBIR system where image-derived scene classification layers (thematic maps) are used to index the original EO images, allowing users to perform semantics queries on cross-sensor and time-series images. Single date multispectral images captured by different sensors (e.g. Landsat sensors, Sentinel-2) are stored in an array database system (here: Rasdaman Community Edition - http://www.rasdaman.org/) together with various pre-processed thematic layers (spectral indexes, spectral categories based on physical models (Baraldi et al. 2006), geometry). The

system facilitates users to formulate queries using a user-friendly web-based graphical inference engine.

3. Developing spatiotemporal semantic content-based image querying system

Our goal is to extent the ImageQuerying system with 4D spatiotemporal scene-domain knowledge and to map this knowledge onto a 2D sub-symbolic image domain through a sensor model. The ImageQuerying system shall facilitate the implementation of the following spatiotemporal queries (Yuan, 1999): (1) Single-date query: e.g. retrieve lakes depicted in an EO image within target area x at acquisition time t; (2) Multi-date spatiotemporal queries: e.g. retrieve all images where vegetation growth occurs at location x over a target time interval [t0-tn] (TTI); (3) spatiotemporal behavior queries required to trace the thematic changes of objects through time and space: e.g. trace the extent of flooded area at location x and TTI.

3.1 Modular organization of knowledge required to query time-series images

Spatiotemporal modelling of target classes in the 4D scene domain becomes an important asset given the increasing spatial, spectral and temporal resolutions of EO: e.g. Sentinel-2. Therefore, we propose the organization of the knowledge required for EO image retrieval using ImageQuerying system into three knowledge layers (Figure 1): (1) 4D spatiotemporal scene domain knowledge accounting for the semantics of real-world objects of interest; (2) Image domain knowledge accounting for quantitative visual features in an image such as color, texture or geometry; (3) Knowledge about the transfer functions (translation rules) capable of mapping the scene domain knowledge onto the image domain knowledge.

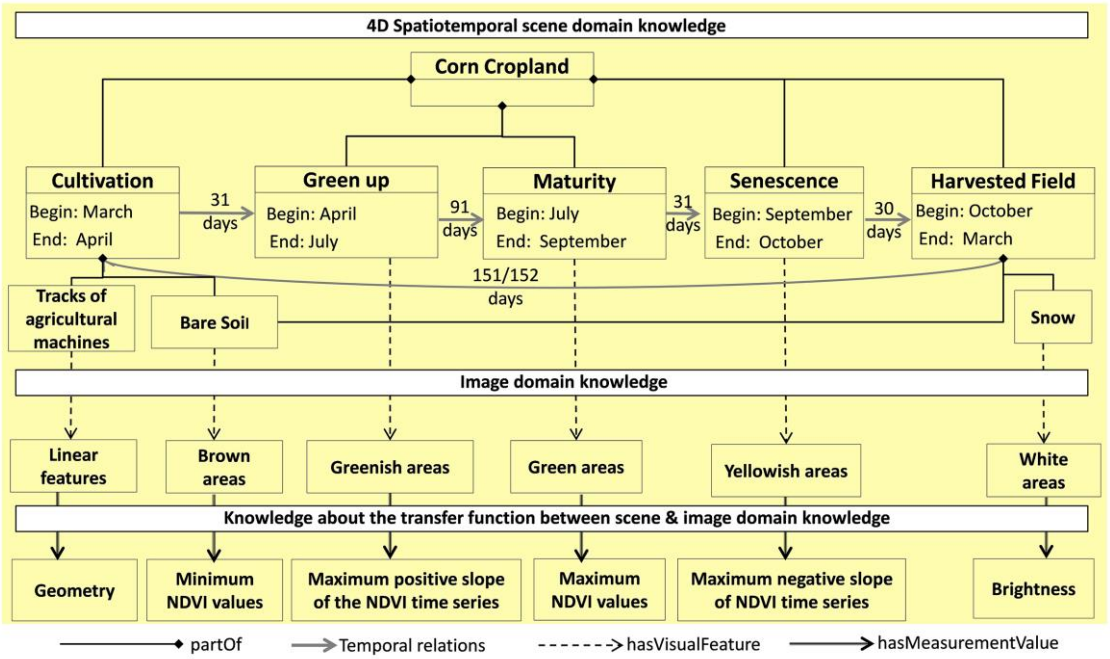


Figure 1. Modelling objects with cyclic behavior – corn cropland example in northern hemisphere.

The following real-world objects and events are modelled in the scene domain knowledge layer: *periodic objects* (objects with cyclic behavior), *persistent objects*, *short-term transition events* and *slowly transient events* (Natali et al. 2011; Nixon and Hornsby, 2010). *Periodic*

objects called also *temporary objects* (Nixon and Hornsby, 2010) consist of a sequence of states subsisting over a certain period of time. Corn cropland for example represents a periodic object with a growth cycle imposed by the agricultural practices specific to a geographic region (see Figure 1). The temporal extents of corn cropland's states impose a certain constrain on the temporal resolution of observations (i.e. EO) are specified by the temporal bounds (begin and end) accounting for cropland growth calendar. The cyclic crop growth states are modelled in the scene domain knowledge layer, whereas the visual characteristics of identified states are specified in the image domain layer: e.g. tracks of tillage and sowing of the agricultural machines appear in the optical domain as linear features. The linear features are measured in the image by using geometry metrics, e.g. compactness, whereas identified growing states are represented using vegetation indices such as Normalized Difference Vegetation Index (NDVI).

Persistent objects are (non-cyclic) discrete objects whose characteristics can change over time, e.g. non-vegetated areas or water bodies (see Figure 2). Both periodic and persistent objects are subject to change caused either by *short-term transition events* or by *slowly transient events*. Vegetated area for example can change due to various policies put into practice by responsible authorities causing slowly transient events such as urban sprawl or due to some short-term transition events such as fire and flooding.

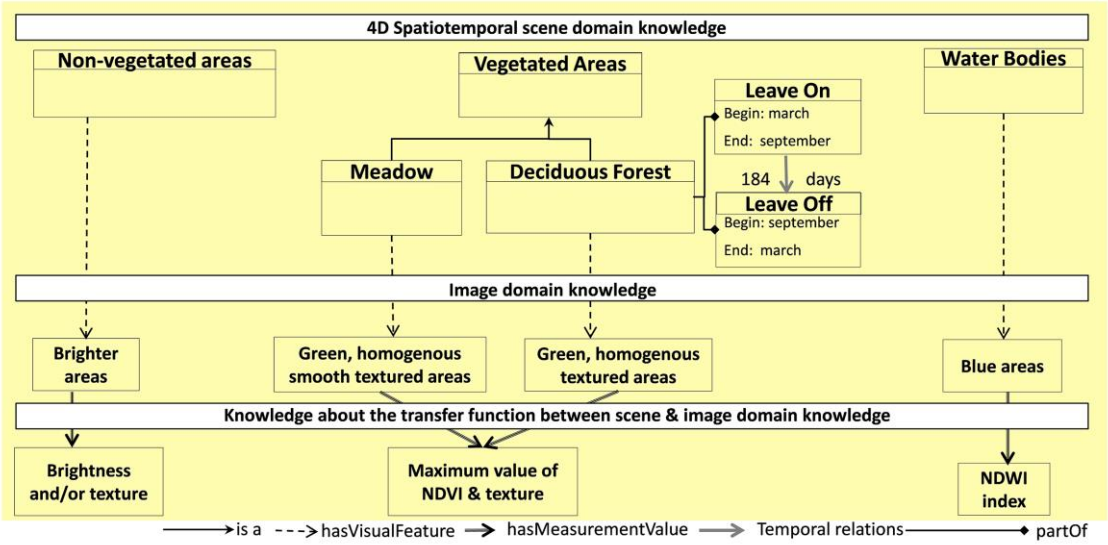


Figure 2. Modelling persistent objects (non-vegetated, water bodies) and periodic objects (decideous forest).

4. Preliminary results

In this section we are focusing on the following image query: retrieve all images where the vegetation growth occurs over specified TTI. The ImageQuerying system facilitates users to re-use queries already stored in the knowledge base (Figure 3: 'Generate map using content based filter: Potential Vegetation Growth'). These queries account for the spatiotemporal scene domain knowledge and the transfer functions required for mapping it onto image domain knowledge: e.g. vegetated areas are areas with high leaf index (1 to 7 classes in our ImageQuerying sytem - Figure 3). Users have to select an area of interest, define a temporal interval and send the query related to the vegetation growth to the system. The system verifies the hypothesis expressed in the query and sends back the discovered images (Figure 3).

5. Conclusion

In this paper, we presented an innovative approach for querying big image data. The strength of this approach is the explicit mapping of spatiotemporal scene domain knowledge onto 2D image domain knowledge such as color, texture, spatial shape or vegetation indices.

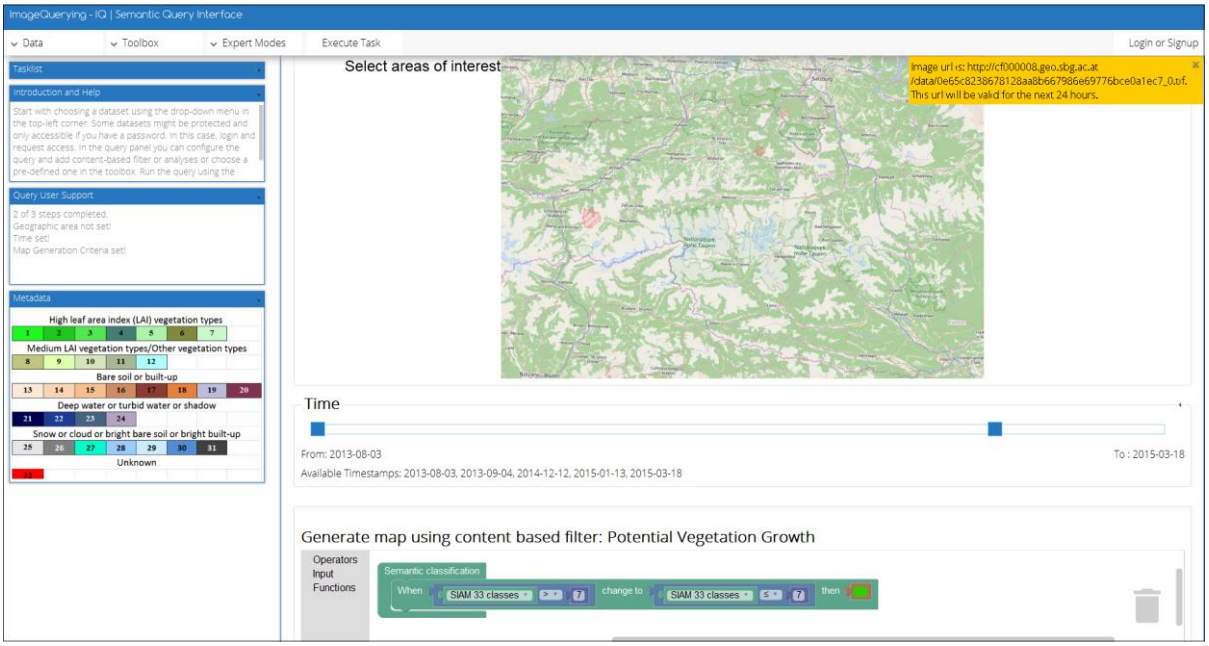


Figure 3. Potential vegetation growth query example with user-defined color coding in the web-based graphical inference engine

Acknowledgements

The research has received funding from Austrian Research Promotion Agency (FFG) under Austrian Space Application Programme (ASAP) within AutoSentinel-2/3 project (contract no: 848009).

References

Baraldi A, Puzzolo V, Blonda P, Bruzzone L, and Tarantino C, 2006, Automatic spectral rule-based preliminary mapping of calibrated Landsat TM and ETM+ images. *Geoscience and Remote Sensing, IEEE Transactions on*, 44 (9): 2563-2586.

Datcu M, Daschiel H, Pelizzari A, Quartulli M, Galoppo A, Colapicchioni A, Pastori M, Seidel K, Marchetti P G, D'Elia, S, 2003, Information mining in remote sensing image archives: system concepts. *Geoscience and Remote Sensing, IEEE Transactions on*, 41:2923-2936.

Durbha S S, King R L, 2005, Semantics-Enabled Framework for Knowledge Discovery From Earth Observation Data Archives. *IEEE Transactions on Geoscience and Remote Sensing*, 43, 2563 - 2572.

Natali S, Beccati A, D'Elia S, Veratelli M, Campalani P, Folegani M, Mantovani S, 2011, Multitemporal data management and exploitation infrastructure, *Proceedings of 6th International workshop on analysis of Multi-temporal Remote Sensing Images (Multi-Temp)*, Trento, Italy, 217-220.

Nixon, V, Hornsby, K S, 2010, Using geolifespans to model dynamic geographic domains. *International Journal of Geographical Information Science*, 24:1289-1308.

Quartulli M, Olaizola I, 2013. A review of EO image information mining. *ISPRS Journal of Photogrammetry and Remote Sensing,* 75:11-28.

Smeulders, W.M.A., Worring, M., Santini, S., Gupta, A., Jain, R., 2000. Content-Based Image Retrieval at the End of the Early Years. IEEE Transanction on Pattern Analysis and machine Intelligence 22, 1349-1380.

Wang M, Wan Q M , Gu L B, Song T Y, 2013, Remote-sensing image retrieval by combining image visual and semantic features. *International Journal of Remote Sensing*, 34:4200-4223.

Yuan M, 1999, Use of a Three-Domain Repesentation to Enhance GIS Support for Complex Spatiotemporal Queries. *Transactions in GIS*, 3:137-159.

Land Use Regression of Particulate Matter in Calgary, Canada

S. Bertazzon[1], F. Underwood[1], M. Johnson[2], J. Zhang[2]

[1]University of Calgary, Department of Geography, 2500 University Dr. NW, Calgary, AB, Canada, T2N 1N4
Email: {bertazzs, feunderw}@ucalgary.ca

[2]Health Canada, Air Health Science Division, 269 Laurier Ave West, Ottawa, ON, Canada, K1A 0K9
Email: markey.johhnson@hc-sc-gc.ca; joyce.zhang@alumni.utoronto.ca

Abstract

Two-week integrated samples of particulate matter ($PM_{1.0}$, $PM_{2.5}$, PM_{10}) were collected in summer and winter in Calgary, Canada. PM concentrations were higher in summer for all size fractions. In both seasons, spatial variation and clustering were moderate. Land use regression (LUR) models were estimated for each PM size fraction and season, yielding $R^2 > 0.75$ for $PM_{2.5}$ and PM_{10} in summer, and $R^2 > 0.45$ for $PM_{1.0}$ in summer and for all winter models. Summer models yielded consistent predictors across size fractions, representing industrial emissions, local traffic, and major arterial traffic. Winter predictors included industrial emissions, major arterial traffic, and distance from open, snow-covered parks. The models suggest industrial pollution covered large areas in both seasons, and was affected by prevailing winds in summer, whereas traffic-related pollution decayed rapidly as distance from roads increased.

1. Introduction

Particulate matter (PM) is a mixture of small particles: acids, organic chemicals, metals, and dust particles (EPA 2016). Coarse particles (PM_{10}) are 2.5−10 micrometers in diameter; fine particles ($PM_{2.5}$) are less than 2.5 micrometers. Particulate pollution is associated with reduced visibility, environmental degradation, and adverse health effects, e.g., respiratory and cardiovascular morbidity and mortality (Rückerl *et al.* 2011), with evidence that health impacts and chemical composition vary by size fraction (Kelly and Fussell 2012). Land use regression (LUR) yields air pollution estimates at fine spatial resolution based on the relationship between air pollution values and land use variables observed at sampled points (Henderson *et al*., 2007). Most LUR literature focuses on NO_2, with a few studies modelling $PM_{2.5}$, ultrafine particles, and PM components (e.g., Henderson *et al*., 2007, Zhang *et al.*, 2015). This paper is the first study comparing models for three PM size fractions. Further novel elements in the well-established LUR literature are the inclusion of prevailing winds and the use of GIScience to advance spatial understanding of air pollution: an example of best practice for a spatial turn in health and environmental research (Richardson *et al*., 2013).

2. Methods

Air monitoring campaigns were conducted in Calgary in August 2010 and January-February 2011. A network of 50 monitors was deployed in each campaign (Bertazzon *et al*. 2015). Due to power outages and equipment failures, the campaigns yielded 27 valid summer PM samples and 29 winter samples. Predictor variables were defined on circular buffers from each sampling point. In addition, windrose variables were defined on buffers modified according to the prevailing winds in each season (Zhang *et al*. 2015).

Getis G and Moran's I spatial statistical tests were conducted to assess spatial clustering and autocorrelation in the variables, based on a row-standardized 3-nearest-neighbours spatial

weights matrix. Model selection was conducted on each PM size fraction: cross-correlation analysis selected one predictor from each category in Table 1, followed by backward variable selection (Bertazzon *et al*. 2015).

Table 1: Model Variables

Response variables	*Unit*
PM1.0, PM2.5, PM10	ug/m3

Land use variables	*Name*	*Unit or description*	*Circular buffers (meters)*	*Windrose buffer*	*dstnc*
Local roads	LRD	Total length of road segments within buffer, in meters	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	√
Major (arterial) roads	MRD		100, 200, ..., 500, 750, 1000	1500, 3000, 5000	√
Primary highways	PHW		100, 200, ..., 500, 750, 1000	1500, 3000, 5000	√
Expressways	EXPW		100, 200, ..., 500, 750, 1000	1500, 3000, 5000	√
Sum:MRD+PHW+EXPW	SMRD	Sum of segments	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Sum: PHW + EXPW	EXPHW	Sum of segments	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Traffic volume	TV	Year avg weekday T	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Population density	POP_den	Pop.in DB×DB buff. prt/ inters. area	100, 200, ..., 500, 750, 1000, 1500, 2000, 2500	1500, 3000, 5000	
Dwelling density	DWL_den	Dwl.in DB×DB buff. prt/ inters. area	100, 200, ..., 500, 750, 1000, 1500, 2000, 2500	1500, 3000, 5000	
Land use: residential	LU_res	Zoning category	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Land use: parks	LU_park	Zoning category	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Land use: institutional	LU_inst	Zoning category	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Land use: commercial	LU_com	Zoning category	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Land use: industrial	LU_ind	Zoning category	100, 200, ..., 500, 750, 1000	1500, 3000, 5000	
Industr. PM emissions	PM_EM	Report emitting pts	1000, 2000, 3000, ..., 6000	1500, 3000, 5000	√

Environmental var.s	*Name*	*Unit*
Elevation	Elev	meters
Wind speed & direction	WS_N, WS_E, WS_S, WS_W	km/hr at 10 m heigth

3. Results

Descriptive statistics for the three sets of independent variables are summarized in Table 2.

Table 2: Standard and Spatial Descriptive Statistics

		Sample	Min.	Max.	Range	Mean	S. D.	S-W	p (SW)	Moran	p (I)	Getis G	p (G)
PM1.0	summer	27	4.86	7.35	2.49	6.37	0.53	0.96	0.41	-0.05	0.84	0.12	0.30
	winter	29	1.50	7.36	5.86	4.18	1.23	0.98	0.84	0.06	0.09	0.11	0.12
PM2.5	summer	27	7.03	10.74	3.71	8.41	0.85	0.94	0.11	-0.01	0.81	0.04	0.26
	winter	29	2.32	9.75	7.43	5.48	1.65	0.98	0.85	0.07	0.42	0.04	0.06
PM10	summer	27	11.30	23.76	12.46	15.16	2.69	0.90	0.01	0.11	0.25	0.04	0.16
	winter	29	4.17	16.33	12.16	9.13	3.09	0.97	0.60	0.15	0.25	0.04	0.03

Particulate matter levels exhibited higher mean values in the summer. Spatial autocorrelation and clustering were never significant according to Moran's I and Getis G tests. The negative sign of Moran's I for $PM_{1.0}$ and $PM_{2.5}$ in summer suggests a dispersed, rather than clustered, spatial pattern. The Shapiro-Wilks test indicated normality for most distributions, except for summer PM_{10}. Histograms and q-q plots did not indicate large

anomalies; therefore, after analyzing the log-transformed variables, all models were run on the raw variables. Seasonal LUR models for each pollutant are summarized in Table 3.

Table 3: Summer and Winter LUR Models for $PM_{1.0}$, $PM_{2.5}$, and PM_{10}

Summer PM1.0	std. β	t value	partial R^2
Intercept	6.60	40.35	
LU_indwr_3000	0.48	3.13	0.21
LRD_dist	-0.47	-3.08	0.19
EXPW_dist	-0.05	-1.56	0.02
R^2	0.49	Adj. R^2	0.42
AIC	33.20	Res. SE	0.40
Res Moran I	-0.15	p (RI)	0.78
BP test	0.44	p (BP)	0.93

Summer PM2.5	std. β	t value	partial R^2
Intercept	7.87	39.49	
LU_indwr_3000	0.78	7.35	0.56
LRD_dist	-0.34	-3.13	0.10
SMRD750	0.28	2.60	0.09
R^2	0.75	Adj. R^2	0.72
AIC	39.71	Res. SE	0.45
Res Moran I	0.02	p (RI)	0.62
BP test	2.81	p (BP)	0.42

Summer PM10	std. β	t value	partial R^2
Intercept	14.19	33.47	
LU_indwr_5000	0.64	5.53	0.48
LRD_dist	-0.24	-2.28	0.06
EXPW400	0.34	2.90	0.22
R^2	0.75	Adj. R^2	0.72
AIC	101.25	Res. SE	1.42
Res Moran I	-0.12	p (RI)	0.57
BP test	2.15	p (BP)	0.54

Winter PM1.0	std. β	t value	partial R^2
Intercept	3.38	8.62	
PM_EM6000	0.38	2.45	0.21
MRDwr_3000	0.29	1.87	0.13
LU_park200	-0.32	-2.13	0.13
R^2	0.47	Adj. R^2	0.41
AIC	84.87	Res. SE	0.95
Res Moran I	-0.11	p (RI)	0.62
BP test	1.52	p (BP)	0.68

Winter PM2.5	std. β	t value	partial R^2
Intercept	4.25	11.34	
PM_EM6000	0.31	1.84	0.17
LU_ind300	0.46	2.76	0.28
MRD200	0.27	1.95	0.06
R^2	0.51	Adj. R^2	0.45
AIC	99.68	Res. SE	1.22
Res Moran I	-0.04	p (RI)	0.46
BP test	2.59	p (BP)	0.46

Winter PM10	std. β	t value	partial R^2
Intercept	8.31	10.90	
PM_EM6000	0.30	1.82	0.16
LU_ind300	0.32	1.91	0.20
LU_park200	-0.35	-2.40	0.18
R^2	0.54	Adj. R^2	0.48
AIC	134.3	Res. SE	2.22
Res Moran I	0.00	p (RI)	0.33
BP test	3.19	p (BP)	0.36

Summer models yielded better results for coarser particulate, with $R^2 > 0.75$ for $PM_{2.5}$ and PM_{10}, and $R^2 = 0.49$ for $PM_{1.0}$. These models contained very similar sets of predictors. *Industrial-land-use* was the largest contributor to all models, on very large buffers, ranging from 3,000- to 5,000-meter radii, their shape affected by the prevailing wind (i.e., windrose). The second contributor, local traffic, was represented by the same variable in all models: *Distance-from-local-roads*. The third contributor was *Expressways* for $PM_{1.0}$ and PM_{10}, and *Sum-of-major-roads* for $PM_{2.5}$, on circular buffers ranging from 400- to 750-meter radii. The rank-order of local vs. arterial traffic was reversed in the PM_{10} model.

Winter models yielded R^2 values between 0.47 and 0.54, with slightly higher values for coarser particulate. The R^2 value was consistent with the summer value of $PM_{1.0}$, and substantially lower for $PM_{2.5}$ and PM_{10}. Industrial emissions were the main contributor to all three models, represented by *Particulate-matter-emissions*, constantly on very large, 6,000-meter radius buffers. Two of the three models featured a second prominent predictor representing industrial activities: *Industrial-land-use*, on a much smaller 300-meter radius buffer. Major arterial traffic was significant for $PM_{2.5}$ and marginally significant for $PM_{1.0}$. As in the summer models, its buffer was small for $PM_{2.5}$, but large and affected by prevailing winds for $PM_{1.0}$. *Park-land-use-within-200-meter-buffer* was significant for $PM_{1.0}$ and PM_{10}.

Standard regression diagnostics and residual tests for all models provided no evidence that any model assumptions were violated. Spatial clustering or autocorrelation in all model residuals were not significant according to the Lagrange multipliers and Breusch-Pagan tests.

4. Discussion

Spatial analyses confirmed $PM_{1.0}$, $PM_{2.5}$, and PM_{10} as regional pollutants, characterized by moderate spatial variation, with non-significant spatial clustering and autocorrelation in both seasons. Recorded particulate concentrations were lower in the winter. Summer models

yielded higher goodness of fit, but winter models were more consistent across size fractions. Analytical results were consistent and interpretable, despite the low sample size.

Although model selection was conducted independently for each pollutant, it led to a remarkably consistent set of predictors, particularly for the summer models. Predictors of the summer models indicate significant association of particulate matter with industrial activities and with traffic, at the local and arterial levels. The correlation of PM with large windrose industrial buffers suggested that particulate matter of industrial origin was found at large distances from the source, with movement affected by prevailing summer winds. Conversely, correlation with relatively small circular traffic buffers suggested traffic-related PM, on local and major roads, decays rapidly as distance from roads increases.

Winter models suggested the association with industrial emissions was even stronger, particularly for coarser sizes, as $PM_{2.5}$ and PM_{10} models contained two predictors representing industrial activities. Like in summer, industrial predictors were selected on very large buffers. By contrast, winter buffers were circular, suggesting a lesser role of the wind on the widespread pattern of PM pollution of industrial origin. Association of PM with traffic was somewhat weaker in the winter, as local traffic was never significant, whereas arterial traffic was only significant for $PM_{2.5}$ and marginally significant for $PM_{1.0}$. Nonetheless, the spatial pattern of traffic pollution was consistent with the summer, with small buffers indicating rapid pollution decay as distance from roads increased. Distance from parks and open spaces, on very small buffers, was significant in the winter for $PM_{1.0}$ and PM_{10}. With most areas of the city typically covered by snow, this may indicate that particulate levels were lower over snow-covered open spaces in the winter.

5. Conclusion

Recorded PM concentrations were higher in the summer. LUR models yielded $R^2 > 0.75$ for $PM_{2.5}$ and PM_{10} in the summer, and $R^2 > 0.45$ for summer $PM_{1.0}$ and for all PM size fractions in the winter. Summer predictors were industrial emissions, local traffic, and major arterial traffic. Winter predictors included industrial emissions, industrial land use, major arterial traffic, and distance from open, snow-covered spaces. For all size fractions, the models suggested that industrial pollution extended over large areas in both winter and summer, and was affected by prevailing winds in summer; whereas traffic-related pollution, both on local roads and on major roads, decayed rapidly as distance from roads increased, in both seasons. These results are being shared with clinicians and used to inform in the creation of more environmentally-advanced models in a second study currently underway.

References

Bertazzon, S., Johnson, M., Eccles, K., & Kaplan, G. (2015). Accounting for spatial effects in land use regression for urban air pollution modelling. *Spatial and Spatio-temporal Epidemiology,* **14–15**: 9–21.

EPA, United States Environmental Protection Agency (2016). Criteria Air Pollutants. https://www.epa.gov/criteria-air-pollutants. Retrieved May 2, 2016.

Henderson, S., Beckerman, B., Jerrett, M., & Braurer, M. (2007). Application of land use regression to estimate long-term concentration of traffic-related nitrogen oxides and fine particulate matter. *Environmental Science and Technology*, **41**(7), 2422–2428.

Kelly, F. and Fussell, J. (2012). Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter. *Atmospheric Environment*, **60**: 504–526.

Richardson, D., Volkow, N., Kwan, M.-P., Kaplan, R., Goodchild, M., & Croyle, R. (2013). Spatial turn in health research. *Science*, **339**(6126), 1390-1392.

Rückerl, R., Schneider, A., Breitner, S., Cyrys, J., & Peters, A. (2011). Health effects of particulate air pollution: A review of epidemiological evidence. *Inhalation Toxicology*, **23**(10): 555–592,

Zhang, J., Sun, L., Barrett, O., Bertazzon, S., Underwood, F., & Johnson, M. (2015). Development of land-use regression models for metals associated with airborne particulate matter in a North American city. *Atmospheric Environment,* **106**: 165–177.

Fast Computation of Continental-Sized Isochrones

Paolo Bolzoni[1], Sven Helmer[1], Oded Lachish[2]

[1] Faculty of Computer Science, Free University of Bozen-Bolzano, 39100 Bolzano, Italy
Email: {firstname.lastname}@unibz.it

[2] Dept. of Computer Science and Information Systems, Birkbeck, University of London, London WC1E 7HX, United Kingdom
Email: oded@dcs.bbk.ac.uk

Abstract

We propose an approach to speed up the computation of isochrones, which are maps showing the reachability of locations given a starting point and a time constraint. The core idea of our technique is to materialize large parts of an isochrone, demonstrating how this can be achieved for multi-modal transport networks in a scalable way. We illustrate the effectiveness of our method with the help of an experimental evaluation.

1. Introduction

Isochrone maps, which given a starting location show the reachability of places within a certain time span, have been around for more than a hundred years. Before the introduction of computer systems and the digitization of map data, creating these artifacts was a very time-consuming task. The easy availability of geographical information systems (GIS), such as PostGIS, and map data, such as OpenStreetMap, has sparked a renewed interest in isochrones. The applications of isochrone maps are manifold. For example, in urban and regional planning they can be used to determine the location of public services, such as hospitals, schools, police and fire stations, making sure that catchment areas cover an adequate part of the population. When planning new transport links, isochrones can help in identifying zones that are not well-connected. Establishing evacuation routes for emergency situations can be facilitated as well. In real estate applications potential buyers and tenants can check which accommodations are within easy reach of workplaces and schools. Finally, planning trips on the fly to quickly determine which places are reachable in an acceptable time frame is also made easier.

Currently, there are no isochrone algorithms that scale to very large and detailed maps. Our goal is to compute isochrones very efficiently, in the ideal case in real-time. While there are efficient algorithms for route planning, most of them are uni-modal, i.e., only one mode of transportation, e.g. by car, is used. Usually, people change their mode of transportation a few times when moving from one location to another, though. There are only a few sophisticated algorithms for multi-modal route planning, e.g. Bast *et al.* (2015), but they only compute shortest paths from point to point.

While a lot of the early work on algorithms for computing isochrones, and also some of the more recent work, assumes that the underlying network is fairly homogeneous, Gamper *et al.* (2012) specifically investigate isochrones for multi-modal transportation networks. These approaches are implemented on top of database systems utilizing geographical information system features. On the one hand, this makes it easier to implement the computation of isochrones, as some general-purpose database system functionality can be re-used. On the other hand, not even geographical information systems support or are optimized for the direct computation of isochrones, so there is still a lot of code outside of the database system that needs to be written.

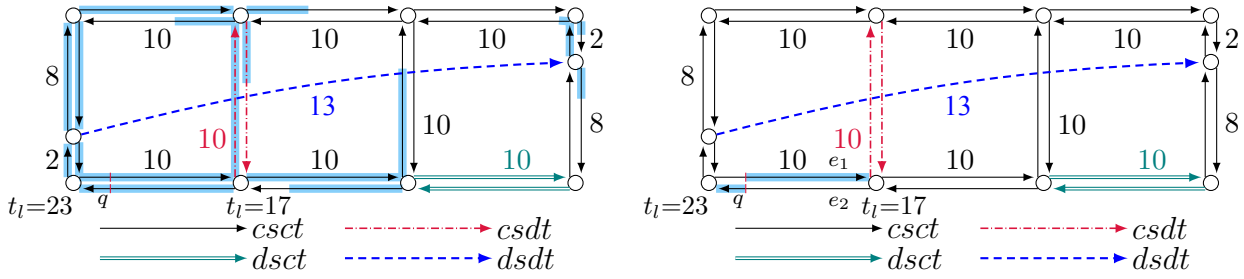
Our technique offers a highly scalable approach for computing isochrones on multi-modal transportation networks efficiently. In summary, we make the following contributions: we develop a data structure that precomputes and materializes a large part of an isochrone map; as a result, our algorithm can assemble large parts of the answer by sequentially scanning the data structure and in the case of a

single-mode scenario only a single sequential scan is required; we evaluate our technique experimentally using real-world data extracted from OpenStreetMap for Europe, showing that we can compute large isochrones quickly.

2. Problem Definition

Let us briefly define the multi-modal transportation network we use (it is similar to the one used by Booth *et al.* (2009)). Vertices represent noteworthy points in the street network like crossings or bus stops, arcs represent connections between these points. Given an arc e we define $d(e)$ as its destination, and $s(e)$ as its source. Moreover, we need a set M of transportation modes that are associated with arcs. Each mode $m \in M$ has a label (e.g. pedestrian, car, bus, train) and a description whether the mode is discrete or continuous in space (ds or cs) and discrete or continuous in time (dt or ct). An example for a continuous space and time ($csct$) mode is a pedestrian network. Any point can be reached at any time. Public transport systems, such as trains and buses, are discrete in space and time ($dsdt$), as they only run at specific times and can only be boarded or left at certain locations, e.g. stations and bus stops. Examples for a mode that is discrete in space and continuous in time ($dsct$) are escalators and tunnels: they operate continuously, but a person cannot get off before reaching the end. Finally, a mode continuous in space and discrete in time ($csdt$) are roads closed at specific times. Due to the multiple different modes, we actually have a multiset of edges, i.e., parallel arcs are possible (e.g. by walking or taking the bus between two stops).

In order to compute an isochrone *Iso*, we need a starting location q, a time threshold t_{max}, and a starting time t^s (needed for the departure times of ds arcs). The isochrone for a query includes all parts of the network that can be reached from q within t_{max}. The answer to q very likely consists of partial arcs around the boundary of the isochrone (see Figure 1(a) for a query q, located in the lower left corner, with $t_{max} = 25$; please also note the different transportation modes, in fact the bus on the $dsdt$ arc is intended to depart at $t^s + 9$ and it is impossible to cross the discrete arcs in the bottom right). Furthermore, the location q does not have to be a vertex: it can lie on continuous space (cs) arcs and then has to be converted to queries starting from the endpoints of the arcs. In Figure 1(b) q lies on the arcs e_1 and e_2, both of which have a weight of 10. Moreover, we know that q has a distance of 8 to $d(e_1)$ and a distance of 2 to $d(e_2)$, which means that at $d(e_1)$ we have 17 minutes left and at $d(e_2)$ 23 minutes.



(a) Example **(b) Starting queries**

Figure 1: Isochrones

3. Our Approach

The basic idea of our approach is to precompute and materialize a large part of isochrones, namely those along the pedestrian network (or any other large $csct$ network). For every node $x \in V$ we create

a list L_x of triplets $((e, \delta(x, s(e)), w_e)$, where $e \in A$ (A being a multiset of directed edges, or arcs), w_e is the weight of the arc e, and $\delta(u, v)$ computes the minimal distance between $u \in V$ and $v \in V$. Basically, we compute the distance δ from node x to the starting node $s(e)$ of every arc e. This also includes starting nodes of arcs that are not part of the $csct$ network. However, to compute $\delta(x, s(e))$ we only use $csct$ arcs and the triplets in a list are sorted in increasing distance of $s(e)$ from x. This data structure allows us to do a very fast computation of the $csct$ components of an isochrone: we just need to sequentially scan L_x. Every time we encounter a non-$csct$ arc, though, we have to trigger a new subquery. In our case we use the algorithm by Johnson (1977)[1] applied to the pedestrian network computing the all-pairs shortest paths to determine the values for δ.

However, applying our technique in a straightforward way does not scale: in the worst case we need storage space quadratic in the number of nodes. Therefore, we partition the graph and apply our technique to every partition (connecting the individual partitions in the process). We employ METIS by Karypis and Kumar (1999), a fast and readily available state-of-the-art algorithm for partitioning graphs; we use the kMETIS variant. We configured it to create partitions minimizing the number of cut arcs under the constraints of producing contiguous partitions and balancing them (in our case the allowed difference in size is at most 3%).

Querying works in the following way. We store the queries to be processed in a priority queue Q, which is initialized with the starting point q (or two starting points, in case q lies on an arc), the starting time t^s, and t_{max}. Q sorts the queries by partition identifier (in order of appearance), breaking ties via the duration of a query in descending order. In other words, if the partition of the starting node v_i of a query q_i that is added to Q is already in Q, then q_i is grouped with these queries (ordering the queries in this group by the remaining time t_i^l). Otherwise, q_i is appended to the end of the queue. Grouping queries by partitions means we can keep the processing localized, not jumping back and forth between different parts of the graph. Ordering queries by duration allows us to do effective pruning. If we have already run a query from a starting node v_i with more time remaining, then there is no reason to run one from the same point with less time remaining. The former query will always cover a greater area.

While the query in front of the queue is still in the same partition as the previously executed one, we process it, remove it from the queue, add new (partial) arcs to the isochrone, and enqueue any new queries resulting from the processing of the current one (no new queries will be scheduled if the distance of the nodes in the materialized list becomes too great). If the next query to be processed accesses a different partition, we switch to a new partition. Once Q runs empty, the algorithm terminates.

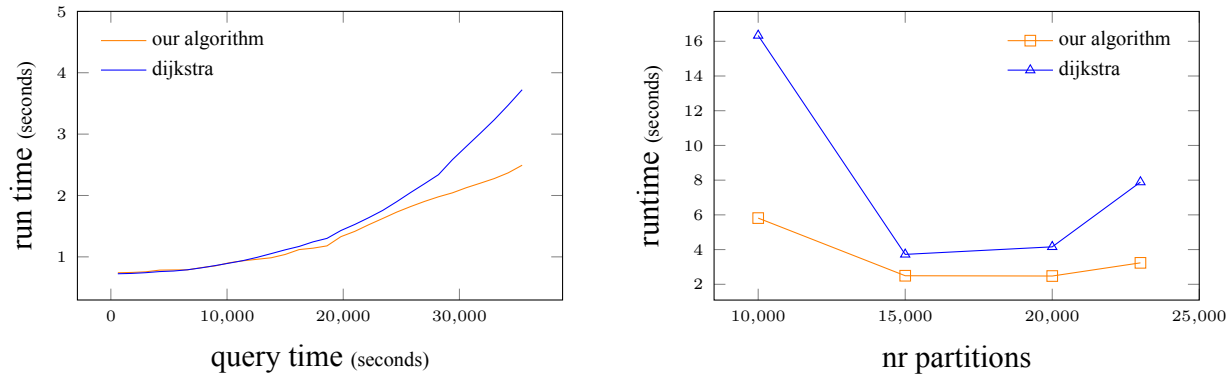
Processing a triplet also depends on what kind of arc we are facing: $csct$, $csdt$, $dsct$, or $dsdt$. The simplest case is $csct$, as this means that we are staying within the materialized network. We merely have to figure out which part of the arc we cover. We calculate the fraction l_e of the arc we can reach, given the time that is left after arriving at $s(e)$ and the (partial) arc l_e is added to the isochrone. If the arc e crosses into another partition, we have to schedule a new query, which is added to Q. The other types of arcs are not much more difficult to handle. $csdt$ arcs, which are closed at certain times, may introduce waiting times, but apart from this are treated in the same way as $csct$ arcs. The discrete space arcs ($dsct$ and $dsdt$) can only be traversed fully or not at all.

4. Experimental Evaluation

We evaluated our algorithm on a PC with an Intel i7-4800MQ CPU running at 2.7 GHz, 24 GBytes RAM, and a 120GB solid state disk. The dataset, taken from OpenStreetMap, is a map of continental Italy comprising 3.7 million vertices and 9.7 million edges.

Figure 2(a) shows a comparison of our technique to running Dijkstra's algorithm to find an isochrone

[1] For sparse networks, Johnson's algorithm is more efficient than Floyd-Warshall.



(a) Run time comparison **(b) Effect of partitioning**

Figure 2: Experiments

using 20,000 partitions. We make a couple of observations. Gamper *et al.* (2012) implemented Dijkstra's algorithm by loading the whole network into main memory, resulting in a suboptimal performance. For an isochrone of ten hours Dijkstra took 100 seconds and their algorithm MINEX even longer than that. We adapted Dijkstra to profit from the graph partitioning as well, leading to a much more competitive approach. Nevertheless, our algorithm still outperforms it for large isochrones, making it much more scalable. We also investigated the effect of the number of partitions on the performance (see Figure 2(b) for ten hour queries). As it turns out, there is a sweet spot: a small number of large partitions means we load too much unused data during query processing, a large number of small partitions means we lose time due to too many I/O operations. We suppose that the optimal number of partitions is related to the number of (urban) agglomerations, but at the moment we determine this number experimentally.

5. Conclusion and Future Work

We show how to implement the computation of isochrones in a scalable way, i.e., we can determine large isochrones on OpenStreetMap graphs very quickly (e.g. a ten hour isochrone for a map of Italy takes on average less than three seconds). We manage to do so by precomputing and materializing parts of the solution. For future work, we see optimization potential in parallelizing the algorithm and developing more intelligent partition loading strategies. Additionally, we want to work on efficiently rendering the full contours of an isochrone.

References

Bast, H., Delling, D., Goldberg, A. V., Müller-Hannemann, M., Pajor, T., Sanders, P., Wagner, D., and Werneck, R. F. (2015). Route planning in transportation networks. *Computing Research Repository (CoRR)*, **abs/1504.05140**.

Booth, J., Sistla, P., Wolfson, O., and Cruz, I. F. (2009). A data model for trip planning in multimodal transportation systems. In *Proc. of the 12th Int. Conf. on Extending Database Technology (EDBT'09)*, pages 994–1005, Saint Petersburg, Russia.

Gamper, J., Böhlen, M. H., and Innerebner, M. (2012). Scalable computation of isochrones with network expiration. In *Proc. of 24th Int. Conf. Scientific and Statistical Database Management (SSDBM'12)*, pages 526–543, Chania, Crete, Greece.

Johnson, D. B. (1977). Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM*, **24**(1), 1–13.

Karypis, G. and Kumar, V. (1999). A fast and highly quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, **20**(1), 359–392.

A Regional Approach for Modeling Dog Cancer Incidences with Regard to Different Reporting Practices

G. Boo[1, 2], S. Leyk[3], S. I. Fabrikant[1], A. Pospischil[2]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
Email: {gianluca.boo; sara.fabrikant}@geo.uzh.ch

[2]Collegium Helveticum, Zurich, Switzerland
Email: apos@vetpath.uzh.ch

[3]Geography Department, University of Colorado, Boulder, CO, USA
Email: stefan.leyk@colorado.edu

Abstract

Underreporting is a persistent limitation in research on environmental risk factors for dog cancer, impeding potential comparative investigations with human cancer. To address this challenge, we propose a regional modeling approach accounting for different reporting practices across the study area. In doing this, we demonstrate the need for new modeling strategies to improve statistical performance through more systematic assessments of spatial non-stationarity of statistical associations that can be linked to underreporting.

1. Introduction

Humans and dogs have been sharing their habitat for millennia by being exposed to similar environmental conditions over time. An interesting aspect of this co-evolutionary process is that the development of cancer in humans and dogs might be comparable. Environmental exposures associated with cancer in dogs could thus serve to timely identify risk for humans (Pinho *et al.* 2012). However, to date, only few studies have addressed possible linkages between environmental risk factors and dog cancer. This gap is due to uncertainty in most existing dog cancer databases, typically because of underreporting. For this reason, in this study, we investigate underreporting of dog cancer, and propose a regional modeling approach to account for different reporting practices across the study area.

We develop a case study based on the Swiss Canine Cancer Registry (SCCR), a unique dog cancer database that has been assembled for comparative investigations with humans. We model dog cancer incidences at the municipal level using demographic variables and indicators accounting for underreporting. Based on the model residuals, we decompose Switzerland into regions of similar model fit and build regional models of dog cancer incidences. We finally compare statistical performance and spatial distributions of model residuals to identify changes in the statistical associations across the study area. In doing this, we aim to demonstrate that underreporting challenges the use of statistical models of dog cancer incidences, and that a regional modeling approach improves statistical performance by mitigating spatial non-stationarity of statistical associations.

2. Data

The SCCR stores diagnostic records collected in Switzerland between 1955 and 2008, and it is in the process of being updated to more recent years. Comprising more than 120,000 records, this is the largest and most durable animal cancer database, to date. The present study is based on the 3,509 cancer examinations performed in 2008, which have been enumerated within the 2,351

Swiss municipal units to protect the privacy of the dog owners. We retrieved dog census data for the same year to assess the impact of demographic characteristics accountable for cancer development. This census has been established in 2006 following the nationwide obligation of dog registration. For our study purposes, we summarized information about age and sex of the dog population at the municipal level.

We used indicators of urban character and socio-economic status, as our prior research has shown that these variables successfully account for underreporting of dog cancer. The urban character is estimated based on human population densities in 2008, as provided by the Swiss Federal Statistical Office. The socio-economic status is derived from average national income tax information collected by the Swiss Federal Tax Administration in 2008. We also derived information on access to veterinary care from a hectometric distance-raster based on the addresses of the 938 veterinarians active in 2013. The distance-raster is constrained along roads to compute travel distances, and values are averaged to estimate access to veterinary care at the municipal level. We used more recent addresses of registered veterinarians because historical data are not available.

3. Methods

We model dog cancer incidences through a Poisson regression because this method is commonly used to identify risk factors for the spatial distribution of rare diseases (Frome 1983). We choose this model, in spite of possible over-dispersion, as the spatial distribution of model residuals can inform about potential misspecifications. We fit dog cancer incidences at the municipal level through the following variables: Dog Average Age, Female Dog Ratio, Average Income Tax, Human Population Density, and Distance to Veterinary Care. Dog Population is used as offset. Statistical performance is evaluated with the McFadden pseudo R-squared and the analysis of variance reduction, including a spatial decomposition based on model residuals (Cameron and Windmeijer 1997). We used Pearson residuals because absolute values exceeding 2 highlight lack of model fit.

The spatial distribution of model residuals is used to define regions of similar model fit, and inform about spatial non-stationarity of statistical associations (Brunsdon *et al.* 2008). We identify contiguous regions by means of a connectivity graph algorithm (minimum spanning tree), based on Queen contiguity to determine adjacent municipal units (Duque *et al.* 2007). The optimal number of regions is selected through the pseudo F-statistic, indicating the number of regions with maximum internal similarity and external dissimilarity (Ketchen and Shook 1996). We then fit models of dog cancer incidences for each region separately. We finally compare the statistical performance of the regional models with the global model and explore the spatial distributions of Pearson residuals to assess whether spatial stationarity in the statistical associations has improved.

4. Results and Discussion

For the global model, the McFadden statistic shows a value of .50. This means that 50% of the total variability in dog cancer incidences is explained by the model. The variables accounting for underreporting of dog cancer are statistically significant ($p<.01$) and contribute to 87% of the overall variance reduction. The remaining 13% of variance reduction is explained by the demographic variables, which are also statistically significant ($p<.01$). Our results confirm the expected effects of indicators accounting for well-known sources of underreporting of dog cancer. Nevertheless, the spatial distribution of the model residuals presented in Figure 1a shows

several regions of poor model fit, suggesting that the global model might be affected by spatial non-stationarity of statistical associations.

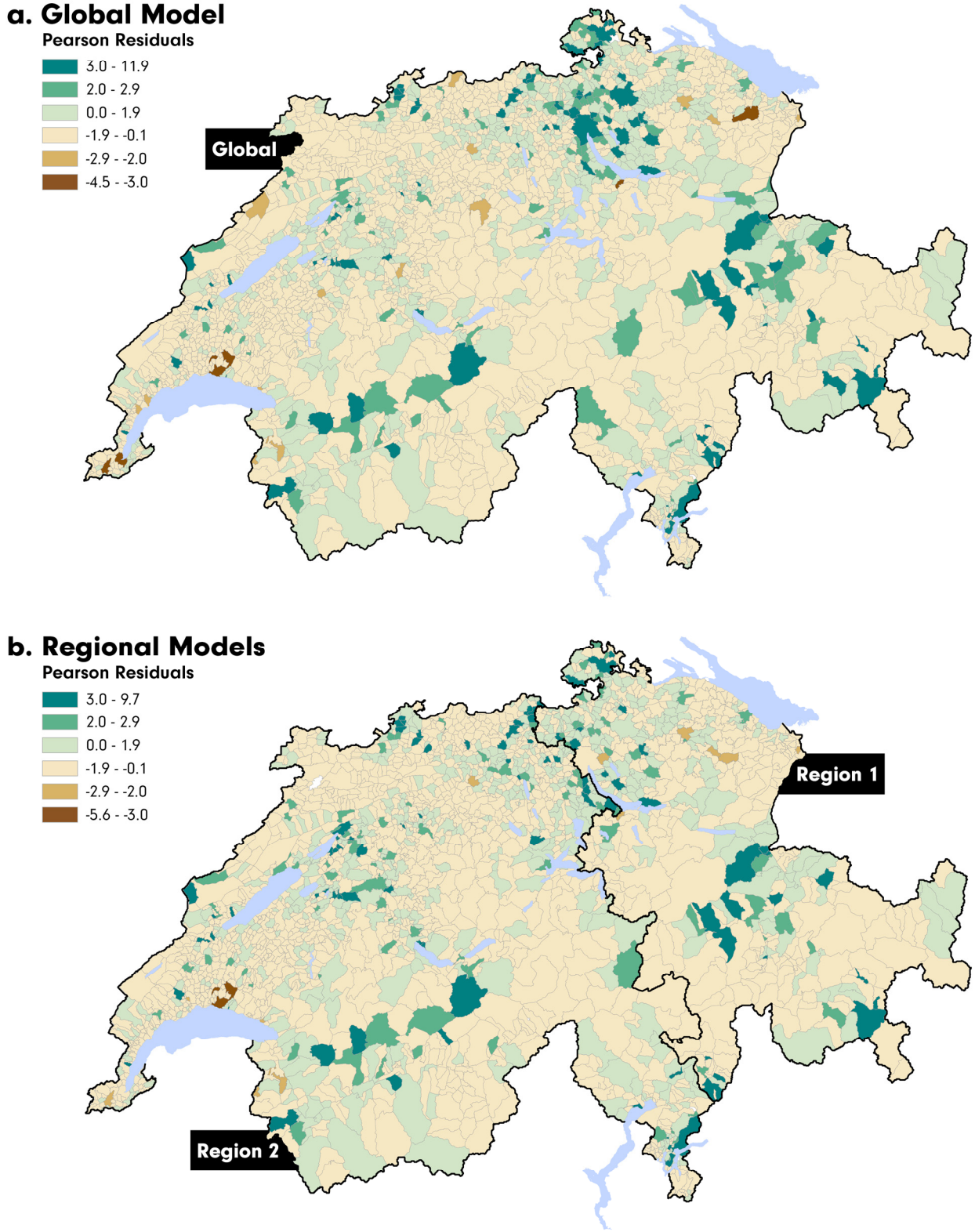


Figure 1. Spatial distribution of Pearson residuals for the global model (a) and the two regional models (b).

Our regional approach produced two macro-regions of distinct model fit: Region 1 and Region 2, consisting of 548 and 1,803 municipalities, respectively. When regional models of dog cancer incidences are fit to the two regions, most variables are statistically significant ($p < .01$) with a McFadden statistics of .58 for Region 1 and .46 for Region 2. Interestingly, the overall variance reduction explained by indicators of underreporting drops to 72% for Region 1, while it increases to 95% for Region 2. This confirms a decreased impact of underreporting of dog cancer for Region 1, but an increase for Region 2. Figure 1b presents the spatial distribution of Pearson residuals across both regions and shows a general improvement of model fit when compared with Figure 1a. This suggests that effects of spatial stationarity in statistical associations could be mitigated to some extent using the proposed regional modeling approach.

5. Conclusions and Outlook

The results of this case study demonstrate that the presence of different reporting practices across the study area challenges the use of statistical models for fitting dog cancer incidences. To address this issue, we propose a regional modeling approach to specifically delineate underreporting regimes across the study area. In our study, we identify two distinct regions with different characteristics. While model fit for Region 1 shows improved statistical performance, likely due to a reduced impact of underreporting, model fit for Region 2 deteriorates, suggesting increased underreporting and thus less reliable statistical associations. Still, employing two regional models reduces non-stationarity of statistical associations observed in the global model.

In future research, we aim to further refine our regional modeling approach by deepening the understanding of spatial non-stationarity of statistical associations linked to underreporting (see for example Leyk *et al.* 2012). In doing so, we intend to propose a new framework for systematically investigating regional changes in statistical associations across the study area, as a strategy to mitigate effects of spatial non-stationarity of statistical associations.

References

Brunsdon C, Fotheringham AS and Charlton ME, 1996, Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4):281–298.

Cameron AC and Windmeijer FAG, 1997, An R-squared Measure of Goodness of Fit for Some Common Nonlinear Regression Models. *Journal of Econometrics*, 77(2):329–342.

Duque JC, Ramos R and Suriñach J, 2007, Supervised Regionalization Methods: A Survey. *International Regional Science Review*, 30(3):195–220.

Frome EL, 1983, The Analysis of Rates Using Poisson Regression Models. *Biometrics*, 39(3):665–674.

Ketchen DJ and Shook CL, 1996, The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, 17(6):441–458.

Leyk S, Norlund PU and Nuckols JR, 2012, Robust Assessment of Spatial Non-Stationarity in Model Associations Related to Pediatric Mortality due to Diarrheal Disease in Brazil. *Spatial and Spatio-temporal Epidemiology*, 3:95–105.

Pinho SS, Carvalho S, Cabral J, Reis CA and Gärtner F, 2012, Canine Tumors: A Spontaneous Animal Model of Human Carcinogenesis. *Translational Research*, 159(3):165–172.

AN ONTOLOGICAL ANALYSIS OF WATER FEATURES

Boyan Brodaric[1], Torsten Hahmann[2], Michael Gruninger[3]

[1]Geological Survey of Canada, Ottawa, Canada K1A0E9
Email: boyan.brodaric@canada.ca
[2]School of Computing and Information Science, University of Maine, Orono, ME 04469, USA
Email: torsten.hahmann@maine.edu
[3]University of Toronto, Toronto, Canada M5S3GB
Email: gruninger@mie.utoronto.ca

Abstract

Water features are understood and represented heterogeneously in a wide variety of settings, including in data standards, polices and regulations, and amongst different cultures and languages. Ontologies aim to reduce this heterogeneity by representing commonalities across such settings. In this paper we build upon existing work in hydro ontologies and philosophical ontology to enhance the conceptualization and representation of water features. This results in a new taxonomy for water features, which helps identify and organize their essential parts. The results are represented as a first-order logic extension of the DOLCE ontology as well as an independent ontology fragment, and these are intended to serve as a reference ontology for the hydro domain as well as an aid to data interoperability.

1. Introduction

Water features are entities that are essentially composed of water and variably other things. Prototypical examples include lakes, rivers, puddles, and clouds, but can also include aquifers. They play a key role in many human activities, such as those related to health, climate and weather, agriculture, energy, recreation, and transportation. Research and operations in these domains are heavily dependent on digital representations of water features, but the inherent conceptualizations can vary widely. Examples of heterogeneity abound, and can be found when comparing international water data standards (Boisvert & Brodaric 2012; Dornblut & Atkinson 2013; INSPIRE 2013; 2014), national catalogs of hydrographic features (Duce & Janowicz 2010), ontological considerations (Galton & Mizoguchi 2009; Santos *et al*. 2005; Sinha *et al*. 2014; Wellen & Sieber 2013), and database structures (Maidment, 2002; Strassberg, *et al*., 2011). This is problematic as it inhibits some uses, especially their integration, which is typically an important precursor to regional scientific analysis such as water availability, or complex societal decision-making such as water allotment. At the heart of the problem is a disparity about the fundamental nature of a water feature, as different aspects are variously emphasized in distinct conceptualizations. These aspects include most notably the water body, its water matter, its container or void (the space it occupies), or even an immaterial spiritual entity (Mark *et al*. 2007; Wellen & Sieber 2013). The emphases exist perhaps to enable diverse uses, for example, reasoning about the presence of a water body facilitates navigation of rivers that might have wet or dry segments; reasoning about the constitution and flow of water matter informs contamination scenarios, as does reasoning about the permeability of the container; and reasoning about the container's void informs storage and overflow scenarios. Yet, it is still somewhat surprising that an entity of such significance is so widely construed and often vaguely defined. In this paper we undertake an ontological analysis of water features and develop a new conceptualization and representation that encompass the key aspects. This is achieved by extending and uniting two significant approaches to physical ontology, namely Hayes' ontology of liquids (1978) and Fine's theory of parts (1999). The results contribute to the design of the HyFO reference

ontology, which is being developed for the hydro domain to help identify semantic heterogeneities, aid interoperability, and inform ongoing representation initiatives.

2. Background and Related Work

As part of Hayes' seminal ontological analysis, liquid features are delineated using several criteria, chief amongst them being containment and support. Containment refers to liquid being topologically surrounded, while support refers to it being held against a surface. These criteria help distinguish most of the representative examples: water bodies in lakes and rivers are contained and supported, in puddles they are uncontained and supported (assuming puddles are spills resting on relatively flat surfaces), while in clouds they are uncontained and unsupported. Note that essential aspects can then be identified for different water features: contained water features must have a container and a void, but uncontained features must not; supported water features must have a supporting boundary, but unsupported features must not. However, aquifers are not distinguished from rivers using these criteria, inasmuch as water bodies can be contained and supported by both aquifers and rivers, under general notions of containment and support. Hence, these notions alone are insufficient to delineate all the representative water features.

In addition to its failure to delineate the full range of water features, Hayes' ontology of liquids also does not provide great detail about the inner structure of a water feature, that is, its main components, their relations, and the relation to the water feature itself. This is advanced somewhat by Fine's notions of temporary and timeless parts, which respectively form variable and rigid wholes. Using rivers as a prototypical example, Fine segments the wet aspect of a water feature into two things linked as whole to part: (1) a persistent water body comprising a variable whole, and (2) its changing (temporary) water matter parts. It is the water body that, over time, persists and rises or falls within a container, and which consists of all the water matter in the container at a timepoint, while it is some specific water matter part that moves at variable speed within the container. However, several things remain unaddressed by Fine. Most notably, the relation of a water body to other aspects is unanswered, such as the relation to a container or to the water feature itself, and rigid wholes are not considered in relation to water features despite their potential application.

Related work on water feature ontologies (e.g. Galton & Mizoguchi 2009; Santos *et al*. 2005; Sinha *et al*. 2014; Wellen & Sieber, 2013) individually incorporate some, but not all, of the seemingly fundamental distinctions made by Hayes and Fine. In particular, the three entities emerging from the above work, namely the water feature, water body and water matter, are not distinguished by any one approach, and the complete range of representative water features is also not delineated. This work then complements the related efforts, by encompassing the full range of representative features and refining their internal structure.

3. Water Feature Conceptualization and Representation

A new conceptualization for water features is achieved by two additions to Hayes' and Fine's efforts: a notion of dependence is introduced to further categorize contained water features, and water features themselves are distinctly recognized and characterized as rigid wholes.

In previous work (Hahmann & Brodaric 2013), detachable and dependent containment denote whether the topological attachment between physical entities is accidental or necessary, respectively: for example, an amount of water matter is accidently contained by a riverbed because it could possibly be in a different riverbed, but the river channel (the void) is necessarily contained by its host riverbed and could not possibly be displaced elsewhere without being a different channel. Likewise, an aquifer and the stuff that it is made of—the rock plus water matter—are dependently contained, because they are necessarily spatially co-located such that if one boundary changes then so does the other. Using this distinction,

aquifers can be distinguished from rivers, because the water in aquifers is dependently contained, and the water in rivers is detachably contained. Interestingly, aquifers also can be variably supported (confined) or unsupported (unconfined) as per the permeability of their boundaries. This leads to an enhanced taxonomy of water features, as shown in Figure 1, in which the representative examples are delineated first by whether the water matter is contained or not, then supported or not, and finally by the type of dependence.

Further understanding of water features as rigid wholes rigorously grounds the water feature-body-matter distinction in Fine's whole-part theory, and it means that each water feature has a stable configuration of specific essential parts. In particular, each water feature has a water body as an essential part, but as implied by the taxonomy below, different water features can also have other essential parts. A cloud has a water body as its only essential part, whereas a puddle has a water body and supporting surface as essential parts, and a river and aquifer have their container, void, supporting surface, and water body as essential parts, all arranged in a particular topological relation unique to each type of feature. Importantly, rigid and variable wholes can be nested: indeed, each of the stable water feature parts is in turn a variable whole, meaning that each can itself have changing parts in time, such as a water body with changing water matter, or a riverbed with changing segments due to the effect of physical processes over time. In essence, while a water feature has fixed essential parts configured uniquely, each such part can be dynamically exchanged over time.

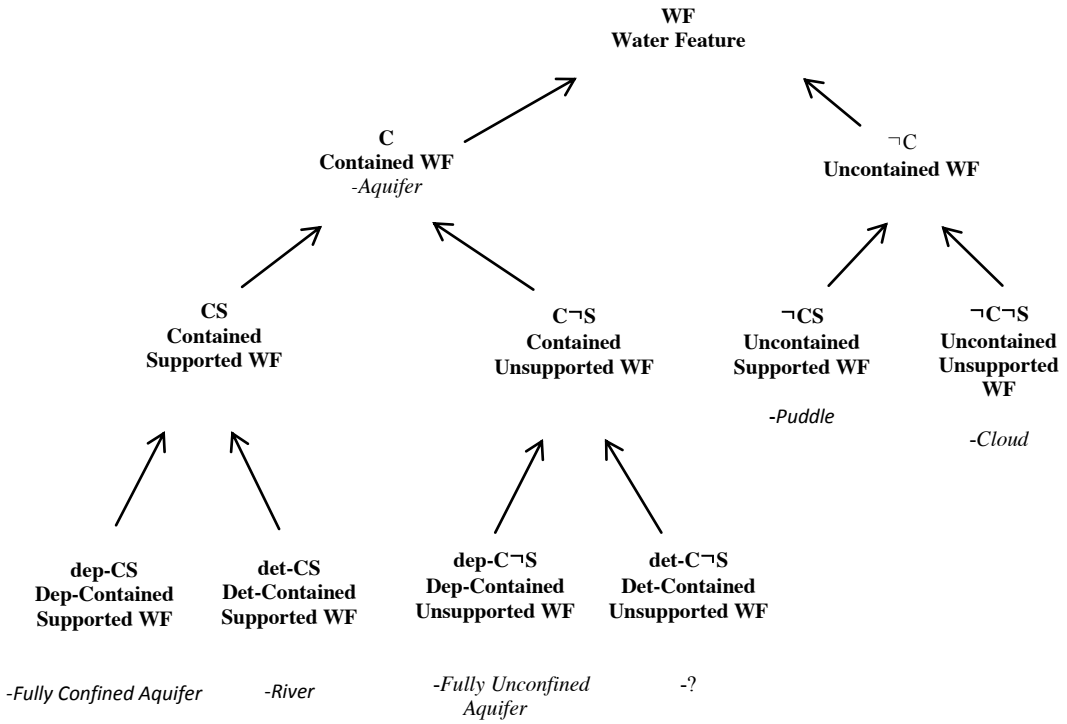


Figure 1. Taxonomy of water features.

This conceptualization of water features is being incorporated into the HyFO ontology, which is expressed in first-order logic as an extension of the DOLCE foundational ontology (Masolo *et al*. 2003): water features and bodies specialize DOLCE physical objects, water matter specializes DOLCE matter, while voids and supporting boundaries specialize DOLCE features. Important relations that bind these entities together include containment, void hosting, constitution, and dependence. The conceptualization can also be represented independently of DOLCE, as a so-called ontology design pattern. Unfortunately, neither representation can be adequately elaborated or illustrated here due to lack of space.

4. Summary and Future Directions

A new conceptualization for water features is developed by extending Hayes' ontology of liquids and applying Fine's theory of parts. The conceptualization is represented as a first-order logic extension of the DOLCE foundational ontology and as a lightweight ontology design pattern. It forms the core of the HyFO ontology, which is being developed as a reference ontology for the hydro domain and as an aid to interoperability. The conceptualization, via HyFO, is currently being tested in various ways, including via mappings with other hydro representations. Two interesting avenues remain to be explored in its design: applying dependence to the notion of support, to further distinguish types of support, and the inclusion of immaterial essential parts such as spiritual entities, as identified in ethnophysiographic studies. The conceptualization seems to be well positioned for enhancement in these directions. It remains to be seen, though, upon further testing, whether other notions deemed insufficiently fundamental will warrant inclusion, such as rate of water flow or degree of water consolidation, as originally suggested by Hayes.

Acknowledgements

The authors gratefully acknowledge support for this work obtained from the Groundwater Program of Natural Resources Canada and the Maine Economic Improvement Fund.

References

Boisvert E and Brodaric B, 2012, GroundWater Markup Language (GWML) – enabling groundwater data interoperability in spatial data infrastructures. *Journal of Hydroinformatics*, 14(1):93–107.

Dornblut I and Atkinson R, 2013, *Hy_features: a geographic information model for the hydrology domain*. Technical Report GRDC 43r1, Global Runoff Data Centre, November 2013.

Duce S and Janowicz K, 2010, Microtheories for Spatial Data Infrastructures – Accounting for Diversity of Local Conceptualizations at a Global Level. In: SI Fabrikant, T Reichenbacher, MV Kreveld and C Schlieder (Eds), *6th International Conference, GIScience 2010*, Springer, LNCS 6992, 27-41.

Fine K, 1999, Things and Their Parts. *Midwest Studies in Philosophy*, XXIII, 61-74.

Galton A and Mizoguchi R, 2009, The Water Falls but the Waterfall does not Fall: New perspectives on Objects, Processes, and Events. *Applied Ontology*, 4(2):71-107.

Hahmann T and Brodaric B, 2013, Kinds of full physical containment. In: T Tenbrink, J Stell, A Galton and Z Wood (Eds), *11th International Conference on Spatial Information Theory, COSIT-13*. LNCS 8116, Springer, 397-417.

Hayes PJ, 1978, *Naive Physics I: Ontology for Liquids*. Working Papers, No. 35, University of Geneva, 66 pp.

INSPIRE Thematic Working Group Geology, 2013, *D2.8.II.4 INSPIRE Data Specification on Geology – Draft Guidelines, v3.0 rc3*. Technical report, INSPIRE, 369 pp.

INSPIRE Thematic Working Group Hydrography, 2009, *D2.8.I.8 INSPIRE Data Specification on Hydrography – Guidelines, v3.0*. Technical report, INSPIRE, 175 pp.

Maidment DR, 2002, *Arc Hydro: GIS for Water Resources*. ESRI Inc, 203 pp.

Mark DM, Turk AG and Stea D, 2007, Progress on yindjibarndi ethnophysiography. In: S Winter, M Duckham, L Kulik and B Kuipers (Eds), *8th International Conference on Spatial Information Theory, COSIT 2007*. LNCS 4736, Springer, 1–19.

Masolo C, Borgo S, Gangemi A, Guarino N and Oltramari A, 2003, *Wonderweb Deliverable D18 – Ontology Library (Final Report)*. Technical report, National Research Council - Institute of Cognitive Sci. and Technology, Trento, 349 pp.

Santos P, Bennett B and Sakellariou G, 2005, Supervaluation Semantics for an Inland Water Feature Ontology. In: *Proceedings International Joint Conference on Artificial Intelligence*, *IJCAI-05*, 564–569.

Sinha G, Mark D, Kolas D, Varanka D, Boleslo ER, Feng C-C, Usery E, Lieberman J and Sorokine A, 2014, An Ontology Design Pattern for Surface Water Features. In: M Duckham, E Pebesma and K Stewart (Eds), *8th International Conference, GIScience 2014*. LNCS 8728, Springer, 187-203.

Strassberg G, Jones NL and Maidment DR, 2011, *Arc Hydro Groundwater : GIS for Hydrogeology*. ESRI Inc, 160 pp.

Wellen CC and Sieber RE, 2013, Toward an Inclusive Semantic Interoperability: The Case of Cree Hydrographic Features. *International Journal of Geographical Information Science*, 27(1):168-191.

Walk and Learn: An Empirical Framework for Assessing Spatial Knowledge Acquisition during Mobile Map Use

A. Brügger[1], K.-F. Richter[1], S. I. Fabrikant[1]

[1] Department of Geography, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
{annina.bruegger; kai-florian.richter; sara.fabrikant}@geo.uzh.ch

Abstract

We gladly use automated technology (e.g., smart devices) to extend our hard working minds. But what if such technology turns into mind crutches we cannot do without? Understanding how varying levels of automation in mobile maps might impact navigation performance and spatial knowledge acquisition will provide important insights for the ongoing debate on the potentially detrimental effects of using navigation systems on human spatial cognition. We need to identify the right balance between system automation (support) and user autonomy (self-reliance). Preliminary results of a pilot study performed within a novel empirical framework indicate that it is possible to increase user autonomy and spatial knowledge acquisition without negatively impacting navigation performance and usefulness of the system.

1. Introduction

Various research fields have investigated how system automation might influence human knowledge and abilities. It is generally agreed that computers often make decisions originally performed by humans in a (more) efficient way. This has positive and negative impacts on humans as, for example, automation can reduce their physical and cognitive effort (Sheridan 2002). This also holds for navigational tasks, where acquiring spatial knowledge is crucial to orient and move in space without getting lost. Recent developments in self-driving vehicles highlight the need for better understanding human behavior, especially when humans have to take over from automated systems during system failure (Merat *et al.* 2014). The ideal human-system interaction would be to use the best of both human and technology (Sheridan 2002), which we aim for in our research. Specifically, how do we balance the advantages of system automation and the need for human autonomy to maximize both navigation efficiency and knowledge acquisition?

2. Balancing Assistance and Engagement

Research investigating mobile navigation aids identified negative impacts on spatial knowledge acquisition, despite being very effective for efficient navigation (e.g. Willis *et al.* 2009). The consequences of automated guidance seem to be a disengagement of navigators' attention from their surroundings (Gardony *et al.* 2013), and split attention between mobile device and the traversed environment (Willis *et al.* 2009).

However, mobile navigation devices should enable pro-active engagement with the environment, which will lead to better spatial knowledge acquisition (Chung *et al.* 2016; Parush *et al.* 2007), as systems might break down, or users might lose the device and suddenly depend on their own abilities (Hirtle and Raubal 2013). The means to design such systems are yet unclear. Systems would need to provide efficient wayfinding support (sufficient system

automation) while at the same time engage users during the wayfinding process, such that they learn something about and from the environment (sufficient user autonomy).

Our research aims at finding the right balance between system automation and user autonomy. We are constructing an empirical framework in which we will explore various design solutions for mobile maps. This experimental setup aims to establish how system design decisions might affect users' spatial knowledge acquisition while also measuring navigation performance. The latter is important in ensuring that our experimental designs do not render the navigation task too difficult or too tedious.

A key aspect is that empirical studies with pedestrians are conducted in urban outdoor environments. We ask participants to follow a given route with the help of mobile map applications, which will vary the level of system automation for one or several cognitive processes relevant in navigation (e.g., self-localization or route planning). Adopting a between-subject design, we plan to always test at least two participant groups with different levels of automated features: either automation is permanently present or the user needs to initiate the required cognitive process. Intermediate levels will also be considered.

Subsequent to the assisted route-following task, participants are asked to find the exact same way back without any system assistance. Walking back will assess participants' acquired spatial knowledge. This is a hard task, as wayfinding decisions have to be reversed, and the navigator's perspective of the traversed environment will change. Strategies to encode and decode spatial knowledge vary across individuals and groups (Ishikawa and Montello 2006). A key factor in our analysis will thus be the assessment of differences in spatial abilities.

In order to support our findings, we further measure the navigator's eye movements using a mobile eye tracker to determine the influence of mobile map design on participants' environmental perception. Analyzing areas of interests defined for the route-following task allows for more systematically studying participants viewing behavior, for example, when fixating environmental features with changing perspectives (e.g., original route and return journey).

We hypothesize that increasing user autonomy as a result of lowering system automation will lead to increased spatial knowledge acquisition due to increased active engagement with both the navigation application and the environment.

3. Pilot Study

We are currently conducting a field study based on our novel empirical framework which tests the effects of two system automation levels on participants' self-localization process. Constant location updating on the mobile map seems to consume a user's attention (Willis *et al.* 2009), which changes how humans perceive the environment during navigation, and leads eventually to a loss of the fundamental skill of environmental information collection (Parush *et al.* 2007). Here, we report on preliminary results of a pilot study.

3.1 Method

Six participants (average=25.5 years) participated in the pilot study. On arrival, all participants filled in a demographic questionnaire, donned the mobile eye-tracker (Figure 1a), and conducted a training session with the application. Half of the participants used a mobile map, which constantly updated their location on the map ('always-on' group). The second half was instructed to press a button to get their location displayed on the map for ten seconds ('on-request' group). All participants followed the route shown in Figure 1b using one of the two map application

types. Once participants reached the destination, they were asked to walk the same route back to the starting point without any mobile map assistance. We recorded hesitations, stops, mistakes, task time, eye movements for both navigation directions, and all interactions with the application (e.g., zoom). After these route-following tasks, participants were to respond to the Building Memory test to assess their spatial memory abilities.

Figure 1. Participant wearing a mobile eye-tracker while using a tablet device (a) showing the route on the mobile map (b).

3.2 Results & Discussion

Preliminary results indicate that participants with a low level of system automation ('on request') are more actively involved during the navigation process. They interact more with the mobile map, and hesitate and stop more often along the way, possibly to verify that they are still on the right track. The 'always on' group hardly hesitated or stopped during the route-following task. Interestingly, completion time is similar for both groups. Overall, the 'on request' group had no problems in walking back and finding the start point again, while all 'always on' participants made at least one mistake, and one failed to identify the starting point.

Participants in the 'always on' group seem to have slightly lower scores than the 'on-request' group in the Building Memory test. But five of the six participants achieved 20 or more out of 24 possible points, which clearly demonstrates good spatial memory abilities. Still, participants in the 'always on' group did not find the way back without mistakes.

This leads us to conclude at this stage of our research that lowering the level of automation in the self-localization process most likely positively affects spatial knowledge acquisition. It seems possible to increase user autonomy without limiting the assisted navigation process. We still need to confirm this contention once all the empirical data has been collected and analyzed, including the eye-movement recordings.

References

Chung J, Pagnini F, Langer E, 2016, Mindful navigation for pedestrians: Improving engagement with augmented reality. *Technology in Society*, 45, 29-33.

Gardony AL et. al., 2013, How navigational aids impair spatial memory: Evidence for divided attention. *Spatial Cognition and Computation*, 13(4): 319–350.

Hirtle SC and Raubal M, 2013, Many to Many Mobile Maps. In: Raubal M, Mark D and Frank A (eds) *Cognitive and Linguistic Aspects of Geographic Space: New Perspectives on Geographic Information Research*. Berlin, Springer: 141–157.

Ishikawa T and Montello DR, 2006, Spatial knowledge acquisition from direct experience in the environment: Individual differences in the development of metric knowledge and the integration of separately learned places. *Cognitive Psychology*, *52*(2), 93–129.

Merat N et. al., 2014, Transition to manual: Driver behaviour when resuming control from a highly automated vehicle. *Transportation Research Part F: Traffic Psychology and Behaviour*, 27: 274–282.

Parush A, Ahuvia S, Erev I, 2007, Degradation in Spatial Knowledge Acquisition When Using Automatic Navigation Systems. *Spatial Information Theory*, 238–254.

Sheridan TB, 2002, *Humans and automation: System design and research issues*. John Wiley & Sons, Inc.

Willis KS et. al., 2009, A comparison of spatial knowledge acquisition with maps and mobile maps. *Computers, Environment and Urban Systems*, 33(2): 100–110.

Measuring Distance "As the Horse Runs": Cross-Scale Comparison of Terrain-Based Metrics

BP Buttenfield[1], M Ghandehari[1], S Leyk[1], LV Stanislawski[2], ME Brantley[1], and Yi Qiang[1]

[1]Department of Geography, University of Colorado, Boulder CO 80309
Email: {babs; mehran.ghandehari; stefan.leyk; margaret.brantley; yi.qiang} @colorado.edu

[2]U.S. Geological Survey (USGS), Center of Excellence for Geospatial Information Science, Rolla MO 65401
Email: lstan @usgs.gov
Disclaimer: Any use of trade, firm, or product names is for descriptive purposes and does not imply endorsement by the U.S. Government.

1. Introduction

Distance metrics play significant roles in spatial modeling tasks, such as flood inundation (Tucker and Hancock 2010), stream extraction (Stanislawski *et al.* 2015), power line routing (Kiessling *et al.* 2003) and analysis of surface pollutants such as nitrogen (Harms *et al.* 2009*)*. Avalanche risk is based on slope, aspect, and curvature, all directly computed from distance metrics (Gutiérrez 2012). Distance metrics anchor variogram analysis, kernel estimation, and spatial interpolation (Cressie 1993). Several approaches are employed to measure distance. Planar metrics measure straight line distance between two points ("as the crow flies") and are simple and intuitive, but suffer from uncertainties. Planar metrics assume that Digital Elevation Model (DEM) pixels are rigid and flat, as tiny facets of ceramic tile approximating a continuous terrain surface. In truth, terrain can bend, twist and undulate within each pixel.

Work with Light Detection and Ranging (lidar) data or High Resolution Topography to achieve precise measurements present challenges, as filtering can eliminate or distort significant features (Passalacqua *et al*. 2015). The current availability of lidar data is far from comprehensive in developed nations, and non-existent in many rural and undeveloped regions. Notwithstanding computational advances, distance estimation on DEMs has never been systematically assessed, due to assumptions that improvements are so small that surface adjustment is unwarranted. For individual pixels inaccuracies may be small, but additive effects can propagate dramatically, especially in regional models (e.g., disaster evacuation) or global models (e.g., sea level rise) where pixels span dozens to hundreds of kilometers (Usery *et al* 2003). Such models are increasingly common, lending compelling reasons to understand shortcomings in the use of planar distance metrics. Researchers have studied curvature-based terrain modeling. Jenny *et al.* (2011) use curvature to generate hierarchical terrain models. Schneider (2001) creates a 'plausibility' metric for DEM-extracted structure lines. d'Oleire-Oltmanns *et al.* (2014) adopt object-based image processing as an alternative to working with DEMs; acknowledging the pre-processing involved in converting terrain into an object model is computationally intensive, and likely infeasible for some applications.

This paper compares planar distance with *surface adjusted distance*, evolving from distance "as the crow flies" to distance "as the horse runs". Several methods are compared for DEMs spanning a range of resolutions for the study area and validated against a 3 meter (m) lidar data benchmark. Error magnitudes vary with pixel size and with the method of surface adjustment. The rate of error increase may also vary with landscape type (terrain roughness, precipitation regimes and land settlement patterns). Cross-scale analysis for a single study area is reported here. Additional areas will be presented at the conference.

2. Data and Study Area

The study area is 7,885.94 square kilometers (sq km), located in western North Carolina, (35.798 degrees N and 81.473 degrees W) spanning the Pisgah National Forest. Its location at

the southern edge of the Appalachian mountain range is a humid, hilly landscape, averaging 51 inches (129.5 cm) annual precipitation, with elevations ranging from 209 to 1602 meters (m). Distances are measured on 10, 30, 100, 1000, and 5000m resolution DEMs and compared with the 3m lidar benchmark data. The first three were downloaded from Geospatial Data Gateway (https://gdg.sc.egov.usda.gov/). The source for 100m and 1000m DEMs was Shuttle Radar Topography Mission (SRTM) (http://dds.cr.usgs.gov/srtm/version2_1/). The 5000m DEM was resampled from 100m data, and provided courtesy of USGS. DEMs are in NAD1983 UTM Zone 17N.

3. Methods

Five straight-line transects were registered on each DEM ranging in length from 39.58 km to 107.26 km as measured on the 3m lidar data (Figure 1). Each transect was sampled at 3m intervals to create a set of test points for comparing the various distance metrics (Figure 2a).

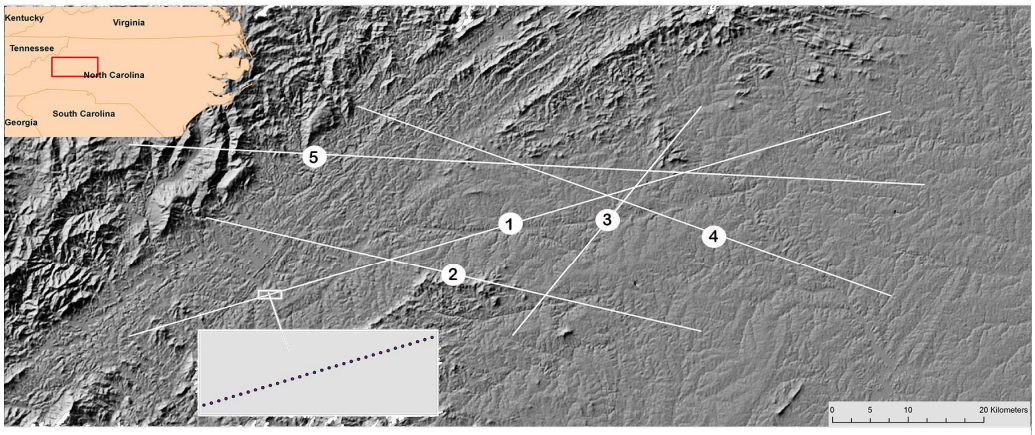


Figure 1. 3m LiDAR DEM with five transect lines overlaid. Transect lengths (in kilometers) used for validation are as follows: #1: 103.31; #2: 68.00; #3: 39.58; #4: 75.50; and #5: 107.37. Gray rectangle enlarges a section of one transect to show point samples taken at 3m spacing.

The tested methods all incorporate elevation, but differ in contextual information about surrounding pixels to gain a progression of surface adjustment. For example, Pixel-to-Pixel distance traverses sampled points in sequence along each transect. A 3D Euclidean calculation sums distances between sampled points but ignores adjacent pixels (Figure 2b). Four additional tested methods utilize elevation and spatial context. These are:

- *Closest Centroid* distance assigns the elevation at the pixel centroid to any point along the selected path that falls within that pixel (Figure 2c) and will be used to measure lengths on the lidar benchmark to validate transect distances at coarser resolutions.
- *TIN* distance is an ArcGIS® command partitions the DEM and interpolates elevations for points within each triangular facet from the three local vertices.
- *Natural Neighbor* distance (ArcGIS® command) partitions Thiessen polygons. Sampled transect points seed a second layer of Thiessen polygons. The proportion of overlap between the two layers weights interpolation of the elevations of all Thiessen neighbors.
- *Weighted Average* distance uses the average elevation of the pixel containing the point and the eight surrounding pixels, weighted by the distance to corresponding pixel centroids (Figure 2d).

Three additional methods fit local polynomials with varying degrees to progressively incorporate elevation, slope and curvature (full surface adjustment) into the distance metric. Bilinear distance fits a first order polynomial to four adjacent pixels. Biquadratic fits a second order equation to eight pixels. Bicubic distance fits a third order polynomial to sixteen surrounding pixels. Figures 2e-2g show example configurations.

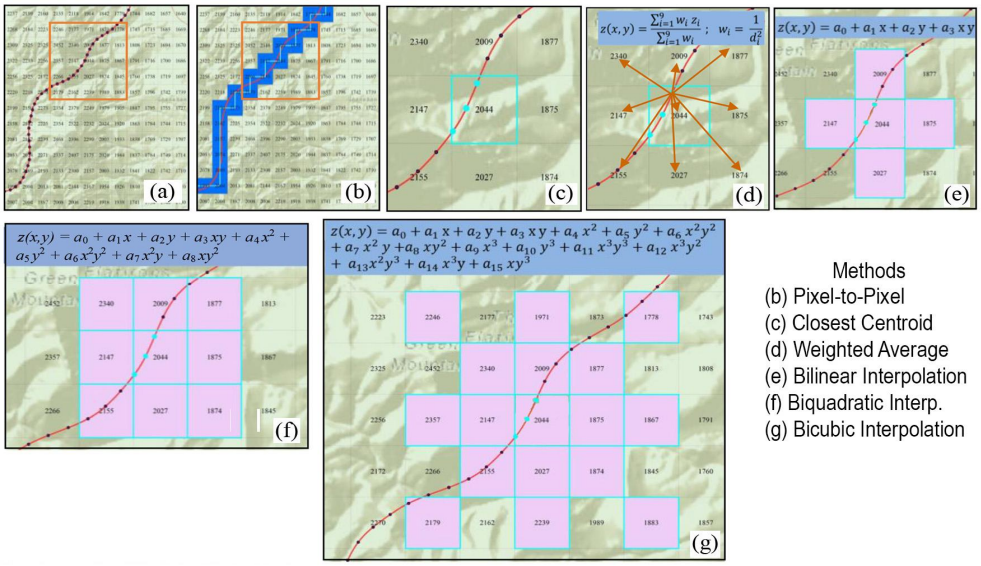
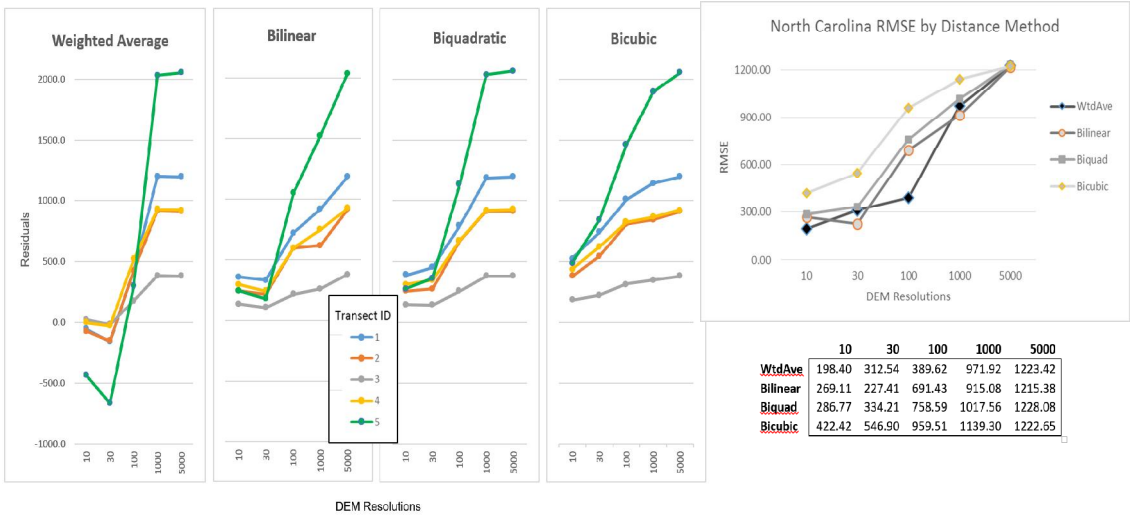


Figure 2. (a) 3m samples along transect; (b)-(g) distance computation methods. Orange vectors (d) show distance to adjacent centroids for sampled point lying at their intersection. Orange boxes (a)-(b) show areas enlarged in other panels. Elevations for magenta pixels (e)-(g) included in polynomial computations.

3. Results

All methods deviate from the benchmark, with Pixel-to-Pixel and Closest Centroid showing highest magnitude errors and consistent over-estimation for all transects at 10, 30 and 100m resolutions. Surprisingly, Arc® TIN and Arc® Natural Neighbor show nearly identical residuals for most transects across all resolutions, for reasons that are not clear. Due to space limitations, detailed distances are not reported here. Figure 3 plots absolute residuals by transect for four methods whose results are closer to the benchmark. Residuals show a general trend of increase at coarser resolutions, as additional sample points fall within a single DEM pixel. This is most pronounced for the two longest transects (#1 and #5). Residuals progress at different rates for each transect because of the varied character of terrain that each one spans. The 5th transect shows the most extreme error pattern because it crosses both rough and smooth terrain. These imply that transect length and terrain type are important factors needing further testing.



	10	30	100	1000	5000
WtdAve	198.40	312.54	389.62	971.92	1223.42
Bilinear	269.11	227.41	691.43	915.08	1215.38
Biquad	286.77	334.21	758.59	1017.56	1228.08
Bicubic	422.42	546.90	959.51	1139.30	1222.65

Figure 3. Residuals Analysis. Residual (lidar minus test DEM distance) and RMSE plots in meters.

Figure 3 also shows RMSE values for the four selected distance methods, with lowest RMSEs for either Weighted Average or Bilinear polynomial at each DEM resolution. However, the Weighted Average seems to be the least precise method of the four, furnishing a larger range of residual values, as well as mixing over- with under-estimation.

4. Discussion

Chronic mis-estimation resulting from planar distance metrics impacts models that rely on terrain for science, planning and decision support. One goal of this work is to determine if surface adjustment can obviate the need to work at lidar resolutions, given the scarcity of lidar data in rural areas and developing regions. This research demonstrates that surface adjustment generates terrain-based distance measurements that approach results for finer resolution data, with some caveats. Results depend on DEM resolution and method of surface adjustment. The choice of point sample spacing is likely scale dependent. At coarser resolutions a multitude of point samples fall within a single DEM pixel: for some distance methods (e.g., Closest Centroid) redundant samples can distort transect distances. Thinning point samples proportionate to test resolutions is an option under investigation. Resampling the 500m DEM also distorted measurements, and should be avoided for tasks requiring precise distance. Results also depend on terrain homogeneity. Transect #5 that crosses rough to smooth terrain exhibits a unique error pattern relative to those crossing more uniform terrain, and this warrants examination of heterogeneous landscapes as well as considering a larger sample of transects, which would permit statistical validation. Polynomials prove reliable, but the improvement from bicubic equations seems barely worth the added complexity when the bilinear performs similarly, and the weighted average equally well at coarse resolutions. Ongoing research explores whether findings hold true in other landscapes.

Acknowledgements

This research is supported by the Grand Challenge Initiative “Earth Lab” funded by the University of Colorado http://www.colorado.edu/grandchallenges/. We acknowledge the USGS Center for Excellence in Geospatial Information Science (CEGIS) for analytic advice.

References

Cressie NAC, 1993, *Statistics for spatial data*, Wiley, New York.

d’Oleire-Oltmanns S, Marzolff I, Tiede D, and Blaschke T, 2014, Detection of gully-affected areas by applying object-based image analysis in the region of Taroudannt, Morocco. *Remote Sensing 6(9):* 8287-8309.

Harms TK, Wentz EA and Grimm NB, 2009, Spatial heterogeneity of denitrification in semi-arid floodplains. *Ecosystems* 12(1): 129-143.

Gutiérrez M, 2012, Chapter 15.3 Applied Geomorphology in Periglacial Regions, *Geomorphology.* (Translated by Bobeck P.). Taylor & Francis: 601-613.

Kiessling F, Nefzger P, Kaintzyk U, and Nolasco JF 2003. *Overhead Power Lines: Planning Design, Construction.* Berlin: Springer Science and Business Media.

Jenny B, Jenny H, and Hurni L, 2011, Terrain generalization with multi-scale pyramids constrained by curvature. *Cartography & Geographic Information Science* 38(1): 110-116.

Passalacqua P, Belmont P, Staley DM, Simley, JD, Arrowsmith JR, Bode CA, Crosby C, DeLong SB, Glenn NF, Kelly SA, Lague D, Sangireddy H, Schaggrath K, Tarboton DG, Wasklewicz T, and Wheaton JM, 2015, Analyzing high resolution topography for advancing the understanding of mass and energy transfer through landscapes: A review. *Earth Science Reviews* 148: 174-193.

Schneider B, 2001, On the uncertainty of local shape of lines and surfaces. *Cartography & Geographic Information Science* 28(4): 237-247.

Stanislawski LV, Falgout, J and Buttenfield BP, 2015, Automated Extraction of Natural Drainage Density Patterns for the U.S. through High Performance Computing. *The Cartographic Journal* 52(2): 185-192.

Tucker GE & Hancock GR, 2010, Modeling landscape evolution. *Earth Science Processes/Landforms* 35:28-50

Usery EL, Finn MP, Cox JD, Beard T, Ruhl S, and Bearden M, 2003, Projecting Global Datasets to Achieve Equal Areas. *Cartography and Geographic Information Science* 30(1):69-79.

Using dynamic geospatial ontologies to support information extraction from big Earth observation data sets

Gilberto Camara[1], Adeline Maciel[1], Victor Maus[1], Lubia Vinhas[1], and Alber Sanchez[1]

[1] Image Processing Division, National Institute for Space Research, Av dos Astronautas 1758, Sao Jose dos Campos, 12227-001 Brazil.
Email: {gilberto.camara, adeline.maciel, victor.maus, lubia.vinhas, alber.ipia}@inpe.br

Abstract

This paper presents the spatiotemporal interval logic formalism and shows how to use it for reasoning about land use change using big Earth observation data. This formalism improves our ability to extract information from large land remote sensing data sets.

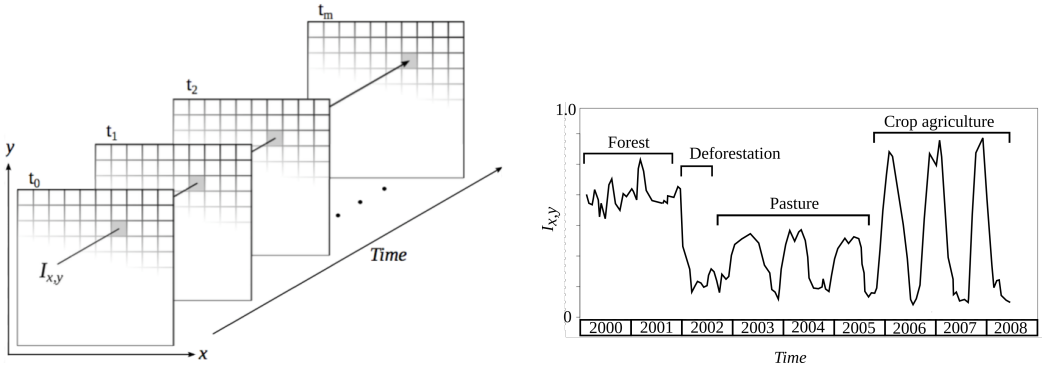
1 Events as key concepts for describing land use change

Remote sensing satellites are the only source that provides consistent data about the Earth's land and oceans. The open availability of big Earth observation data has led to an opportunity to improve information on land changes in the planet. However, most studies that use remote sensing images to detect change still adopt a *snapshot* approach. Image from a sequence are classified one by one; results are compared to account for change. There is no actual representation of the occurrences of change, but only of their effects. Two land areas with different change trajectories whose initial and final states are the same cannot be distinguished. With access to big data sets, researchers need better ways to describe and understand change. The challenge is to make best use of big Earth observation data sets to represent change.

This paper uses the concept of *'events'* from dynamic spatial ontologies to describe land use change. Events are complete entities on their respective time intervals; their lifetime is limited while objects persist in time and are complete in space (Galton and Mizoguchi, 2009; Worboys, 2005; Hacker, 1982). Since events are intrinsically related to the objects they modify, a geospatial event calculus should specify not only what happens, but also which objects are affected by such changes. We present an event calculus formalism for reasoning about land use change. The formalism is general enough to be applied in other geospatial domains.

To define events in big Earth observation data sets, multiple satellite observations of an area are mapped to 3D arrays in space-time. A pixel location (x, y) in consecutive times $t_1, ..., t_m$ makes up a satellite image time series (Figure 1a). One can extract land use change information

for each pixel, considered as an atomic 'land object'. Data mining techniques such as Time-Weighted Dynamic Time Warping (Maus et al., 2016) match temporal patterns of events to their actual occurrence in remote sensing time series (Figure 1b). The results are the temporal boundaries of events associated to a land object. For example, Figure 1b shows four major events extracted from a remote sensing time series, expressed in terms of the intervals they happen. From 2000 to 2001 the area was a forest that was deforested in 2002. From 2003 to 2005 the area it was used for pasture and from 2005 to 2008, as cropland. Since classifying all pixels in a space-time array produces a large set of events, we need an event reasoning formalism to extract information.



(a) Time series measures (e.g. EVI index) of a pixel location (x, y)

(b) Land-use/cover change events associated to a pixel location (x, y)

Figure 1: A 3-dimensional array of satellite data and events describing change at a particular location. Adapted from Maus et al. (2016).

2 The spatiotemporal interval logic

The main elements of a temporal reasoning formalism include the primitive time unit (*instants or intervals?*) and the granularity (*is time continuous or a sequence of discrete elements?*). For describing land use change trajectories from remote sensing data, we consider that an interval-based approach with discrete granularity is better than instant-based formalisms such as the Event Calculus (Kowalski and Sergot, 1989). Thus, we propose to extend Allen's interval temporal logic (Allen, 1984) to build a general framework to reason about events. Allen (1983) defines a set of mutually exclusive primitive relations between temporal intervals. Each of these is a predicate over intervals: *during*, *starts*, *finishes*, *before*, *overlap*, *meets*, and *equal*. These predicates have become widely used in many areas of computing.

In this work, we propose a spatiotemporal interval logic that includes geospatial objects explicitly. Geo-objects are intrinsically tied to space, and events change their properties. The elements of the formalism are a set of *discrete geo-objects* ($O = o_1, o_2, ..., o_n$), *discrete time intervals* ($T = t_1, t_2, ..., t_n$), and *properties of objects* ($P = p_1, p_2, ..., p_n$). Extending the ideas from Allen (1984), we introduce the predicate $holds(o, p, t) \rightarrow bool$, to denote the assertion

that the property p of geo-object o holds over interval t. We also introduce the predicate $occur(o, p, T_e) \rightarrow bool$ to denote that, given an interval $T_e \subset T$, the property p of geo-object o is true over the whole subset T_e. Some of the basic axioms of our spatiotemporal interval logic are presented in Table 1. In these axioms, we use the notation $T_e \subset T$ to denote a temporally connected proper subset of T.

Table 1: Basic axioms of spatiotemporal interval logic

Events happen over a given interval
$\forall o \in O, occur(o, p, T_e) \wedge in(T'_e, T_e) \implies \neg occur(o, p, T'_e)$ *where*
$in(T'_e, T_e) \Leftrightarrow during(T'_e, T_e) \vee starts(T'_e, T_e) \vee finishes(T'_e, T_e)$
Events do not change over an interval
$\forall o \in O, occur(o, p, T_e) \implies \forall t \in T_e, holds(o, p, t)$
Events are unique
$\forall o \in O, occur(o, p, T_e) \wedge meets(T_e, T'_e) \wedge occur(o, p', T'_e) \implies p \neq p'$

3 Reasoning about land use change

While the full development of the spatiotemporal interval logical applied to land classification is beyond the scope of this paper, we show some queries useful to reason about *land use trajectories*. Informally, a *land use trajectory* is a path from one land use state to another, for example when a forest area is converted to pasture. Formally land use trajectories are expressed as logical expressions over an event data set.

As an example, consider a study that investigates the agreement known as the Brazil's Soy Moratorium, signed by major commodity traders agreeing not to purchase soybeans grown on lands deforested after July 2006 in the Brazilian Amazonia (Gibbs et al. (2015)). Farmers abiding by the Soy Moratorium agree not to directly replace forest by soybean plantations. However, the agreement does not preclude indirect land use changes, as when a farmer buys land previously deforested that is being used as pasture. In this case, the cattle rancher may sell his land and move elsewhere, causing deforestation without violating the Soy Moratorium. Thus, we want to discover not only direct land use changes, where forest is replaced by soybeans, but also indirect land use changes. The queries in Table 2 point out how to elicit both direct and indirect land use change caused by soybeans in Amazonia.

Table 2: Using the spatiotemporal interval logic to map land use change trajectories in Brazil

Which forest areas have been replaced by soybeans?
$occur(o, \text{“}forest\text{”}, t_1) \wedge meets(t_1, t_2) \wedge occur(o, \text{“}deforestation\text{”}, t_2)$
$\wedge\, meets(t_2, t_3) \wedge occur(o, \text{“}soy\text{”}, t_3)$
Which forest areas have been replaced by pasture and later turned into soybean?
$occur(o, \text{“}forest\text{”}, t_1) \wedge meets(t_1, t_2) \wedge occur(o, \text{“}deforestation\text{”}, t_2) \wedge meets(t_2, t_3)$
$\wedge\, occur(o, \text{“}pasture\text{”}, t_3) \wedge meets(t_3, t_4) \wedge occur(o, \text{“}soy\text{”}, t_4)$

4 Conclusions

This paper presents the spatiotemporal interval logic, which is a spatial extensions of the temporal interval logic proposed by Allen (1984). The formalism considers the nature of events detectable using Earth observation data, which are discrete transitions where one land cover type is replaced by another. The proposed logic allows reasoning about land use trajectories in regional and global areas. To be useful, this formalism needs to be supported by efficient data mining techniques, capable of extracting event data sets from big data. When such event data be available, the spatiotemporal interval logic improves information extraction from large remote sensing data sets.

Acknowledgements

The authors are supported by FAPESP e-science program (grants 2014-08398-6 and 2016-03397-7). Gilberto Camara is also supported by CNPq (grant 3121512014-4).

References

J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

J. F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23(2):123–154, 1984.

A. Galton and R. Mizoguchi. The water falls but the waterfall does not fall: New perspectives on objects , processes and events. *Applied Ontology*, 4:71–107, 2009.

H. K. Gibbs, L. Rausch, J. Munger, et al. Brazil's soy moratorium. *Science*, 347(6220):377–378, 2015.

P. Hacker. Events and Objects in Space and Time. *Mind*, XCI(361):1–19, 1982.

R. Kowalski and M. Sergot. A logic-based calculus of events. In *Foundations of knowledge based management*, pages 23–55. Springer, 1989.

V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. de Queiroz. A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP(99):1–11, 2016.

M. Worboys. Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19:1–28, 2005.

Comparing digital traces of modern travellers to journeys of two 18th-19th century British poets

O. Chesnokova[1], I. N. Gregory[2], R. S. Purves[1]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland
Email: {olga.chesnokova; ross.purves}@geo.uzh.ch

[2]Department of History, Lancaster University, Lancaster LA1 4YD, UK
Email: i.gregory@lancaster.ac.uk

Abstract

The growth of interest in georeferenced media brings with it a wealth of possibilities for exploring digital media, and importantly, for advancing research concerned with domain driven research questions. In this paper, we compare previous work, which extracted and analysed the spatial traces of two 18th-19th century poets in the English Lake District with modern data digital traces in the form of Flickr images. We explore the semantics of the modern day data through use of Latent Dirichlet Allocation, and analyse the extent to which modern day tourism mimics (or indeed follows) the foundations laid by Samuel Coleridge and Thomas Gray. Our results show that tourists, just like the poets, describe mountain landscapes from below and visit popular locations commonly perceived as beautiful.

1. Introduction

The attractiveness of the English Lake District as a destination, both historically and in the present, is reflected by a wealth of poetry, travel diaries, guidebooks, and other forms of contemporary and historical written texts. Cooper and Gregory (2011) recognised the richness of such sources and performed an analysis of texts by two 18th-19th century British poets, Samuel Coleridge and Thomas Gray. In their analysis they explored not only how and what was described, but also the locations visited by the two writers.

Modern day visitors to the Lake District also leave traces, not always in the elegant prose of their forebears, but often through photographs and associated descriptions shared on image hosting web sites such as Flickr or Panoramio, often linked to specific locations through coordinates (c.f. Girardin et al. 2008). These images allow us to explore patterns of landscape description and semantics, and in this paper we explore two questions, and link these to Gray and Coleridge's historical exploration of the Lake District:

1: Do images and metadata allow similar semantic and geometric analysis to that previously carried out by Cooper and Gregory (2011)?

2: Do the spatial traces of modern travellers resemble those of Gray or Coleridge?

2. Methods

2.1 Data preprocessing

Using the Flickr API[1], we retrieved 135649 photographs and their metadata (user, coordinates, time taken and tags) taken by 6855 users in the Lake District. We removed photographs uploaded by single-posters and the most prolific contributors as well as tags with a coefficient of variation greater than 500% (Purves et al. 2011). Furthermore, we also removed tags containing toponyms or camera parameters.

[1] https://www.flickr.com/services/api/

2.2 Spatial and semantic analysis

The connection between highly-photographed places (the number of unique users taking a picture at each location) and the attractiveness of places is used in several studies (e.g., Casalegno et al. 2013; Girardin et al. 2008). One way to explore the reasons for this attractiveness is to analyse the underlying semantics captured by tags. We performed an unsupervised classification of tags using Latent Dirichlet Allocation (LDA) (Blei, Ng and Jordan 2003) where a user-defined number of classes from the text is extracted. Each Flickr tag is given a probability of belonging to a class. In this way, the similarities of regions and their typical characteristics can be described.

Several authors have proposed stratifying photographs into those taken by tourists or locals, often using the length of time a user is found in a region. If this time is short, the user is classified as a tourist (Girardin et al. 2008; Straumann et al. 2014). Here we use 30 days to separate tourists from locals.

3. Results

In Figure 1, we compare the distribution of photographs taken by all users (a), tourists (b) and the places mentioned in Gray (c) and Coleridge's accounts (d). Pictures taken by tourists and locals are concentred around Keswick, Grasmere and Windermere, whereas the regions to the north and east of the Lake District are the least photographed.

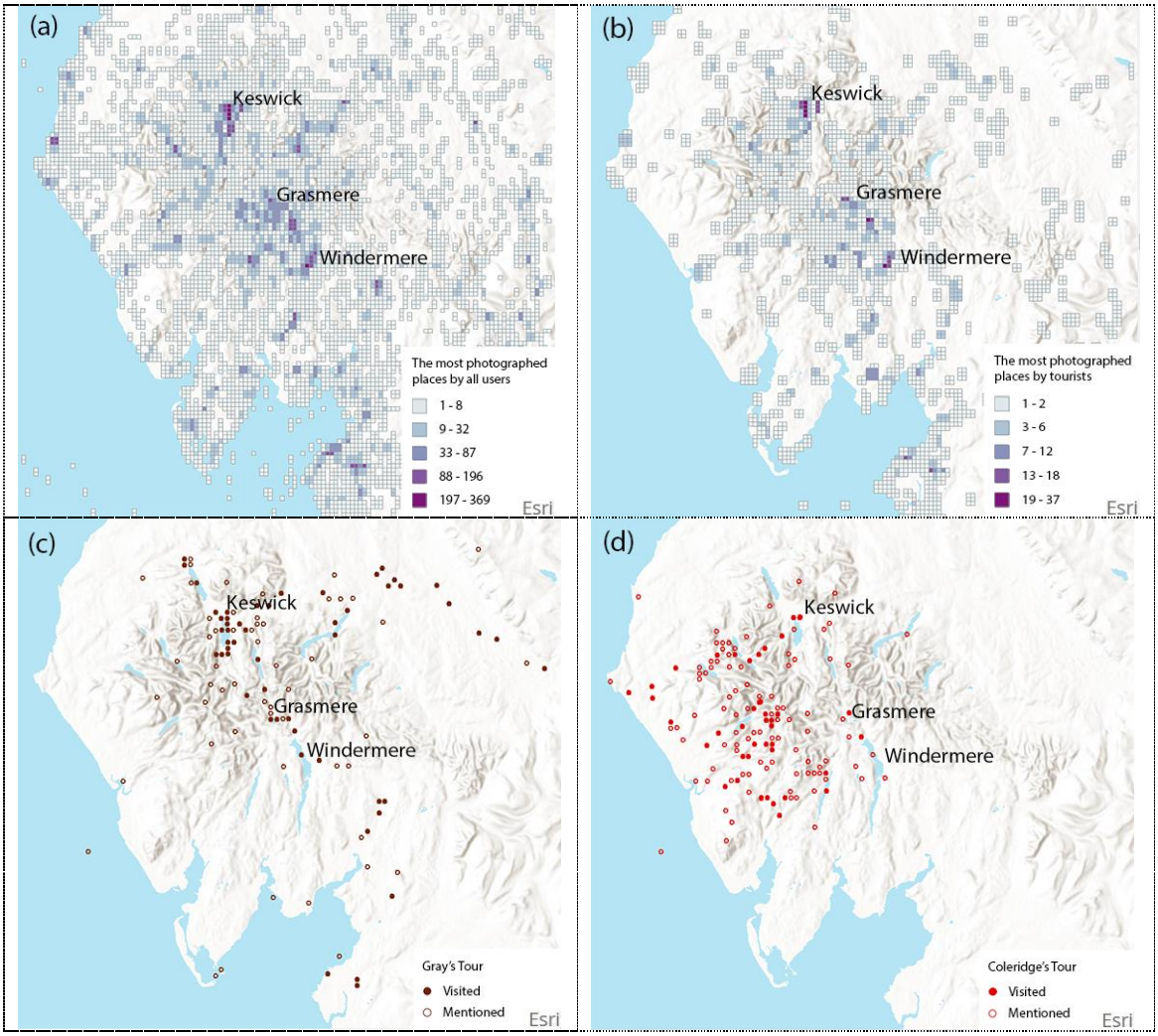


Figure 1. A comparison of the most photographed places by all users and by tourists and the spatial patterns of Gray's and Coleridge's travels.

58% of the photos of today's tourists were taken below 100 m and 92% were taken below 300 m. To draw a comparison with the journeys of Gray and Coleridge, we analysed the altitude associated with descriptions of upland landscapes. To do so we extracted images whose tags described either toponyms found above 600 m or terms such as 'summit,' 'fell,' and 'mountain'. We found, analogously to Cooper and Gregory (2011), that modern day tourists describe the uplands from below, with 21% of upland related descriptions captured below 100 m and 68% below 300 m.

Gray has been described as a "Picturesque" traveller, visiting popular locations commonly perceived as beautiful, while Coleridge was characterised as an "environmental insider and post-Picturesque traveller" (Cooper and Gregory 2011, p. 94) who thus took less well travelled ways. By analysing the number of unique users found within 2 km of places visited by Gray and Coleridge we find that this Picturesque view (as propagated by Gray) of the Lake District continues to be preferred by modern tourists (χ^2 test, $p < 0.05$).

Using LDA, we extracted, and labelled, 10 classes of landscapes from Flickr tags (Table 1). We hypothesised that these classes might reveal relationships between landscapes and observers. For instance, two quite different regions to the north and south are linked through the 'animal class', and turn out to be the locations of wildlife parks. A third location in Penrith features in this class, and a park where red squirrels are commonly photographed is the deciding factor for its membership. Similarly, on lake shores photos of seabirds are influential in linking these locations to this class (Figure 2).

Table 1. Ten classes of landscapes by the LDA unsupervised classification.

Indicative names	The most probable words per class (ranked by probability)
"railway"	railway, boats, train, station, viaduct, steam, hotel, harbour, trains, railways
"animals"	nature, birds, wildlife, north, animal, landscapes, old, spring, animals, northwest
"winter"	winter, snow, rocks, coast, stone, stones, settle, ice, circle, midlandhotel
"holidays"	holiday, sea, beach, sun, architecture, sand, classic, sculpture
"view"	clouds, sunset, countryside, mountains, cloud, jetty, view, scenery, valley, tree
"hiking"	mountains, autumn, walking, hiking, colour, rain, camping, farm, leaves, coasttocoast
"summer"	boat, tree, summer, wedding, sunrise, family, pub, places, horse
"nature"	mountain, river, bridge, reflection, stream, village, path, swan, evening, duck
?	grass, sheep, light, reflections, panorama, black, wall, night, field, building
?	walk, seaside, bus, dog, stagecoach, cold, music, long, dusk, frost

The class defined by the tags 'railway,' 'train' and 'station' overlaps with the railway. Different seasons are shown in four colours (Figure 2), but we found no major temporal differences.

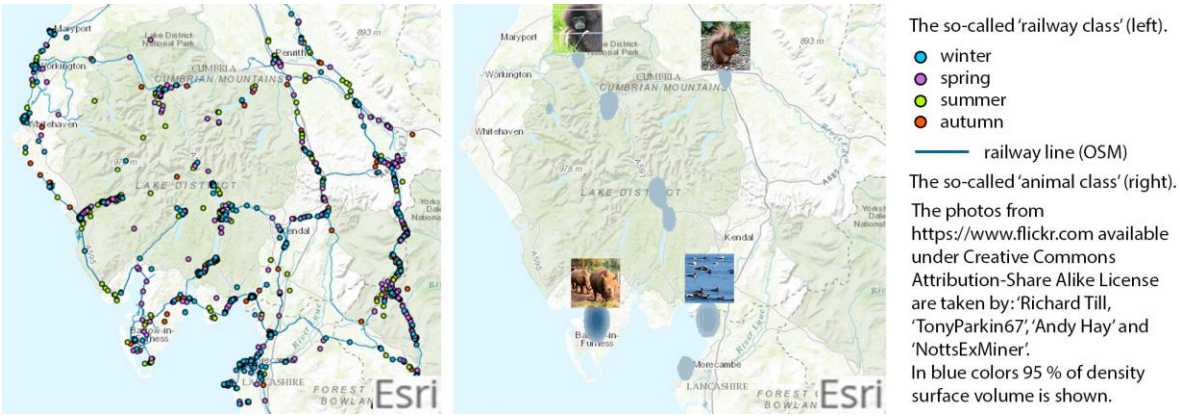


Figure 2. Spatial distribution of results.

4. Concluding discussion

Our motivation was to demonstrate the extent to which it is possible to use Flickr data for the extraction of modern travelling spatial and semantic patterns.

Firstly, our analysis shows that today's tourists have similar travelling patterns to the poets of the 18th-19th century, especially with respect to the Picturesque travelling style of Gray. The villages of Keswick and Grasmere are still considered most attractive and regions in the north remain little visited. Furthermore, the habit of describing the uplands from a low-lying vantage point persists to the present day.

Secondly, we observe that nouns in particular have a high probability in each class (c.f. Purves et al. 2011). This means that we can extract some unexpected patterns as observed with our 'animal class,' but we are not able to define the adjectives that describe the classes (c.f. Cooper and Gregory 2011). An additional difficulty lies in assigning labels to classes, as indicated by question marks in Table 1.

Our results demonstrate that, even with the limited semantic depth available in Flickr images, it is possible to compare not only travel patterns, but also the ways in which landscapes are described, with historical sources. In future work we will extend our approach using richer, natural language, contemporary sources such as travel blogs and guidebooks.

Acknowledgements

Ian Gregory's contribution was funded by the ERC under grant agreement number 283850.

References

Blei DM, Ng AY and Jordan MI, 2003, Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3(1):993–1022. doi:10.1162/jmlr.2003.3.4-5.993.

Casalegno S, Inger R, DeSilvey C and Gaston KJ, 2013, Spatial Covariance between Aesthetic Value & Other Ecosystem Services. *PLOS ONE* 8(6). http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0068437.

Cooper D and Gregory IN, 2011, Mapping the English Lake District: A Literary GIS. *Transactions of the Institute of British Geographers* 36:89–108. doi:10.1111/j.1475-5661.2010.00405.x.

Girardin F, Blat J, Calabrese F, Dal FF and Ratti C, 2008, Digital Footprinting: Uncovering Tourists with User-Generated Content. *IEEE Pervasive Computing* 7(4):36–44. doi:10.1109/MPRV.2008.71.

Purves RS, Edwardes AJ and Wood J, 2011, Describing Place through User Generated Content. *First Monday. Peer-Reviewed Journal on the Internet* 16(9). http://firstmonday.org/ojs/index.php/fm/article/view/3710/3035.

Straumann RK, Çöltekin A and Andrienko G, 2014, Towards (Re)Constructing Narratives from Georeferenced Photographs through Visual Analytics. *The Cartographic Journal* 51(2):152–165. doi:10.1179/1743277414Y.0000000079.

A moan, a discursion into the visualisation of very large spatial data and some rubrics for identifying big questions

A. Comber[1], C. F. Brunsdon[2], M. Charlton, R. Harris[3]

[1] School of Geography, University of Leeds, LS2 9JT, UK
Email: a.comber@leeds.ac.uk

[2] NUI Maynooth, Maynooth, Co Kildare, Ireland
Email: {christopher.brunsdon; martin.charlton} @nuim.ie

[3] School of Geographical Sciences, University of Bristol, BS8 1SS.
Email: rich.harris@bris.ac.uk

Abstract

This short paper links 2 areas of big data science in the context of GIScience: inferential analysis and visualisation. It discusses ideas around integration and analysis of large, spatial referenced datasets and considers how results of these can best be visualised. It advocates a critical approach to big data visualization and warns of the inherent dangers of simply identifying patterns, whether through data mining, modeling or visualization. It adds to ongoing debates by suggesting techniques and rubrics, possibly even hinting at a manifesto.

1. Introduction

There is an increasing amount of data of all kinds available to scientists that provide opportunities to gain novel insights about all kinds of phenomena. The availability of these data is being driven by 2 factors. 1) The large amount of open data and wider recognition of the value that can be added to that data (Molloy, 2011) by linking it to other data and by developing novel data analyses; 2) The many new forms of data generated every day by citizens, either passively or actively (See *et al.*, 2016) on GPS- and web-enabled tablets, devices has resulted in an explosion of citizen contributed, crowdsourced or volunteered data.

Much has been written about the characteristics of these very large datasets: from the 3 or 5 or is it 7Vs? to the 3Ds (Dynamic, Diverse, Dense to which Dirty should be added), and perhaps more interestingly, their existence has stimulated a number of theoretical and practical considerations. This ranges from the need to revisit classic measures of and tools for statistical inference (Brunsdon, in press) to the need to redesign some of the more commonly used software tools to handle the data volumes. The role of GIScience relates to location, which may be precise in the form of latitude and longitude or approximate for example using a small census area reference or a post-code. However, despite being in this age of so called 'Big Data' the real challenge is to identify and answer 'Big Questions' which so far the research community, including the GIScience community, has failed to do.

2. Large quantities of spatially referenced data

There are large quantities of spatially referenced data of many different types, describing many different phenomena. This provides opportunities for new forms of knowledge. A typical data-mining / computer science approach is encapsulated by the following quote: '*Scouring databases and other data stores for insight is often compared to the proverbial search for a needle in a haystack, but … big data turns that idea on its head*' and quoting Viktor Mayer-Schönberger '*With big data, we don't know what the needle is. We can let the data speak and use it to generate really intriguing questions*'[1].GIScience offers suites of

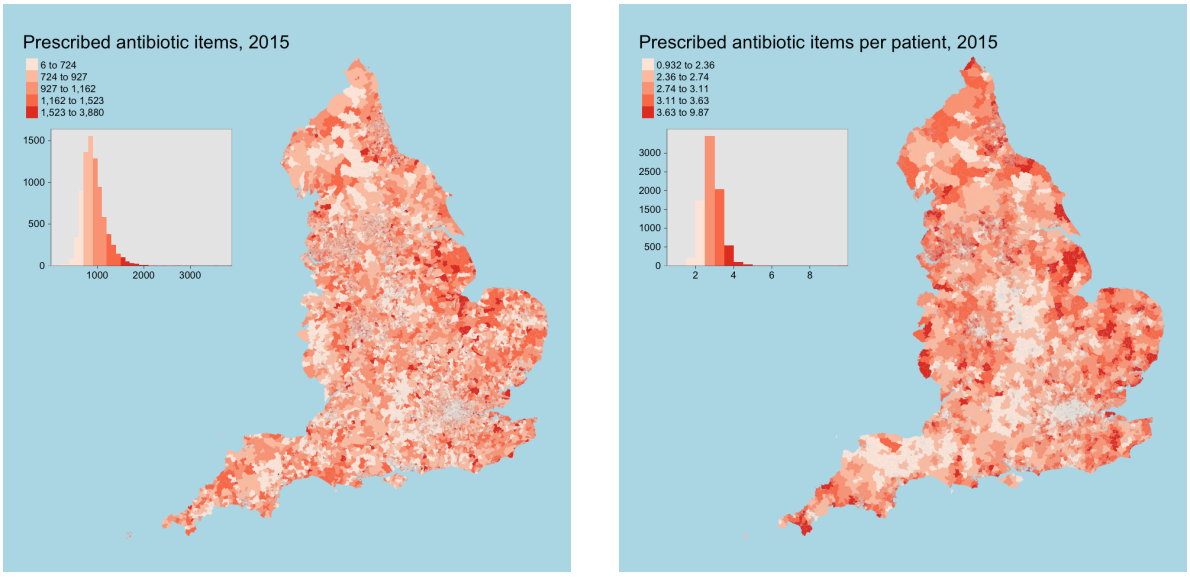
[1] http://data-informed.com/big-datas-value-much-larger-than-specific-business-questions/

methods and techniques for integrating such data over defined geographic areas (van der Zee and Scholten, 2014), for example by summing the counts of some phenomenon over a census areas and for analysis. Consider for example, the medical prescription data provided by the UK government for England (Figure 1). It records individual prescriptions, the prescribing GP practice or hospital, the cost and month of prescribing, with 10.1m records for January 2015. The GP practice postcode is provided separately and can be used to locate the data.

```
>
> head(data.1)
  SHA PCT PRACTICE        BNF_CODE                                BNF_NAME ITEMS   NIC ACT_COST QUANTITY PERIOD
1 Q44 RXA    N81646 0102000N0AAABAB Hyoscine Butylbrom_Tab 10mg              1  1.13     1.16       21 201501
2 Q44 RXA    N81646 0401010Z0AAAAAA Zopiclone_Tab 7.5mg                     15  3.28     4.72       57 201501
3 Q44 RXA    N81646 0401020K0AAAHAH Diazepam_Tab 2mg                        35 99.91   104.64     2662 201501
4 Q44 RXA    N81646 040201060AAALAL Olanzapine_Tab 15mg                      3  1.29     1.53       21 201501
5 Q44 RXA    N81646 0403010B0AAAHAH Amitriptyline HCl_Tab 25mg              16  1.52     3.20       38 201501
6 Q44 RXA    N81646 0403010B0AAAIAI Amitriptyline HCl_Tab 50mg              44  4.04     8.68      101 201501
>
```

Figure 1. Example of prescription level data.

Links to other information are needed to extract added value. Consider the aim identify factors related to Antibiotic resistance (ABR) from prescriptions. One approach is to using data mining to identify possible relationships between antibiotic prescribing patterns with a plethora of socio-economic factors. Another is to consider specific factors, for example, related to ill-health, deprivation, old age etc, on the basis that these social groups may have higher anti-biotic prescribing rates and therefore may be more likely to exhibit ABR at some point in the future. To illustrate this, prescribing data for 2015 was linked to census data for the 32,844 lower super output areas (LSOAs) in England plus 7 random created variables (r1 to r7). Figure 2 shows antibiotic prescribing rates items per patient for LSOAs in England.



a) b)

Figure 2. a) Antibiotic prescribing rates, and b) a cartogram of the same for LSOAs

When a large number of socio-economic attributes are used to construct a predictive model of prescribing rates (items per census area), many factors may be significant and with a reasonable model fit (Figure 3a). Figure 3b maps the LSOA outliers. This raises 2 problems. First, the potential meaninglessness of the variables identified as significant although in this case we could try to explain many of them. But the point is that this would be a *post hoc* rationalization. It is easier to tell convincing stories based on what is found whilst

oblivious to the potential sensitivity of the results to the methods by which the data have been compiled and collected. The traditional modelling framework doesn't work well for big data but without that framework what are we left with? How do we establish what is or isn't credible? Do we develop routines to examine specific groups of data in order cluster areas? A geo-asthenic classification on the combinations of drugs prescribed, for example? PCA could help here as potentially useful as inputs to PAM (Partitioning Around Medoids) or k-means.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.131e+01  7.376e+00   4.245 2.20e-05 ***
r1               -1.711e-04  7.882e-05  -2.171 0.029924 *
r2               -1.681e-04  7.881e-05  -2.133 0.032951 *
r3               -8.712e-05  7.882e-05  -1.105 0.269056
r4               -2.665e-05  7.882e-05  -0.338 0.735321
r5               -1.073e-05  7.882e-05  -0.136 0.891678
r6                3.541e-05  7.885e-05   0.449 0.653318
r7               -1.417e-05  7.884e-05  -0.180 0.857377
LTSickDisab      -9.079e-01  9.736e-02  -9.326  < 2e-16 ***
Stud              2.219e-01  3.137e-02   7.074 1.54e-12 ***
Inactive          9.933e-01  5.584e-02  17.788  < 2e-16 ***
InactiveOther    -1.300e+00  6.845e-02 -18.992  < 2e-16 ***
Pop               1.597e-01  1.531e-02  10.431  < 2e-16 ***
LL                5.586e-02  3.203e-02   1.744 0.081173 .
BadHealth        -7.740e-02  6.329e-02  -1.223 0.221360
EconPop          -9.556e-02  1.782e-02  -5.363 8.26e-08 ***
Unempl            1.930e-01  5.860e-02   3.293 0.000994 ***
Retired          -3.513e-01  6.239e-02  -5.631 1.80e-08 ***
StudInactive     -1.445e+00  5.774e-02 -25.024  < 2e-16 ***
IMD               5.327e+00  5.128e-01  10.389  < 2e-16 ***
Income           -5.289e+02  6.944e+01  -7.616 2.68e-14 ***
Employment        1.802e+01  6.827e+01   0.264 0.791798
Education         9.718e-01  1.095e-01   8.872  < 2e-16 ***
HealthDisability  7.094e+01  2.215e+00  32.028  < 2e-16 ***
Crime            -2.515e+01  1.674e+00 -15.019  < 2e-16 ***
BarriersHousing  -1.635e+00  1.177e-01 -13.890  < 2e-16 ***
LivingEnv        -1.588e+00  8.796e-02 -18.052  < 2e-16 ***
IncomeChildren   -1.010e+02  2.288e+01  -4.416 1.01e-05 ***
IncomeElderly    -3.261e+02  1.734e+01 -18.809  < 2e-16 ***
patients          3.755e-01  3.865e-03  97.150  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 135.4 on 32814 degrees of freedom
Multiple R-squared:  0.6592,    Adjusted R-squared:  0.6589
F-statistic:  2188 on 29 and 32814 DF,  p-value: < 2.2e-16
```

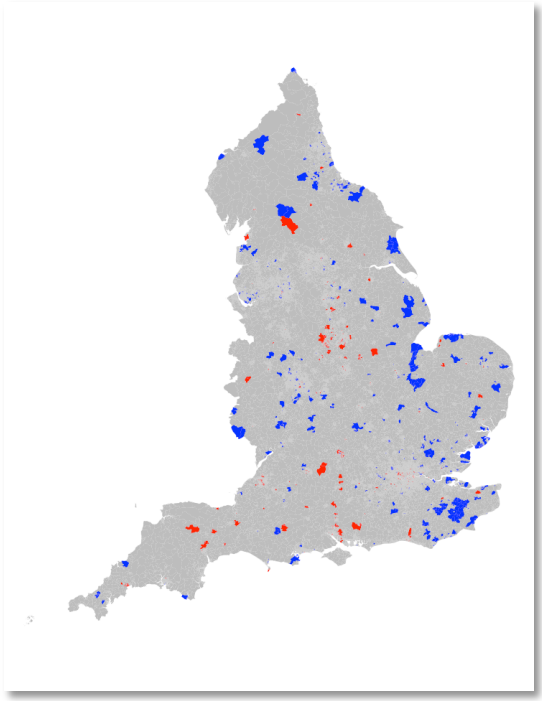


Figure 3. A model of prescribing rates (coefficient estimates) and a map of outliers

Second, the difficulty in identifying spatial patterns or trends using standard mapping approaches, so oft cited as being the panacea for big spatial data integration (Keim *et al.*, 2013). The problem is that LSOAs with larger areas dominate visually. Cartograms could be used to rescale the LSOA areas proportional to the LOSA coefficient estimate for long term sick and disabled (‘LTSickDisab’) applied to the value for each LSOA (Figure 4a). But it is difficult to interpret them because the LSOAs themselves are unfamiliar. This is in contrast to the classic cartograms of GDP developed in the 1970s and 1980s to highlight income inequalities between developed and less developed countries. So what are the options here? One might be to use regular structures such as hexbins to pool the data as in Figures 4b but to continue to undertake the analysis at the finer resolution and then to report the results using the new structures. Another approach being investigated by the authors is to transform the cartogram data but to maintain the topology of the original structures.

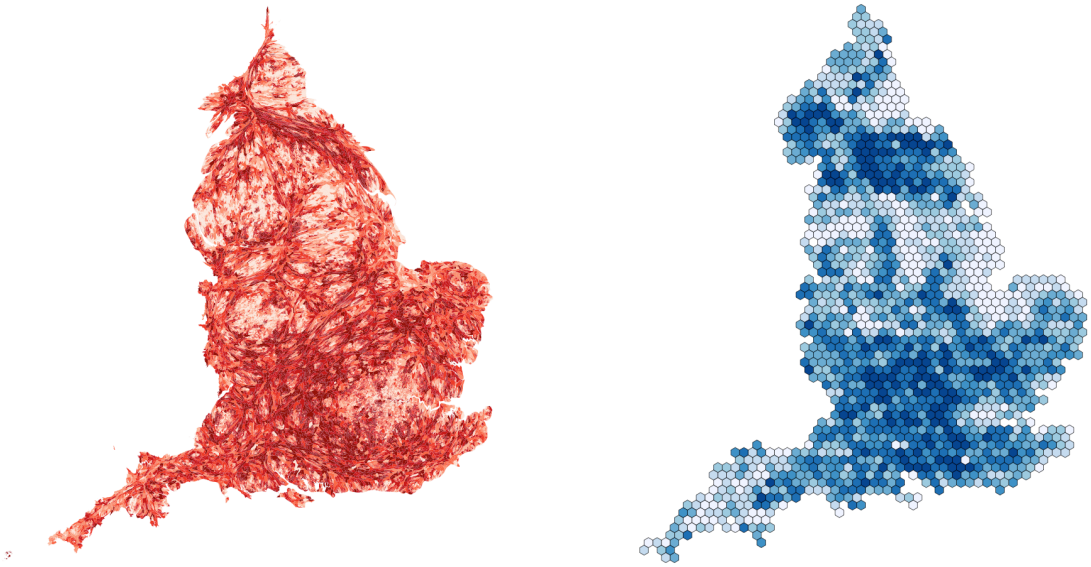
3. Rubrics

The map is a very powerful and familiar communication tool. We use them or any other visualisation to convey some information and to do that often involves also providing some critical context to the information to give salience. Context in mapping often seeks to provide a feeling of place using information that is familiar and therefore interpretable by the target audience. Maps that do this are persuasive and mapping geographical data will reveal geographical differences. However, it has long been understood in statistics that differences between values, between places are not, in themselves, evidence of anything particularly

important. Frequentist statistical testing (e.g. the t-test) puts the emphasis on statistical significance, irrelevant in the context of big data where everything will be significant due to sample size. So perhaps we need to consider how we meaningfully visualise (map) something that is meaningful. Do we for example consider dynamic linked graphics (Wood *et al.*, 2011) with immersive displays (Turkay *et al.*, 2014) to visually explore the data dimensionality?

What is needed is a balance between effective visualisation and acknowledgment that that visualisation becomes part of the problem if it legitimises dubious data and dubious research practices. This suggests the need for critical approaches big data visualisation. We need better protocols for visualisation and analysis but also to recognise that traditional visualisations may obscure detail that is important locally, although perhaps not globally, in similar way to the need for novel inferential approaches to replace those that assume small samples and situations where nil and null hypothesis testing are not relevant, such as those under big data.

What matters is whether we are learning anything particularly useful through analysis and visualisation of big data. This requires consideration of how and why the data are generated. Perhaps a test of result credibility is needed, for example by analysing random samples of the big data to see if the same results are generated. Or testing against a geographical shuffling of the data. Are similar spatial patterns observed (albeit in different places)? It is not good enough to simply take a data set, discover something and then to map and publish the results. This only serves to increase publication bias and if we only ever publish things that purport to be surprising but are actually due to luck and/or dodgy data then we will continue to publish the exception rather than the rule, which will not benefit wider society.



a) b)

Figure 4. a) shaded cartogram of showing areas where antibiotic prescribing is related to long term sick and disabled, and b) aggregated over 10km hexbins

Spatial big open data undoubtedly has the capacity to support better science and to benefit of society through increased transparency, reproducibility, efficiency etc (Molloy, 2011) and protocols to knit data together have been suggested (e.g. van der Zee and Scholten, 2014). Some have suggested a role for digital geography/GIScience not only in integrating and analysing large amounts of spatial data, but also in the engagement of the '*hard work of theory*' (Pickles, 1997, p. 370) to wrestle information control away from a very small group of corporate entities (Thatcher, 2014). However, others have noted the gaps and myths that exist between the rhetoric of big spatial data analyses and the reality (Janssen *et al.* 2012). Of

course, experimental design is also important and inferential statistics originally considered this as a part of the process of discovery, in the following way:

1. Formulate a research question.
2. Identify what data to collect and how to collect it.
3. Perform some statistical tests to determine whether any effects/associations are unlikely to have occurred by chance.
4. Get an answer to the question.

The big data paradigm turns experimental design / inferential theory on its head (the minus signs are intentional):

-1. Collect lots of data about anything.
-2. Perform some kind of data mining.
-3. Get some kind of answer.
-4. Decide what question it was an answer to.

The danger is that some researchers advocate omitting stage -4. This is problematic: if the research question is not specified then the results are answers to arbitrary questions. The means that instead of finding an answer to 'does drug X reverse type 2 diabetes' you may get an answer to a completely irrelevant question (e.g. 'do grey cats fart more often than ginger ones?') and frequently lacking an inferential dimension (information on cats farting might be spurious). We probably do need to exploit big data, but perhaps we should develop better protocols for their analysis. That is we should have some idea of what questions are important and have a better way of validating any findings. Considered, representatively robust visualisations (i.e. that communicate space and value, locally and globally) may help in this.

Until we act in this way then we cannot know whether we should expect the Big Questions to be deep in the Big Data? Or whether playing with Big Data will help us to answer Big Questions we currently have (and regardless of the answer, the MAUP does not go away). Our final observation in relation big data is that if the aim is to find a needle in a haystack then making the haystack bigger does not make the job any easier. If we don't know what kind of needle that we are looking for it helps even less. Critiques of question and context free spatial data analysis, so called *databating*, persist.

Acknowledgements

We acknowledge the stimulation of the IGU Applied Geography workshop on *the Application of Big data in Geography*, Rhodes 7-10th May 2016 that led to extensive discussions of these ideas with the bones being put together at 34,000 feet on EZ8790.

References

Brunsdon, C (in press). Quantitative methods II: Issues of inference in quantitative human geography. *Progress in Human Geography*

Janssen M, Charalabidis Y & Zuiderwijk, A (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258-268

Keim, D, Qu, H, & Ma, KL (2013). Big-data visualization. *Computer Graphics &Applications*, *33*(4): 20-21

Molloy JC (2011). The open knowledge foundation: open data means better science. *PLoS Biol*, *9*(12) e1001195

Pickles J (1997). Tool or science? GIS, technoscience and the theoretical turn. *Annals of the AAG*, 87: 363-372

See L et al (2016). Crowdsourcing, Citizen Science or Volunteered Geographic Information? The Current State of Crowdsourced Geographic Information *ISPRS International Journal of Geo-Information*, 5: 55

Thatcher J (2014). Big Data, Big Questions| Living on Fumes: Digital Footprints, Data Fumes, and the Limitations of Spatial Big Data. *International Journal of Communication*, *8*:19

van der Zee E, & Scholten H (2014). Spatial dimensions of big data: Application of geographical concepts and spatial technology to the internet of things. In *Big Data and Internet of Things* (pp. 137-168). Springer.

Turkay C, Slingsby A, Hauser H, Wood J & Dykes J (2014). Attribute signatures: Dynamic visual summaries for analyzing multivariate geographical data. *IEEE Trans on Vis and Computer Graphics*, 20: 2033-2042

Wood J et al (2011). Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica*, 46: 239-251

Local variation in hedonic house price, Hanoi: a spatial analysis of SQTO theory

A. J. Comber[1], P. Harris[2], T. Q. Nguyen[3], P. K. Chi[4], H. Tran[4] and H. H. Phe[5]

[1] School of Geography, University of Leeds, LS2 9JT, UK
Email: a.comber@leeds.ac.uk

[2] Rothamsted Research, North Wyke, Okehampton, EX20 2SB, UK
Email: paul.harris@rothamsted.ac.uk

[3] National University of Civil Engineering, Hanoi, Vietnam
Email: quannt@nuce.edu.vn

[4] GeoViet Consulting Co. Ltd, Hanoi, Vietnam
Email: chi.geoviet@gmail.com

[5] Vinaconex R&D, Pham Hung St, Hanoi City, Vietnam
Email: hoanghuuphe@googlemail.com

Abstract

This paper applies a local analysis to model and predict hedonic house price in Hanoi, Vietnam. It applies a locally compensated geographically weighted ridge regression to data survey data collected to support the Status Quality Trade Off theory proposed by Phe and Wakely (2000). This has an inherently local flavour is therefore suitable for local statistical approaches such as GWR (Brunsdon *et al*., 1996). The locally compensated ridge regression accounts for the observed local collinearity. The results provide a spatially nuanced model of status poles associated with areas of desirable housing. Some key areas for future work are suggested.

1. Introduction

The SQTO (Status Quality Trade Off) theory (Phe and Wakely, 2000) seeks to explain urban development dynamics and links the temporally dynamic process of urban transformation (*housing status*) with the relatively permanent character of the physical environment (*dwelling quality*). In doing so, SQTO models changes the relationship between residential location choices, for example the social desirability of a particular area for different social groups, and housing physical characteristics such as floor area, number of bathrooms, number of storeys, etc.

SQTO has at its core the idea that desirable residential location patterns are explained by polar structures, where one or several *status poles* represent the highest points of certain kinds of social status held by the residential population relating to particular notions of wealth, political power, business, culture, ethnicity, education specific to particular social or geo-demographic groups. It is underpinned by 2 considerations: that status poles are local and that they vary for different social groups and as such SQTO theory has at its core the essence of Tobler's 1st law of Geography (Tobler, 1970), suggesting the suitability of explicitly local frameworks of analysis such geographically weighted (GW) approaches (Brunsdon *et al.*, 1996). This is in contrast to mainstream economic residential location theories that explain residential location patterns using classic utility maximization theories and budget constraints (Alonso, 1964; Fujita, 1989). In these, access to place of work to minimise commuting costs, for example, is one of the main factors explaining residential location choices for lower-income households and higher-income households choose larger properties in the suburbs

with choices moderated by factors such as school quality (Kim *et al.*, 2005), crime (Weisbrod *et al.*, 1980) and access to other amenities (Parkes *et al.*, 2002).

This research 1) applies a geographically weighted regression to model local variations factors predicting hedonic house price in Hanoi, and 2) applies a locally-compensated ridge regression to create a spatially distributed model of house prices in Hanoi.

2. Methods

A detailed house survey of nearly 1,000 households in Hanoi was undertaken in 2014 collecting data for variables, reflecting both physical (tangible) and social (intangible) factors related to housing choices. In this study 13 variables were selected as input to analyses as listed in Table 1.. The data were cleaned to remove NULL variables resulting in 633 data points whose locations are shown in Figure 1. The dependent variable in this case is *HPRICVND* - the house price in millions of Vietnamese Dong. The spatial distribution of this data is reasonable, providing good coverage of the study area.

Table 1. The SQTO variables. Type = 1 indicates a tangible variable related to dwelling quality, Type = 2 indicates an intangible variable related to housing status.

Variable	Description	Type
HPRICVND	Price of house in Millions of Vietnamese Dong	Target
AIRCON	Air-Conditioner (Yes, No)	1
GFA	Total floor area (incl. mezzanine) (m^2)	1
PLOTAREA	Total plot area (m^2)	1
SHOPFRNT	Shop Front (Yes, No)	2
PLUMBING	Plumbing Quality (Good, Other)	1
HOUSEGRADE	Permanent, Other	1
CAR	Car ownership (Yes, No)	2
CENTDISR	Measured distance to Centre District	2
DISCENDI	Perceived travel time to the Centre District	2
EDYEARS	Time in education of the interviewee (years)	2
OCCUP_PRIVBIZ	Private Business owner (1=Yes, 0=No)	2
SCHOOQLT	School Quality (Good, Other)	2
STRTYPE	Type of street (Business, Residential)	2

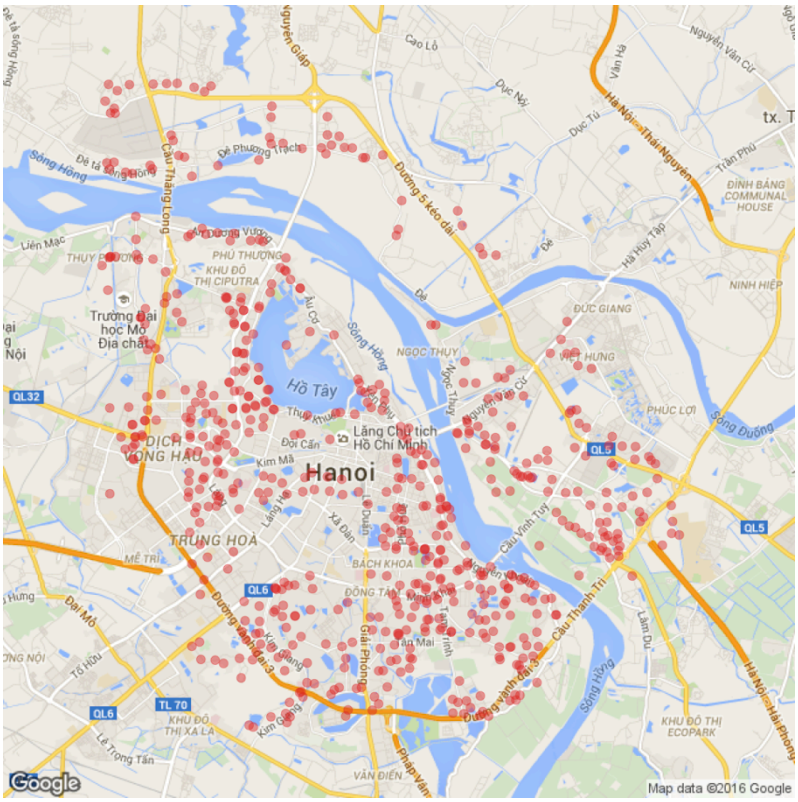


Figure 1. The study area and the 633 survey data points, shaded with a small transparency term and a Google Maps background.

Geographically weighted approaches uses a moving kernel and data under the kernel are used to make a *local* calculation of some kind, such as a regression. The data are weighted by their distance to the kernel centre and local variable collinearity was tested for. The analysis consisted of the following stages:

1. Collation of the optimum bandwidth for the geographically weighted model;
2. An exploratory GW analysis using GW correlations, a GWR fit and various GWR collinearity diagnostics.
3. The calibration of a locally-compensated ridge regression model;

A full description of robust geographically weighted analysis can be found in Gollini *et al.* (2013). In this study a bi-square function was applied and result of the weighting means that data nearer to the kernel centre make a greater contribution to the estimation of local regression coefficients and a weighted least squares regression model is constructed at each regression point *i*. Here, an optimum kernel bandwidth for GWR can be found by minimising a model fit diagnostic and this case a leave-one-out cross-validation (CV) score (Bowman 1984; Brunsdon *et al.* 1996) was used under a bi-square kernel. The standard GWR model is:

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \epsilon_i \qquad \text{(Eqn 1)}$$

where y_i is the dependent variable at location i, x_{ik} is the value of the k^{th} independent variable at location i, m is the number of independent variables, β_{i0} is the intercept term at location i, β_{ik} is the local regression coefficient for the k^{th} independent variable at location i and ε_i is the random error at location i.

Details on how such a global ridge regression model is adapted to a GW form to provide the locally-compensated ridge GWR model can be found in Brunsdon *et al.* (2012); Lu *et al.* (2014); Gollini *et al.* (2013). These studies also describe associated local collinearity diagnostics for GWR, such as localised correlations amongst pairs of predictors, local VIFs

for each predictor, local variance decomposition proportions (VDPs); and cross-product matrix Condition Numbers (CNs).

3. Initial Results and Discussion Points

A GWR collinearity diagnostic procedure was run which returns CNs describing local correlations amongst pairs of predictors. CNs greater than 30 suggest evidence of local collinearity. In this case the CNs were found to range from 12.78 to 941.76, with a mean CN of 38.78, suggesting that there are areas with very high levels of collinearity present in the study area when the data are considered in the context of local analyses such as GWR. The CN values for each data point are shown in Figure 2 and show that regions of strong collinearity exists in areas around the central portion of the study area.

In order to take account of the observed local collinearity, a locally compensated ridge GWR was calibrated and the analysis was conducted over a 200m grid of points covering the study area. The spatially distributed variables were combined using the spatially distributed coefficient estimates arising from the locally compensated GWR in order to construct the model of house price in Figure 3. This shows distinct patterns of predicted house prices reflecting not only the downtown poles that one would expect but also distinct poles of different levels of house price in areas around the centre.

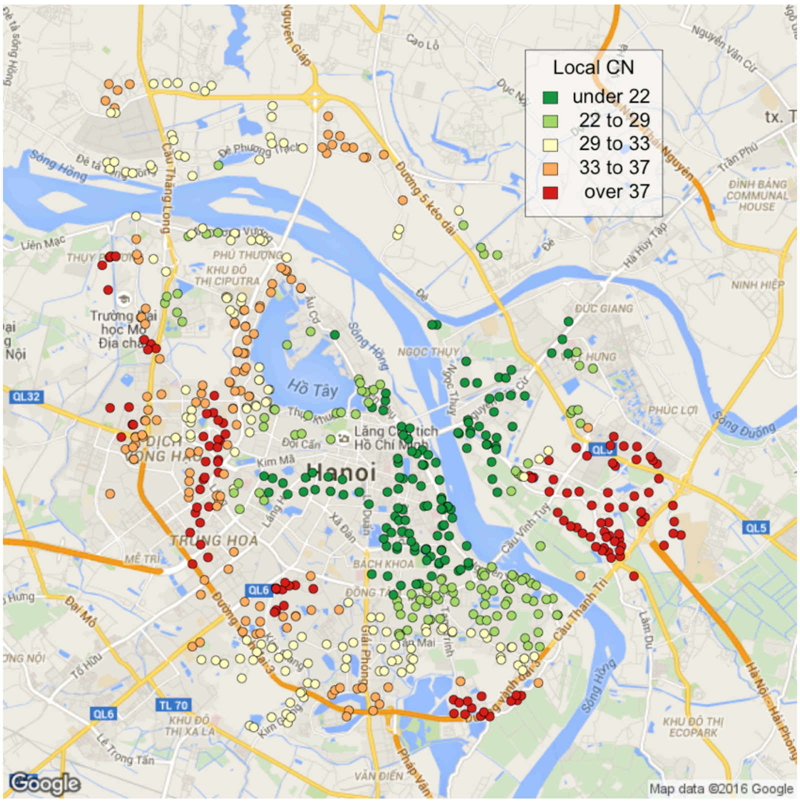


Figure 2. The spatial variation in condition number (CN) in the survey data, with GoogleMaps backdrop.

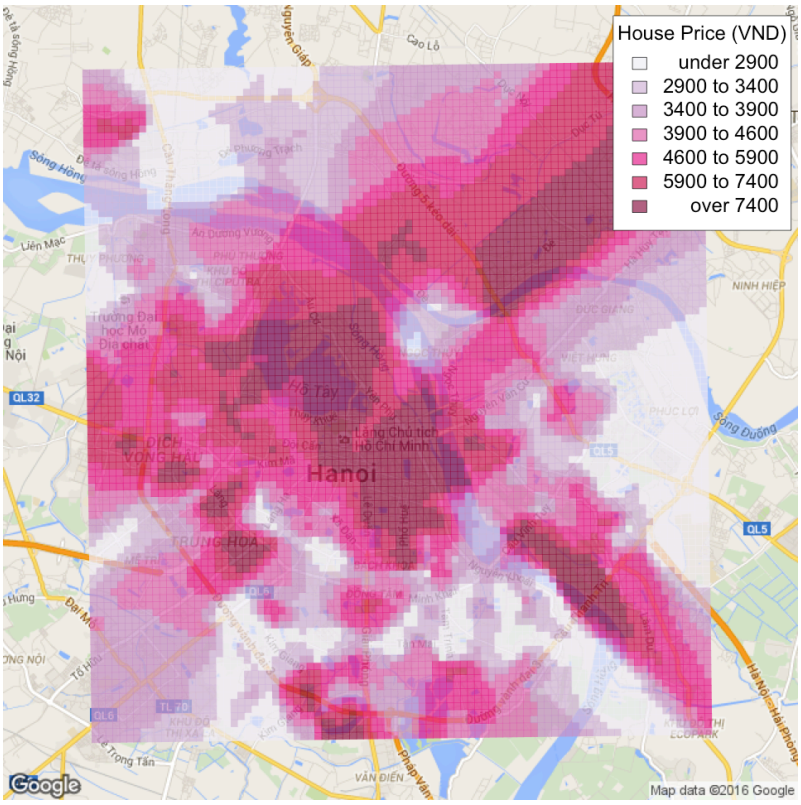


Figure 3. The spatial distribution of house prices in Hanoi derived from a locally compensated GWR model.

The results of this analysis demonstrate that it is important to consider and test for local collinearity even where none is found to exist globally. When local collinearity is found, local models should be applied that are able to handle it. Wheeler (2007; 2009) proposed penalized GWR models with the ridge GWR model and the GW lasso (Wheeler 2009). This study also applied a ridge GWR, but only applied a ridge term where necessary, when the local CN was found to be greater than 30. Thus local ridges were found (Brunsdon *et al.* 2012) and not a global one as applied in the Wheeler (2007) study.

Future work will extend this analysis in a number of directions: network analyses to account for the river running through Hanoi; the application of mixed GWR approaches that include both globally-fixed with locally-varying model coefficients (Nakaya *et al.* 2005) and will disaggregate the poles to identify geo-demographic specific areas of high status. The approach will also be extended to model house price bubbles.

Acknowledgements

This collaborative research was funded by British Academy Newton Advanced Fellowships NG150097, *Developing a housing model based on the status-quality trade off theory*.

References

Alonso W. (1964). *Location and Land Use*. Harvard University Press, Cambridge, Massachusetts.

Bowman A (1984). An Alternative Method of Cross-Validation for the Smoothing of Density Estimates. *Biometrika*, 71, 353–360.

Brunsdon C, Charlton M, Harris P (2012). Living with Collinearity in Local Regression Models. In *Proceedings of the 10th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. Brasil.

Brunsdon, C., Fotheringham, A.S., Charlton, M., (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28(4), 281-298.

Fujita, M. (1989). *Urban Economic Theory, Land Use and City Size*, Cambridge: Cambridge University Press.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., & Harris, P. (2013). GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. *Journal of Statistical Software,* 63.

Kim, J-H., Pagliara, F. and Preston, J., 2005, The intention to move and residential location choice behaviour. *Urban Studies*, Vol. 42, 1621–1636.

Lu, B., Harris, P., Charlton, M., & Brunsdon, C. (2014). The GWmodel R package: further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-spatial Information Science*, *17*(2), 85-101.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in medicine*, *24*(17), 2695-2717.

Parkes, A., Kearns, A. and Atkinson, R., 2002, What makes people dissatisfied with their neighbourhoods? *Urban Studies*, Vol. 39, 2413–2438.

Phe HH and Wakely P (2000). Status, quality and the other trade-off: Towards a new theory of urban residential location. *Urban Studies*, 37(1), 7-35.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography*, *46*, 234-240.

Weisbrod, G. E., Lerman, S. R., and Ben-Akiva, M., 1980, Tradeoffs in residential location decisions: Transportation versus other factors. *Transport Policy and Decision Making*, Vol. 1, 13–26.

Wheeler D (2007). Diagnostic Tools and a Remedial Method for Collinearity in Geographically Weighted Regression. *Environment and Planning A*, 39(10), 2464–2481.

Wheeler D (2009). Simultaneous Coefficient Penalization and Model Selection in Geographically Weighted Regression: the Geographically Weighted Lasso. *Environment and Planning A*, 41(3), 722–742.

An Object Based Approach for Submarine Canyon Identification from Surface Networks

Andrés Cortés Murcia[1,2] , Éric Guilbert [1,2], Mir Abolfazl Mostafavi[1,2]

[1] Dept. of Geomatics Sciences, Université Laval, Québec, G1V 0A6 (QC) Canada
[2]Center for research in geomatics, Université Laval, Québec, G1V 0A6 (QC) Canada
Email: andres.cortes-murcia.1@ulaval.ca
{eric.guilbert; mir-abolfazl.mostafavi}@scg.ulaval.ca

Abstract

In this paper we propose using surface networks to identify submarine canyons from bathymetric data. Identification is done in two steps. First, thalweg lines that fit the canyon definition are extracted; second, the floor around each thalweg is measured to separate steep narrow canyons from broader channels. Results are validated against a classification provided by geomorphologists.

1. Introduction

Submarine canyons are relevant features for geomorphologists because they can explain the origin and evolution of marine landscape. Although there is a common understanding of what a canyon is, its description is often vague and its definition is not applicable for automatic classification. This issue leads to define parameters with arbitrary values that are difficult to establish.

Traditional methods require image segmentation and classification. These approaches compute discrete local descriptors such as the curvature and depend on threshold parameters chosen by the user. Furtermore, image classification can omit global photo-interpretative characteristics. In general, photo-interpreters identify canyons by their overall shape (narrow, elongated, steep slopes) and position (running across the continental slope in a straight line), observed around salient thalwegs.

This paper proposes an approach where thalwegs are extracted from a terrain surface network and canyons are built around them. We move away from pixel classification to an object-oriented approach built on a topological structure. The surface network is a graph where critical points such as pits and peaks are connected by ridges and thalwegs. Its extraction does not require any parameter. Relevant thalwegs are selected by simplifying the surface network. Simplification parameters are not set locally at pixel level but at the structure level. The valley floor is computed around each thalweg and used to classify canyons and other channels. The method is illustrated on a triangulated irregular network generated from multibeam sounding data from the St-Lawrence estuary (Canada). Results were validated against a manual classification performed by geomorphologists (Normandeau *et a.l* 2015).

2. Surface Network Construction

A surface network is a topological graph. Its nodes are pits (local minima), peaks (local maxima) and saddles (points being a local maximum in one direction and a local minimum in another direction). Saddles are connected to peaks by ridges and to pits by thalwegs (Figure 1). The integrity of the surface network is guaranteed by several topological constraints.

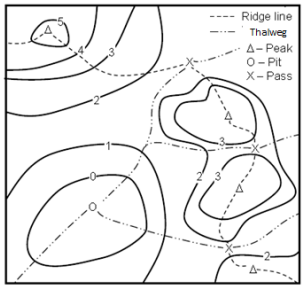


Figure 1. Surface Network Definition.

The surface network was extracted from the TIN following Takahashi *et al.*(1995). Robustness was ensured by setting the following rules: (1) saddles are counted with their multiplicity; (2) if two adjacent points are at the same elevation, their coordinates are compared to provide a total ordering of the points; (3) a virtual pit is added outside the domain and inside each hole corresponding to islands to avoid edge effects. The results are illustrated in figure 2A.

3. Surface Network Simplification

A canyon thalweg is identified by a series of connected thalwegs that cross the continental slope. As such, it starts with a saddle point located at the top of the slope and ends with a pit at its bottom. Extraction of the canyon thalweg is done by removing pits and joining thalwegs into longer lines. Point removal must preserve the topological integrity hence a peak or a pit is always removed with an adjacent saddle point (Rana, 2010). Figure 2 presents a sample of the surface network constructed from the TIN (A) and its simplification (B).

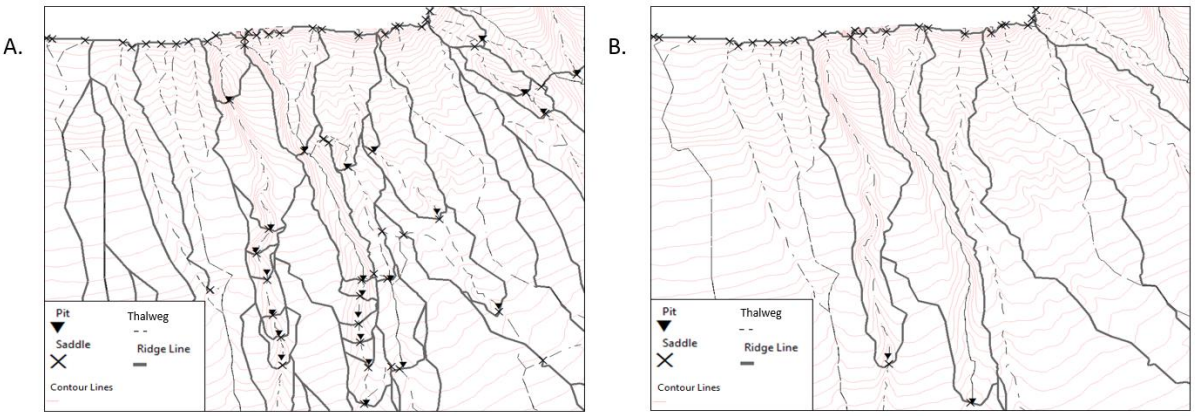


Figure 2. Surface network extracted from the model and its simplification.

Noting p_0, p_1 and p_2 three pits joined by consecutive thalweg lines, three criteria are applied to check if the lines p_0p_1, and p_1p_2 belong to a canyon and p_1 can be removed (figure 3). First, if the slope ratio between p_0p_1 and p_1p_2 is very small or large, there is a change of slope and p_1 must be preserved: pit p_1 can be located at the bottom of a slope. According to experiments, small values are lower than 0.3 and large values are greater than 3. If this value is close to one, slopes are homogeneous. Second, the Euclidian distance and the distance along the thalweg between two consecutive pits are computed. If their ratio is small, the thalweg is meandering and does not belong to a canyon. Finally, only thalwegs located on the estuary slopes are kept. Thalwegs located on the floor of the estuary, characterised by a small height difference, are removed.

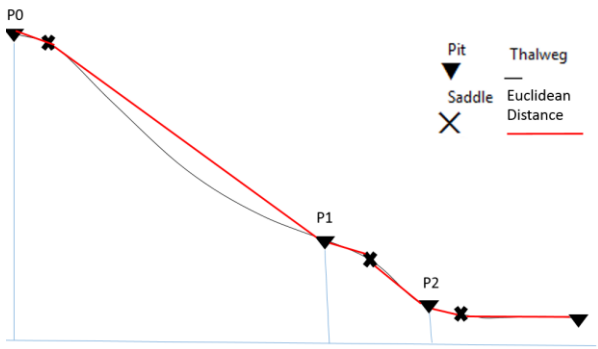


Figure 3. Criteria to identify the canyon thalweg.

4. Valley Floor Extraction

The St-Lawrence estuary contains different types of submarine valleys. Our canyon thalweg definition matches both channels and canyons. In comparison with channels, canyons are narrower with steeper sides. Hence two criteria were identified to separate both types of features: the floor width and the break of slope measured between the floor and the side.

The canyon thalweg is the line of lowest elevation that defines the axis of the canyon. The canyon floor is relatively flat along a direction orthogonal to the thalweg. Hence the floor can be delineated by a growing region approach from the thalweg (Straumann and Purves, 2008). Instead of setting an absolute slope threshold defining flatness, an average cross-sectional slope is measured along the thalweg in its close vicinity and used as a threshold value. A polygon containing the points satisfying the slope criterion is then built around the thalweg (figure 4). This approach allows defining the floor without fixing a parameter value. Thresholds in our study area were measured between 0.5 and 2 degrees. Overall, the average threshold was 1.5 degrees, which agreed with values proposed by Straumann and Purves (2008).

The width of a channel or canyon was measured by the average distance from the thalweg to the border of the floor. A threshold width was set according to geomorphologists' classification to 150 m.

For the second criterion, the angle between the floor and the slope was measured by taking a point on the slope. It was measured by the ratio between the height difference and the distance from the point to the thalweg. If the angle difference is important, there is a break of slope characterising a canyon. On the opposite, a small angle difference corresponds to a smooth change of slope. The maximum ratio for channels was measured at 0.029 while the minimum ratio for canyons was 0.049. The gap being quite significant, this criterion appears more robust than the width criterion.

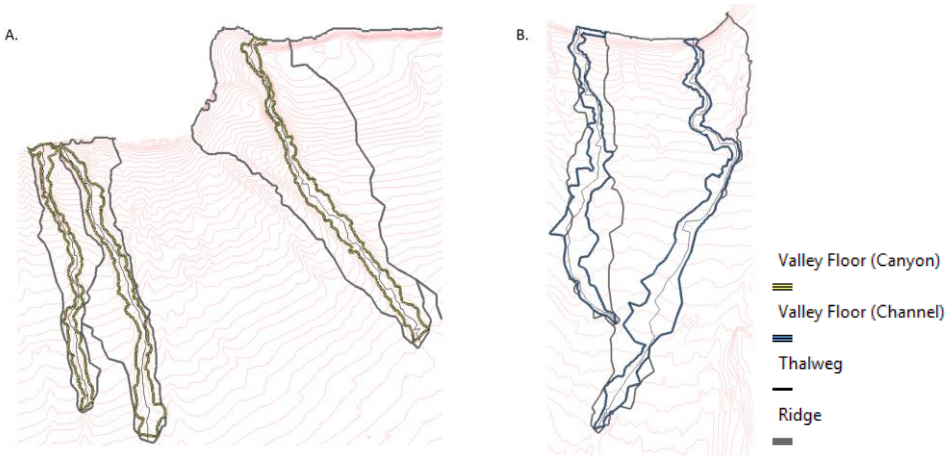


Figure 4. Left: canyons, right: channels.

5. Discussion and conclusions

Our results were compared with the classification obtained by Normandeau *et al*.(2015) who studied canyons in the St-Lawrence estuary based on their geomorphology and feeding sources. Threshold parameters were set through discussion with geomorphologists in order to match their classification. Our approach provides highly accurate landform locations (86%) compared to the manual approach. This measure considers similar results (same location but different length) by each work zone (table 1). As seen from figure 5, thalwegs were shorter because they stop at the bottom of the slope while canyons can extend in the estuary floor depending on how far sediments are transported.

Table 1. Accuracy of our approach

7 Work Zones	Manual Approach		Automatic Approach		Similar Results
	Canyons	Channels	Canyons	Channels	
TOTAL	12	9	11	5	6 of 7 zones
Accuracy			92%	56%	86%

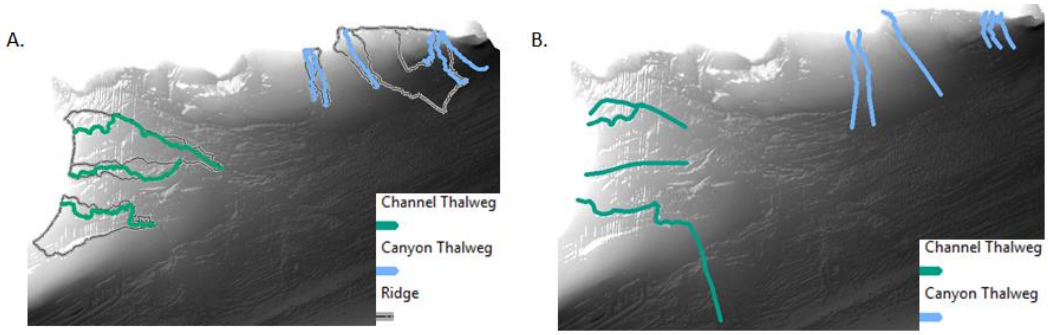


Figure 5. A)our approach, B)manual classification.

While a correct classification was obtained, it relies on the definition of threshold parameters, which were obtained through testing and comparison. In order to extend the model to other domains, further work is required to analyse threshold values and give a contextual definition which may depend on the area considered (including geometrical parameters such as the average slope and size of the continental slope or the shelf) and on geomorphological parameters such as the floor composition or some dynamic process observed over geomorphological scales. On a shorter term, the model shall be extended to describe other features that can be observed in the estuary, not only valleys but also the floor, slopes and shelf in order to provide a description of the morphological structure of the estuary at different levels.

Acknowledgements

Authors would like to thank professor Patrick Lajeunesse and Alexandre Normandeau for their assistance. Data were provided by the Canadian Hydrographic Service.

References

Normandeau, A., Lajeunesse, P. and St-Onge, G., 2015. Submarine canyons and channels in the lower St. Lawrence Estuary (Eastern Canada): Morphology, classification and recent sediment dynamics. Geomorphology 241, pp. 1–18.

Rana S., 2010. Surface Networks: New techniques for their automated extraction, generalisation and application. Enhanced Thesis. University College London, 128 pages.

Straumann, R. K., and Purves, R. S. 2008. Delineation of valleys and valley floors. In T. J. Cova, H. J. Miller, K. Beard, A. Frank, & M. Goodchild (eds), 5th Int. Conf. GIScience (pp. 320-336). Berlin: Springer.

Takahashi, S., Ikeda, T., Shinagawa, Y., Kunii, T., Ueda, M., 1995.Algorithms for Extracting Correct Critical Points and Constructing Topological Graphs from Discrete Geographical Elevation Data. Computer Graphics Forum, 14 (3), C-181-192.

The Encyclopedia Gallica of Events (or Why Geographic Information Science is Not Like Physics)

H. Couclelis

University of California, Santa Barbara
Santa Barbara, CA 93106-4060, USA
Email: cook@geog.ucsb.edu

Abstract

As a contribution towards the further development of user-centered information systems, I present an argument for a contextual and subjective view of events and related concepts in information science that is distinct from the factual view prevalent in empirical science and everyday life. The central notion is that of 'R-Event', where 'R' stands for 'Relevant'. Drawing on representations of process such as dynamic models or AI approaches to problem-solving and planning, R-Events would improve the precision and value of search results by foregrounding and ranking information about events of high relevance to the user. There is no suggestion here that this proposal is easily implementable. It is offered at this stage as a potentially fruitful thought experiment.

1. Introduction

The title is one that Borges would have loved[1]. Had he written the story of the Encyclopedia Gallica of Events, this would have also included processes, perdurants, occurrences, and non-events. And the Encyclopedia in question would have had to be infinite in length, and self-contradictory. I will argue that in an *informational* (as opposed to an empirical) context, (a), the above concepts do not refer to intersubjectively definable individuals and (b), unlike most other abstract concepts, the things these refer to are not free-standing notions but binary relations connecting information source and observer, and depending on the specific interest the latter has in certain detectable state(s) or change(s) in the world. Thus the news of my birth signaled a momentous event for my family (and a very real painful process for my mother) – but for the rest of humanity the fact of my birth was a non-event, barely registering as a blip of plus-one added to some country's population. Similarly, major flooding in south-eastern Tajikistan may be an empirical fact, but it is an event if you are in the business of importing cotton from that area, a process if you are in charge of evacuating local populations, an occurrence if you are tabulating floods in Central Asia, and noise if south-eastern Tajikistan is not among the things you care about. Similar kinds of context-dependency also hold for the endurant-perdurant distinction, and as has been argued repeatedly over the years, also in the case of fields and objects. Information systems sensitive to user-centered semantics should be able to make these determinations.

2. On Science and Metascience

It may be useful to distinguish clearly between the empirical sciences that directly measure and represent phenomena in the world, and the information sciences (which are meta-sciences) that process and present information about these phenomena in ways that meet and support the interests and purposes of information users. These are really two different epistemic layers, with different

[1] See Borges (1939) The Total Library; Borges (1943) The Library of Babel; and the "certain Chinese Encyclopedia" in Borges (1952) The Analytical Language of John Wilkins.

functions. I will argue for the importance of not conflating the two, because as information scientists we are not *doing* hydrology or *doing* forestry or *doing* urban studies, but trying to help answer questions posed by hydrologists, foresters, planners, and any others, in the most appropriate and helpful ways. There is a difference between knowledge representation on the one hand and information (re)presentation on the other, though the latter must of course draw on the former. It is no wonder that some of the smartest people in our field cannot come to an agreement on the definitions of events, processes, and so on. This is because 'it depends'. The challenge as I see it is to take data and transform them into information that is right for particular users coming to these data from particular angles.

Presenting information appropriately forces us to confront and combine three key notions: first, the changing states of the environment as measured by simple, non-cognitive sensors (the 'pre-facts'); second, the intentional users' contextual interest (or lack thereof) in particular states and changes; and third, an information science's duty to help answer questions about facts in ways appropriate for specific kinds of inquiries by these users. To fix ideas, let us consider what may fall into each of these three categories. First are the uninterpreted, uncritical raw measurements provided by sensors: temperature goes up, temperature goes down; one wave-length is sensed, another wave-length is sensed; this many cars trip the counter at time t, this many cars trip the counter at $t+n$. Second are the myriad changes in the world reflected in data at any instant, and the need to sift among them in search of relevance. Even within specific domains of interest, whether and which changes are important when is a subjective and contextual judgement. So – and third – here is the problem: how can an information modelling system be made to take data from the first group, and process them as information within appropriate categories in the second group, in a way that is logically consistent, semantically meaningful, and relevant to the user?

Galton (2015, p.7) wrote "…it is, presumably, the responsibility of the latter [the data modelling system] to extract from this processual flux those hard nuggets of salience which constitute events, and which from a human perspective represent information rather than mere data." Indeed -with the caveat that one person's hard nuggets of salience may be another person's fool's gold. In this vein, following Bateson, I will define an event as 'a difference that makes a difference'. For clarity, within an informational context I will talk about *R-Events* (where R stands for 'relevant'), to distinguish from empirical facts such as landslides or riots, where talk of events, processes, etc. as applied to specific instances is necessary for communication. R-Events are thus Galton's "hard nuggets of salience" tailored to the interests of specific observers. They emerge from (empirical) events that are likely to be significant in particular situations of interest. R-Events increase the precision of search results by picking out information of high value to the user and possibly also ranking it by significance. Thus our cotton importer would want to know right away about the SE Tajikistan flood because of likely impacts on the infrastructure and human resources of the commodity movement chain: critical roads closed, warehouses flooded, workers unable to reach loading sites. Other headlines from the region, e.g. about a toxic spill, a plane crash, or the election of a new governor would be of no direct import and are better left unstated.

3. Speculations

Elsewhere I proposed a sketch of an information system ontology that could perhaps help move such an agenda forward. It consists of three modules or interrelated parts, named after the well-known triad from linguistics: syntax, semantics, and pragmatics. This is more than mere analogy, since an information system must produce meaningful and appropriate statements about facts.

- The *Pragmatics* module reflects the *context* of the inquiry: that is, anything that shapes and constrains user interests and their appropriate mode of satisfaction. The context may be professional, educational, social, etc. and may be constrained by time, space, budget, data, etc.
- The *Syntactics* module handles the sensor and other data (pre-interpretation) and the data models (post-interpretation), following input from the *Semantics* module, and within the constraints flagged by the *Pragmatics* module.
- The *Semantics* module is the core of the system. It receives input from the *Pragmatics* module in the form of a specification of user interests and constraints, and consists of something like the 7-tier structure I proposed some years ago, which is traversed starting with the purpose of the inquiry and moving down all the way to the specification of the most appropriate spatiotemporal frame for results presentation and analysis (Couclelis 2010). At each step the *Semantics* module queries the *Syntactics* module for data relevant to the corresponding level.

As discussed earlier, there are two distinct epistemic layers, the informational and the empirical. The above three modules taken together correspond to the informational layer, while the empirical layer consists of some dynamic representation of the domain of interest. To identify R-Events, the *Semantics* and *Syntactics* modules search the empirical representation. Depending on the domain and the user's intention, as specified by the *Pragmatics* module, this could be a scientific model, a plan, or some other quantifiable pattern of activity at any appropriate level of detail.

In the next section I outline how *plans* may be used to identify changes in a domain of interest that give rise to R-Events. Often a user's intentions are actualized in a plan that may be simple or complex, implicit or explicit. Plans have been studied extensively in AI (Russel and Norvig 2009). They include goals, states, and actions, the latter meant to move states towards the plan's goals. Actions imply agents, which need not be sentient. Plans are often hierarchical, composed of goals, subgoals, targets, etc. as well as activities, tasks, subtasks, etc. In automated planning these are represented in the approach known as Hierarchical Task Networks (HTN).

In connection with plans, R-Events may be defined as changes in the world that cause a breakdown of a plan or its parts. The context may change, causing the interests of the user to change focus; the means to desired goals may change, forcing changes in strategy; the availability of relevant data may change, expanding or restricting the kinds of questions that may be asked and answered. The following kinds of changes are among those likely to give rise to R-events: those preventing planned tasks or routines to be completed; those forcing a change in strategy; those affecting the value to the user of the plan's goal itself; those necessitating a new plan and goals.

4. For Example

A sketchy illustration of the concept of R-Event in a spatial context is provided by the predicament of our hypothetical cotton importer. A successful import operation depends on the timely arrival of shipments from SE Tajikistan to the designated warehouse in the USA. There are two pieces to this operation: (a) plan of action, and (b) a corresponding spatial organization that enables and constrains the elements of that plan. The plan of action may be represented as a network the nodes that correspond to the hierarchy of goals and subgoals in the vertical dimension, while in the horizontal dimension they delimit vectors indicating sequences of activities. Each such sequence supports the goal above it. Thus both goals and activities are being systematically disaggregated as one moves down the hierarchy. In our example the network is headed by a top-level goal, to run a successful import business. Its success depends on everything functioning properly, and in particular, the two broad areas of activity: producing the cotton, and shipping it from source area to destination. Qua subgoals each of these in turn depend on further subgoals and sequences of

activities: picking, sorting, packing, loading cotton, taking it to collection points, trans-shipping it, etc. And further down: getting workers to their workplaces, ensuring appropriate supplies and transport means are available, and so on.

These activities obviously depend on a spatial organization that must function as intended. The producing areas must be physically capable of producing, and the infrastructure must be able to handle the movement of the product from source areas to destination; fields, roads, ports, and warehouses must be in an appropriate spatial configuration; the elements of that configuration must be operational and accessible as needed: e.g., all the segments on shipment routes must be traversable and all necessary warehouses and loading bays must be whole. There is thus a hierarchy of spatial granularity from the global to the very local that reflects and complements that of goals and subgoals. That is, each level of spatial features enables a corresponding level of activities, sub-activities and tasks (Howarth 2008).

Any disruption in the routines described by the plan is a potential R-Event. News of a flood in the region of interest would trigger a manual or automatic search for problems. The *Semantics* module of the information system will check for data in the *Syntactics* module in a sequence roughly as described in Couclelis (2010), which is compatible with the organization of the plan. The sequence is follows: (1) user purpose (as per the *Pragmatics* module), (2) overall function of the intended operation, (3) spatial organization supporting that function, (4) individual elements of the spatial organization, (5) relevant data to be presented to the user, and (6) appropriate spatiotemporal frame for data presentation. R-Events can appear at any level and at several different locations on the hierarchical network representing the plan. Usually the importance of R-Events diminishes as we move down the structure towards more detailed tasks and smaller areas, thus permitting the selected relevant information to also be ranked by degree of criticality.

5. Conclusion

As a contribution towards the further development of user-centered geographic information systems, this paper proposes the following points for discussion: (a), the notion of R-Event as a means of improving the precision of search results as well as the possibility of ranking these by their value to specific users; (b), the further integration of GI science with AI, in connection with the representation of user purposes and the conditions of their satisfaction; and (c), underlying (a) and (b), the potential practical utility of treating the empirical and informational aspects in information systems as distinct epistemic layers. Next steps should include refining the framework using realistic examples, and enriching it with the extensive literature on related topics.

References

Couclelis H, 2010, Ontologies of Geographic Information. *International Journal of Geographical Information Science*, 24(12):1785-1809.

Galton A, 2015, Outline of a formal theory of processes and events, and why GIScience needs one. In: Fabrikant S**,** Raubal M, Bertolotto M, Davies C**,** Freundschuh S, and Bell **S.** (eds.) *Proceedings of the 12th International Conference, COSIT 2015*, Santa Fe, NM, USA, 3-22.

Howarth J, 2008, Landscape and Purpose: Modeling the Functional and Spatial Organization of the Land. Doctoral dissertation, University of California, Santa Barbara.

Russell S and Norvig P, 2009, *Artificial Intelligence: A Modern Approach*. 3d edition, Prentice Hall, Upper Saddle River, NJ, USA

Characterizing Volunteered Geographic Information using Fuzzy Clustering

S. De Sabbata[1], N.J. Tate[1], C. Jarvis[1]

[1]Department of Geography, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom
Email: s.desabbata@le.ac.uk, njt9@le.ac.uk, chj2@le.ac.uk

Abstract

This paper demonstrates the use of fuzzy clustering to characterize Volunteered Geographic Information (VGI). We argue that classifying small areas based on variables related to the amount, type, and currency of VGI can provide a more nuanced understanding of the content. We present a classification of 2011 UK Census Output Areas in Leicestershire (UK) based on content of OpenStreetMap, using a fuzzy *c*-means clustering algorithm, and we compare the resulting classification with a 'standard' socio-economic geodemographic classification.

1. Introduction

The quality of Volunteered Geographic Information (VGI) has long been a focus of research in GIScience (e.g., Haklay, 2010; Goodchild and Li, 2012; Barron et al, 2014). At the same time related questions have been raised concerning the lineage of VGI: who contributes, who is represented and who is not (e.g., Stephens, 2013; Wilson and Graham, 2013; Glasze and Perkins, 2015; Sieber and Haklay 2015). It is evident that a bias exists in VGI, as the majority of content producers seem to be composed of relatively wealthy, younger, western, tech-savvy, male users, and the interests and knowledge of this particular demographic is thus reflected in the produced content. The study of information geographies (Graham et al., 2015) focuses on how the underlying geographies of wealth and access to technology impact the geographic distribution of participation, and in turn the geographies of representation. For instance, Mashhadi et al. (2015) illustrate how population density, wealth, and centrality influence the completeness of OpenStreetMap (OSM) in London, UK. However, as high-quality authoritative benchmarks are not always available, stand-alone approaches based on data mining are necessary to further our understanding of VGI (Senaratne et al., 2016).

The aim of this paper is to characterize VGI through data mining, and explore its relationship to socio-economic variables. We use a fuzzy clustering method for the classification of 2011 UK Census Output Areas (OAs) based on OSM content, and then compare this to a 'stamdard' geodemographic classification: 2011 Output Area Classification (2011OAC) (O'Brien and Cheshire, 2016). We suggest that this process provides a) a more complete and nuanced understanding of OSM content compared to simple density maps and b) it can be the basis for further studies of information geographies, affording both qualitative and quantitative comparisons with socio-economic information, such as census data.

1.1 Numerical clustering, classification, and geodemographics

The objective of a geodemographic analysis (see e.g., Harris et al., 2005; Alexiou and Singleton, 2015) is to classify neighborhoods based on a basket of socio-economic variables, through a process of numerical clustering. Widely used both in social studies and marketing, these classifications are commonly based on census data or surveys, and are constructed through clustering methods such as hard *k*-means (e.g., Singleton and Longley, 2015), or fuzzy *c*-means (e.g., Fisher and Tate, 2015). Longley and Adnan (2016) recently applied these clustering methods to data derived from social media.

2. Methods

Our aim is to characterize VGI for further analysis and comparison, rather than conduct a 'standard' quality assessment, and thus we operate without a benchmark. Although the variables described below do not directly relate to common quality measures, the variables used in this study still aim to capture elements of the completeness, temporal accuracy, and thematic accuracy of the data. The county of Leicestershire (UK) was the focus of our case study. The Planet.osm file was downloaded on March 16th, 2016.

We first computed aggregated values per OA: counts per feature type (i.e., amenity, highway, etc.); total number of features; average number of edits (based on version numbers); the timestamp of the changeset related to the oldest and latest edited object. We then computed z-scores for the total number of features, the average number of edits, and the number of days between the two timestamps and the time of download. As most OAs contain no features of most types, the counts per feature types have been normalized simply as percentages over the total number of features, as a z-score would be skewed by the large amount of zero values.

The clustering procedure was based on 18 variables: z-scores based on the number of features, average edits, time since oldest edited object and latest edited object, and 14 feature types percentages. The percentage of *building* feature type was excluded, due to high negative correlation value with the *highway* feature type (Pearson's $r=-0.79$, $t(3052)=-71.1$, $p<.001$).

Fuzzy clustering was performed using the *cmeans* function of the e1071 R library. We assigned weights to variables, aiming to highlight the amount, type of content, and edits: 0.25 to both number of features and average edits; 0.125 to both time since oldest and latest edited object; 0.25 equally split among the different feature type counts. Based on the analysis of the within-cluster sum of squares, we selected the target number of clusters to be mined to be six.

3. Results and discussion

Among the six classes identified in the classification (OA hard membership based on highest membership value, see Figure 1), one clearly characterizes the city centres ("Central"), with the highest total features density and low *highway* percentages, and another characterizes the countryside ("Countryside") with the lowest total feature density and high percentages of *natural* features. Two classes group most residential areas, one seemly less up-to-date ("Residential A") than the other ("Residential B").

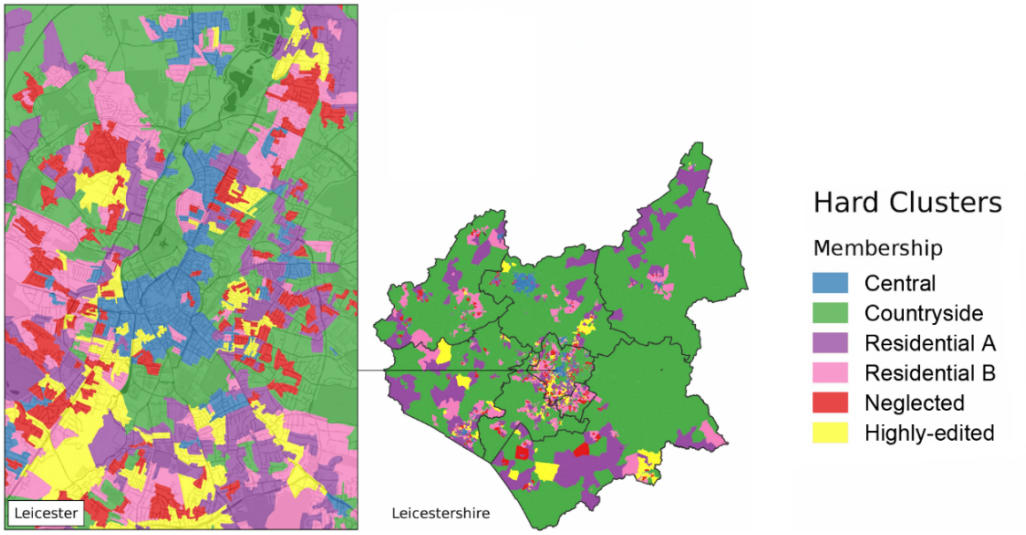


Figure 1. Classification of 2011 OAs based on OSM content. (Contains National Statistics data © Crown copyright and database right, 2016. Contains OS data © Crown copyright and database right, 2016. © OpenStreetMap contributors).

Two other classes are of particular interest in assessing the quality of OSM content. The first is "Neglected", illustrated in Figure 2 (top). The centre of this class has the lowest value of feature density beside the "Countryside" class, and the highest value for the variable related to the time since the last edit has been made. Members of this class are areas which have seen little editing and contain few points of interest. Preliminary analysis suggests that there is a significant association between being classified as "Hard-Pressed Living" in the 2011OAC and being classified as "Neglected" ($\chi^2=15.45$, $p<.001$), with the odds of an area being 'neglected' on OSM being 1.63 (1.26, 2.10) higher if classified as "Hard-Pressed Living".

The second is "Highly-edited", illustrated in Figure 2 (bottom). The centre of this class has a particularly high value for the variable related to the number of edits, which seems to be mostly due to highly edited highways features, or polygons. There seems to be a weak association between being classified as "Cosmopolitans" in the 2011OAC and being classified as "Highly-edited" ($\chi^2=5.83$, $p<.02$), with the odds of an area being 'highly-edited' on OSM being 1.84 (1.05, 3.06) higher if classified as "Cosmopolitans".

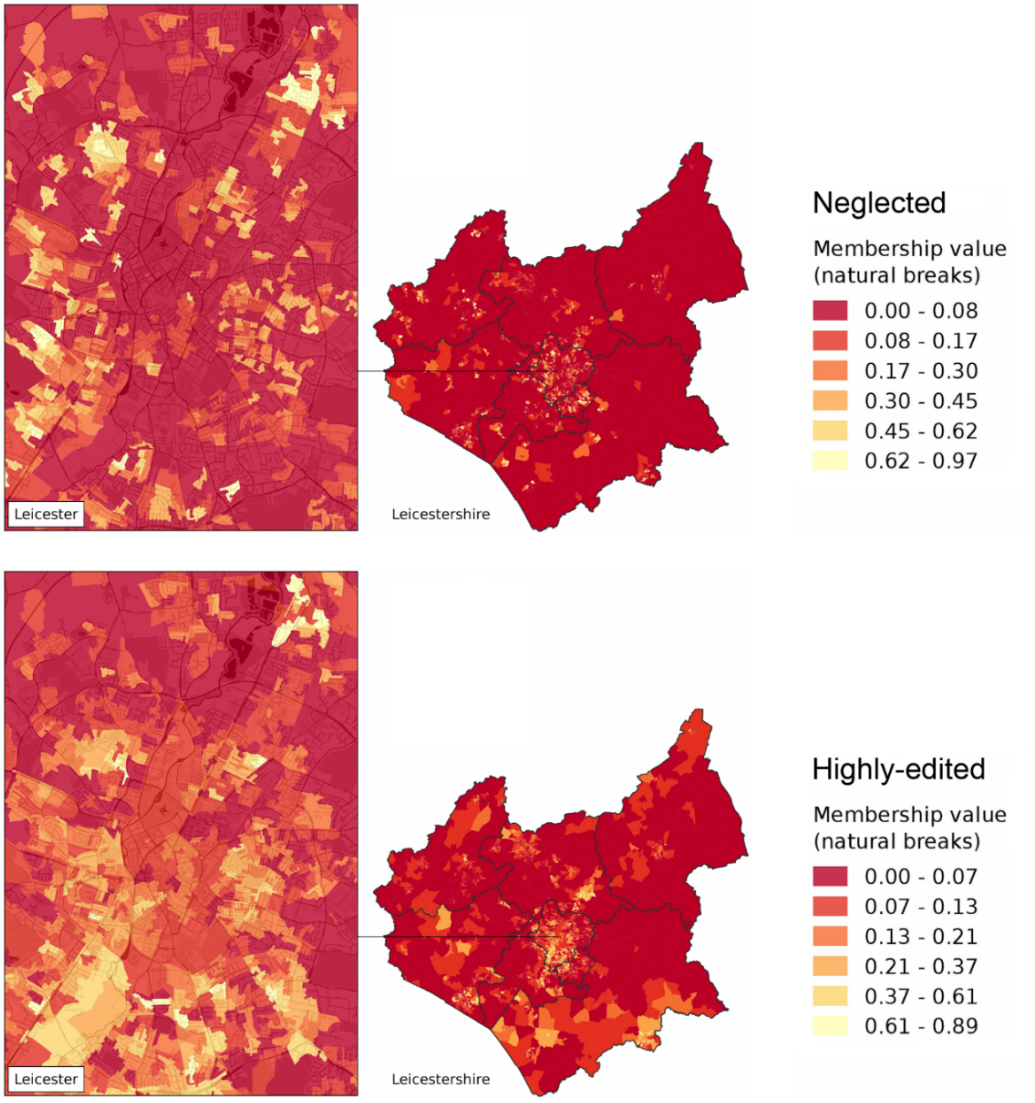


Figure 2. OA membership values for the classes "Neglected" (top) and "Highly-edited"(bottom). (Contains National Statistics data © Crown copyright and database right, 2016. Contains OS data © Crown copyright and database right, 2016. © OpenStreetMap contributors).

Interpreted in the context of the literature discussed above (e.g., Glasze and Perkins, 2015; Mashhadi et al., 2015; Sieber and Haklay 2015), the associations between the classification presented here and the 2011OAC suggest that they could be both related to underlying socio-economic geographies. Future work will focus on a more detailed analysis of the relationships between VGI and socio-economic factors (Senaratne et al., 2016).

The main disadvantages of the methods demonstrated above are the partial arbitrariness of a) the selected attributes, b) the areal units used (and related Modifiable Areal Unit Problem), and c) the interpretation of the resulting clusters. The main advantage is the flexibility of the clustering approach, which allows to create a coherent perspective on the data from a large number of diverse attributes. A broader, open classification has the potential to create a powerful tool for researcher, producers, and users for the analysis, development, and critique of single datasets (as in the present case study), as well as combining multiple VGI sources.

Acknowledgements

The 2011 UK Census Output Area (OA) boundaries and attributes were obtained via the UK Data Service, retrieved from SN:5819. http://discover.ukdataservice.ac.uk/catalogue/?sn=5819
Figure 1 and 2 use map tiles by Stamen Design, under CC BY 3.0. http://maps.stamen.com

References

Alexiou A and Singleton A, 2015, Geodemographic analysis. In: Brunsdon C, and Singleton A, (eds) *Geocomputation: A Practical Primer*, SAGE, London, 137–151.

Barron C, Neis P and Zipf A, 2014, A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS,* 18: 877–895.

Fisher P and Tate NJ, 2015, Modelling class uncertainty in the geodemographic Output Area Classification. *Environment and Planning B*, 42(3): 541–563.

Glasze G, and Perkins C, 2015, Social and political dimensions of the OpenStreetMap project: Towards a critical geographical research agenda. In: Arsanjani JJ, Zipf A, Mooney P, and Helbich M, (eds), *OpenStreetMap in GIScience: Experiences, Research, and Applications*, Springer: Heidelberg, 143–166.

Goodchild MF, and Li L, 2012, Assuring the quality of volunteered geographic information. *Spatial Statistics,* 1: 110–120.

Graham M, De Sabbata S, Zook, MA, 2015, Towards a study of information geographies: (im)mutable augmentations and a mapping of the geographies of information. *Geo:Geography and Environment,* 2(1): 88–105.

Haklay M, 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning B* 37: 682–703.

Harris R, Sleight P and Webber R, 2005, *Geodemographics, GIS and Neighbourhood Targeting*. John Wiley and Sons, Chichester, UK.

Longley PA and Adnan PA, 2016, Geotemporal Twitter demographics. *International Journal of Geographical Information Science* 30(2): 368–389.

Mashhadi A, Quattrone G, and Capra L, 2015, The impact of society on volunteered geographic information: the case of OpenStreetMap. In: Arsanjani JJ, Zipf A, Mooney P, and Helbich M, (eds), *OpenStreetMap in GIScience: Experiences, Research, and Applications,* Springer: Heidelberg, 125–141.

O'Brien O, Cheshire J, 2016, Interactive mapping for large, open demographic data sets using familiar geographical features. *Journal of Maps* 12, 676–683.

Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay M, 2016, A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*. Forthcoming. DOI: 10.1080/13658816.2016.1189556

Sieber,RE, and Haklay M, 2015, The epistemology(s) of volunteered geographic information: a critique. *Geo: Geography and Environment*, 2(1): 122–136.

Singleton AD and Longley PA, 2015, The internal structure of Greater London: a comparison of national and regional geodemographic models. *Geo:Geography and Environment*, 2(1): 69–87.

Stephens M, 2013, Gender and the GeoWeb: divisions in the production of user-generated cartographic information. *GeoJournal*, 78(6): 981–996.

Wilson MW, and Graham M, 2013, Situating Neogeography. *Environment and Planning A*, 45: 3–9.

Time-Geography in Four Dimensions: Potential Path Volumes around 3D Trajectories

U. Demšar[1], J. A. Long[1]

[1]School of Geography & Geosciences, University of St Andrews, Scotland, UK
Email: {urska.demsar; jed.long}@st-andrews.ac.uk

Abstract

An upcoming increase in availability and accuracy of 3D positioning requires development of new analytical approaches that will incorporate the third positional dimension, the elevation and model space and time as a 4D concept. In this paper we propose the extension of time geography into four dimensions. We generalise the time geography concept of a Potential Path Area into a Potential Path Volume around a 3D trajectory and present its mathematical definition. The algorithm for calculating PPVs around 3D trajectories is currently being implemented and will be tested on simulated data and real 3D data from movement ecology.

1. Introduction

Movement data are collected in the form of trajectories, which are sequences of locations collected at certain times. These data are collected using trackers, such as GPS devices, which are capable of recording location in three dimensions. However, typically only the two geographical dimensions (either longitude/latitude or easting/northing) are used for analysis and the third dimension, elevation, is neglected (Belant *et al*. 2012). This is first because the accuracy of GPS elevation measurements is poor, but also because including the third dimension into any kind of geometrical calculations necessary for correct mathematical analysis introduces higher complexity. The upcoming deployment of the European Galileo and the Chinese COMPASS/Beidou systems are expected to improve elevation accuracy (Li *et al.* 2015). Therefore an upcoming critical gap is for analytical methods for 3D movement.

In this paper we propose a generalisation of a well-established movement analysis framework, time geography (Hägerstrand 1970) to consider location in three dimensions. This framework originally operates in the conceptual space of a Space-Time Cube (STC), which consists of a 2D geographic plane and time as the third axis. We propose to extend this concept into a Space-Time HyperCube (STHC), which we define as a 4D space, consisting of a 3D geographic space and fourth dimension - time. By applying the conceptual extension from STC to STHC, we can mathematically generalise time geography into four dimensions.

We focus on one particular time geography concept: the Potential Path Area (PPA), a popular accessibility measure in transport (Patterson and Farber 2015). A PPA is the projection of a Space-Time Prism (STP) onto the geographical plane (Miller 2005). The STP is an accessibility volume within the STC, which represents all the paths which the object could have traversed between two observed positions, P_i and P_{i+1} (Figure 1a). In the 2D case, the PPA is an ellipse. If this ellipse is calculated around each segment of a trajectory, their union can be used to delineate the range of a moving object (Long and Nelson 2012).

To extend this principle into four dimensions, we propose to generalise the PPA ellipse into an ellipsoid located within the three dimensional geographic space. This ellipsoid is the projection of the four-dimensional Space-Time Prism (the 4D accessibility volume between the two observed positions) onto the 3D base space of the STHC. We call this ellipsoid the Potential Path Volume (PPV, Figure 1b) and propose that it can be used in the same way as the Potential Path Area for trajectories where location is measured in three dimensions.

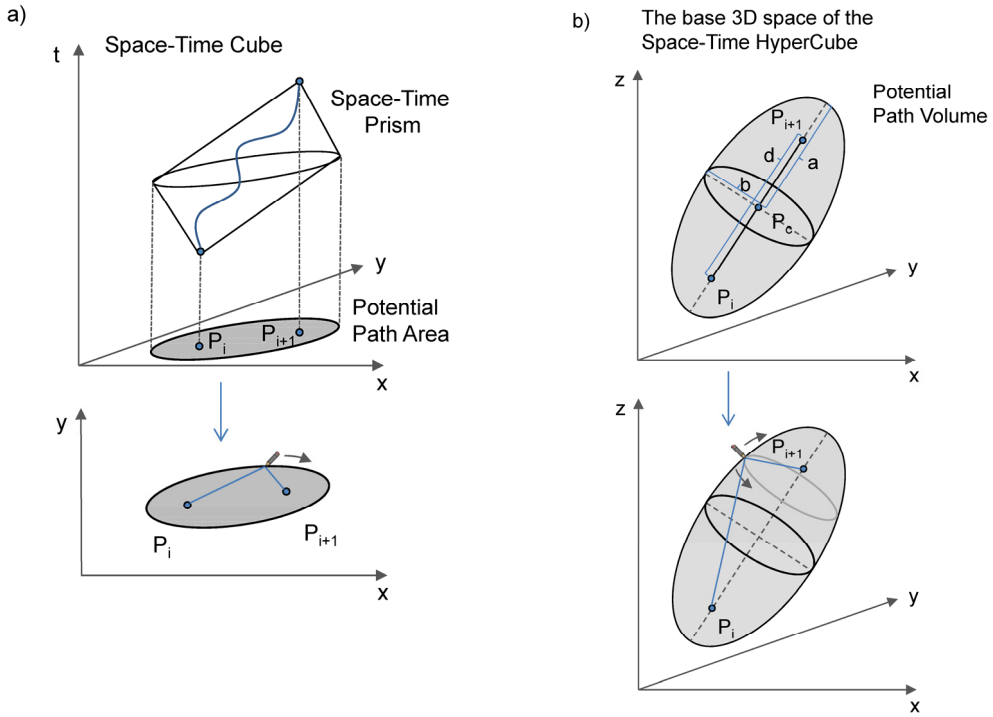


Figure 1: a) The definition of the Potential Path Area as the projection of the Space-Time Prism in an STC and b) its 4D generalisation the Potential Path Volume.

In the remainder of this paper we lay out the mathematical definition of the Potential Path Volume and describe the algorithm for calculation of PPVs for a set of given trajectories. As this is work in progress, implementation and testing of the algorithm are currently on-going.

2. Mathematical definition of the Potential Path Volume

The PPA is an ellipse around a movement segment, so that the two locations (start and end points of a segment, P_i and P_{i+1}) are placed in the foci of the ellipse. Given the maximum possible velocity and time difference between P_i and P_{i+1}, the maximum length of the path between P_i and P_{i+1} can be imagined as a string of this length, fixed in P_i and P_{i+1}. By placing a pen into this string and tracing as far out as possible in all directions, an ellipse is generated that covers all possible paths that the moving object could have passed (Figure 1a).

The same principle can be applied in 3D (Figure 1b). P_i and P_{i+1} are now placed into foci of an oblique ellipsoid. The longest possible length of the path between P_i and P_{i+1} is calculated based on maximum possible velocity and time difference. Then the ellipsoid is generated by tracing the ellipse first in one plane (a plane parallel to movement direction) and then this ellipse is rotated around the axis represented by the direction of movement between P_i and P_{i+1}. This construction results in a special type of an ellipsoid, a prolate spheroid (e.g. the form of a rugby ball), where the two minor axes are identical (Figure 1b), that is, the ellipsoid major axis is a and the two minor axes are b and b. This follows an assumption that movement in the two directions perpendicular to line P_i to P_{i+1} is equally possible. In a more general case, a scalene ellipsoid, where all three axes are different (i.e. major axis a and two minor axes b and c), would be more appropriate – this will be considered in our future work for the types of movements where velocities differ based on direction of movement.

The PPV ellipsoid is defined by the following quantities (Figure 1b): the distance between the two foci (d), the length of the major axis (a), the length of the two minor axes (b), the origin point of the ellipsoid (P_c) and the two rotation angles (α, β) that transform the original coordinate system into a coordinate system defining the ellipsoid (Figure 2a). For

each segment on the trajectory, with start and end points $P_i(x_i, y_i, z_i)$ and $P_{i+1}(x_{i+1}, y_{i+1}, z_{i+1})$, the origin point P_c is given as the central point:

$$P_c = (x_c, y_c, z_c) = (\frac{x_i + x_{i+1}}{2}, \frac{y_i + y_{i+1}}{2}, \frac{z_i + z_{i+1}}{2}) \quad (1)$$

The distance d between P_i and P_{i+1} is Euclidean distance between two 3D points:

$$d = \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2} \quad (2)$$

The major axis of the ellipsoid, a, can be calculated knowing the time difference Δt between P_i and P_{i+1} and the maximum possible velocity v_{max}. For this velocity we could take the maximum observed velocity, however, that would create a degenerate ellipsoid and we therefore follow a more robust calculation of v_{max} (Long and Nelson 2012):

$$v_{max} = 2 \cdot v_m - v_{m-1} \quad (3)$$

where v_m is the maximum observed velocity and v_{m-1} is the next largest observed velocity. Then, a and b are given as this:

$$a = \frac{v_{max} \cdot \Delta t}{2} \qquad b = \sqrt{a^2 - \frac{d^2}{4}} \quad (4)$$

Finally, we also need to find the transformation of the original axes (x, y, z) onto the ellipsoid axes (x', y', z') (Figure 2b). These are created by a combination of the translation of the coordinate origin into the central point P_c and then two rotations, the first one for angle α around the z axis and the second one for angle β around the rotated y-axis (Figure 2a). These two angles are the Tait–Bryan nautical angles of pitch and yaw (because of the symmetry of movement around the axis P_i to P_{i+1}, the roll is not important) and can be calculated as:

$$\alpha = \arctan(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}), \quad \beta = \arcsin(\frac{z_{i+1} - z_i}{d}) \quad (5)$$

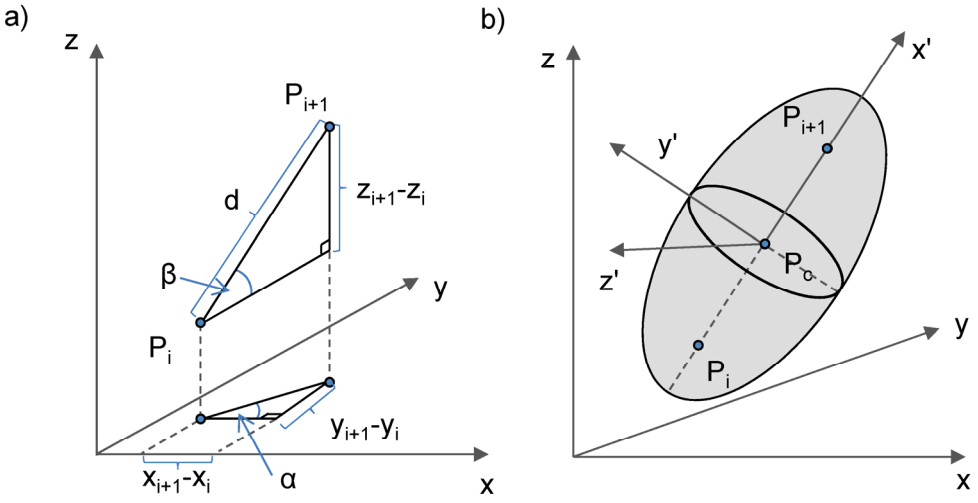


Figure 2. a) Definition of angles α and β and b) transformation of the original coordinate system (x, y, z) into the coordinate system of the ellipsoid (x',y',z').

New coordinates are then calculated as per (6), where $R(\alpha)$ and $R(\beta)$ are rotation matrices (7):

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = R(\beta) \cdot R(\alpha) \cdot \begin{pmatrix} x - x_c \\ y - y_c \\ z - z_c \end{pmatrix} \quad (6)$$

$$R(\alpha) = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad R(\beta) = \begin{pmatrix} \cos\beta & 0 & -\sin\beta \\ 0 & 1 & 0 \\ \sin\beta & 0 & \cos\beta \end{pmatrix} \tag{7}$$

The order of rotations in eq. (6) corresponds to the right to left order in the matrix product, i.e. α first, then β. Given any point *(x,y,z)* in the original coordinate system, we can then determine if the point is within the ellipsoid: this is the case when the following is satisfied:

$$\frac{x'^2}{a^2} + \frac{y'^2}{b^2} + \frac{z'^2}{b^2} \leq 1 \tag{8}$$

We can use this mathematical formulation to define an algorithm that calculates PPVs for a set of 3D trajectories. The algorithm runs through all trajectories and first identifies the largest possible velocity on each trajectory. Then, for each segment of each trajectory it builds the respective PPV as a volumetric representation. That is, it builds a volume where all voxels which, given the respective maximum possible velocity are within the PPV for this trajectory segment (based on (8)), are given value 1 and all others value 0. In the last step, the total PPV volume (for all trajectories and segments) is built as the union of all individual PPV volumes. That is, each voxel of the total PPV volume is assigned value 1 if it belongs to at least one of the individual PPV volumes. The resulting volume delineates the 3D accessibility space that the moving object could have reached. Our algorithm is currently being implemented in R and will be tested on simulated data and real data from movement ecology.

3. Conclusions

A limitation of our algorithm is that the computational complexity is $\mathcal{O}(n \times p \times v)$, where n is the number of trajectories, p the length of the longest trajectory and v the number of voxels. The complexity could be decreased by at each step considering only voxels from the box that bounds the ellipsoid around every segment instead of the entire volume. This has been done previously for space-time densities (Demšar and Virrantaus 2010).

Analogously to the use of PPAs for wildlife trajectories (Long and Nelson 2012), the union of PPVs for all segments could be used as a delineation of a home range for an animal that moves freely in 3D, such as birds or marine animals. We plan to test our approach on a set of real 3D bird trajectories (Tarroux *et al.* 2016). Since the union of PPVs for a set of trajectories defines the volume which an object can reach given its maximum possible velocity, this means that the object could not have been located outside this volume at any time. This could also be used to improve other alternative measures of three dimensional home ranges in ecology, such as the 3D Brownian bridges density (Tracey *et al.* 2014).

References

Belant JL, Millspaugh JJ, Martin JA and Gitzen RA, 2012. Multi-dimensional space use: the final frontier. *Frontiers in Ecology*, doi:10.1890/12.WB.003.

Demšar U and Virrantaus K, 2010. Space-time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10):1527–1542.

Hägerstrand T, 1970, What about people in regional science? *Papers of the Regional Science Assoc.*, 24:7-21.

Li X, Zhang X, Ren X, Fritsche M, Wickert J and Schuh H, 2015, Precise positioning with current multi-constellation GNSS: GPS, GLONASS, Galileo and BeiDou. *Nature Scientific Reports* 5:8328.

Long JA and Nelson TA, 2012. Time Geography and Wildlife Home Range Delineation. *The Journal of Wildlife Management*, 76(2):407-413.

Miller HJ, 2005, A Measurement Theory for Time Geography. *Geographical Analysis*, 37:17-45.

Patterson Z and Farber S, 2015. Potential Path Areas and Activity Spaces in Application: A Review. *Transport Reviews*, 35(6):679-700.

Tarroux A, *et al*., 2016. Data from: Flexible flight response to challenging wind conditions in a commuting Antarctic seabird: do you catch the drift? *Movebank Data Repository*. doi:10.5441/001/1.q206rm6b

Tracey JA, Sheppard J, Zhu J, Wei F, Swaisgood RR and Fisher RN, 2014. Movement-Based Estimation and Visualization of Space Use in 3D for Wildlife Ecology and Conservation. *PloS One*, 9(7):e101205.

High Resolution, Multi-Year Compatible Dasymetric Models of US Population

A. Dmowska[1,2], T. F. Stepinski[1], P. Netzel[1]

[1]University of Cincinnati, Cincinnati, OH, 45221, USA
Email: {dmowskaa;stepintz;netzelpl}@ucmail.uc.edu

[2]Adam Mickiewicz University, Poznan,61-680, Poland
Email: dmowska@amu.edu.pl

Abstract

We developed 30 m resolution grids of the US population in 1990, 2000, and 2010 using a multi-year compatible dasymetric model. These grids are designed to assess population change across the conterminous US at street-level spatial resolution. The model and its novel, computationally efficient implementation in R are described. The grids are available online for interactive exploration and data download using especially developed GeoWeb application SocScape (http://sil.uc.edu/webapps/socscape_usa/).

1. Introduction

The goal of this research is to provide an open and convenient access to high resolution, multi-year compatible population data. Such resource is sought after by academic, government, and industry stakeholders as it can provide access to information needed in numerous applications including social and health services, economic development, and planning. Unfortunately, the US Census data, in its original format of aggregated units, is not multi-year compatible, and, except in the densely populated urban areas, is not given at sufficiently high resolution. Boundaries of small Census aggregates, blocks, block groups, and tracts, change from one Census to another making a direct assessment of change impossible without interpolation. In the rural areas large portions of Census blocks are uninhabited with population restricted to small fragments of the blocks resulting in unrealistic estimates of population density. Numerous other shortcomings of aggregated data have also been identified (Sperling, 2012).

The solution is to calculate a grid-based model of the US population density. Surprisingly, until now no adequate grid-based model had been available. The Socioeconomic Data and Application Center (SEDAC) provides 1 km resolution (250 m for selected metropolitan areas) demographic grids, but they are constructed using an oversimplified model (areal weighting), have insufficient resolution, and are available only for 1990 and 2000. The 90 m dasymetric model (LandScan–USA) has been developed by the Oak Ridge National Laboratory (Bhaduri *et al*., 2007) but it is not available, nor is it expected to be in the public domain when and if it becomes available. Since 2014 we have developed (Dmowska and Stepinski, 2014, 2016) a series of dasymetric models covering the entire conterminous US (CONUS). The first generation of our models was the results of sharpening of SEDAC grids into smaller, 90 m cells; these grids were available only for 1990 and 2000. The second generation grid was the result of direct disaggregation of 2010 Census blocks into 30 m grid using land cover (NLCD2011) and land use (NLUD2010, Theobald 2014) as ancillary data.

Here we report on direct disaggregation of 1990, 2000, and 2010 Census blocks in a manner that makes the three resultant grids comparable to each other thus enabling change assessment. To achieve this we needed to use ancillary data that are compatible between

different years. In addition, we constructed a custom dasymetric model and implemented it in R instead of GIS-specific software in order to make the entire computational pipeline efficient and completely automatic.

2. Data and methods

2.1 Population and ancillary data

The source of the population information is the 1990, 2000, and 2010 decennial Censuses data aggregated to block level. This data consist of two components: shapefiles (TIGER/Line Files), indicating blocks' geographical boundaries, and summary text files which lists population data for each block.

For ancillary data – the data used to guide disaggregation of blocks into the high resolution (hi-res) grid – we use the land cover data. This choice is dictated by the fact that land cover is the only ancillary data for which a single dataset – the National Land Cover Dataset or NLCD – covers the entire CONUS. However, the NLCD1992 (to be used for disaggregation of 1990 blocks) has a legend which is incompatible with the legends of NLCD2001 and NLCD2011 (to be used for disaggregation of 2000 and 2010 blocks, respectively). To enable a direct comparison between NLCD1992 and NLCD2001 the Retrofit Land Cover Change Product (http://www.mrlc.gov/nlcdrlc.php) was developed. In effect, this product consists of two land cover maps with number of land cover classes reduced to eight compatible categories. To transform all three NLCD maps to a common legend we reclassify the retrofit maps for 1992 and 2001 and the NLCD 2011 to just three categories: urban, vegetation, and uninhabited. These reclassified maps were used as ancillary data for dasymetric model.

2.2 Dasymetric model and its implementation in R

The spatial resolution of NLCD is 30 m and it is most convenient to disaggregate blocks to the same resolution. The overall dasymetric model follows the methodology we have developed to obtain 2010 population grid (Dmowska and Stepinski 2016) but we only use modified (see above) land cover as ancillary data. In the first step we rasterize blocks' boundaries shapefiles to 30 m grids. In the second step we determine relationship between land cover categories and population density. Representative population density for each class is established using a set of blocks (selected from the entire US) having relatively homogenous land cover (90% for urban class and 95% for vegetation class). In the third step the population in each block is redistributed to its constituent cells using block-specific weights assigned to the cells having different ancillary category. The weights are assigned based on two factors: relative density of population for each land cover category, and the area of each block occupied by each category. Once the weights are calculated the population in each cell is obtained by multiplying block's aggregated number of people by an appropriate weight. Note that cell's values are population densities in units of people/(area of the cell); they can be fractional, and occasionally quite small. Integrating cells' values over the entire block recovers the aggregated number of people.

The major challenge to calculating 30 m dasymetric model of population density for the entire CONUS is the size of input and output data – we need to disaggregate ~11 million blocks into ~8 billion cells. Traditionally, dasymetric modeling was computed in a GIS environment (ESRI ArcGIS). However, for the model of this size such approach is computationally inefficient. Instead, we implemented our calculations in R (scripts are available at http://sil.uc.edu).

R is a comprehensive computational environment that includes libraries (sp, rgrass7, raster, rgdal) to work with geospatial data as well as libraries to work with standard relational databases (DBI, RSQLite). Data is first preprocessed in GRASS GIS before R is used for actual dasymetric modeling. To manage data storage requirements and to better control the time of computation we process each county separately. We used region concept in GRASS GIS for computationally efficient division of U.S. into separate counties. Raster data for each county is read into SpatialGridDataFrame object which integrates information about its spatial content with Census data into a single relational model. The results of dasymetric modeling for each county is saved in two forms: as SpatialGridDataFrame layer (for mapping) and in a tabular form (for possible subsequent analysis). In the last step dasymetric map for each county is exported from SpatialGridDataFrame object into geotiff (library rgdal) and then maps for individual counties are joined into U.S.-wide map using GDAL tools (gdalbuildvrt, gdal_translate).
Main advantages of using R are: (1) less processing steps are required, (2) increased flexibility and automation, (3) no intermediate layers (4) easily expandable to variables other than total population.

3. Results

The results of our modeling are three population grids, for 1990, 2000, and 2010 which can be used to observe and analyze population change. The only practical mode to access maps of this size is through a GeoWeb application. SocScape (Social Landscape), available at http://sil.uc.edu/webapps/socscape_usa, provides an access to our population grids. In addition to population density grids it also includes grids of racial diversity. We used the weights from dasymetric modelling of total population to disaggregate individual racial groups and combined this information into a single product – a diversity map (Dmowska and Stepinski 2014). SocScape provides fast and intuitive exploration of population and racial diversity patterns starting from the continental scale down to the scale of an individual street. The street map is also included in SocScape, thus, by using an opacity tool, a geographical information can be overlaid on population information. With SocScape one can examine spatial dynamics of population density and racial diversity.

SocScape provides a tool for downloading data for the area of interest. Population density as seen in SocScape is classified to 11 categories but download tool access original, unclassified data. For racial diversity grids, which are classified products, the same layer is displayed and downloaded. Downloaded data are in the GeoTiff format and in the WGS84 Web Mercator (EPSG: 3857) projection. In addition to multi-year compatible data SocScape also provides alternative population and diversity maps for 2010; these maps are the result of more comprehensive dasymetric model (Dmowska and Stepinski 2016) which utilizes additional information available only for 2010. All data is also available for download on county by county and metropolitan area basis at http://sil.uc.edu/cms/index.php?id=socscape-data

Figure 1 shows an example of data downloaded from SocScape. In this example, the population grids for the Las Vegas, NV are shown for 1990, 200, and 2010 revealing spatial details of explosive population growth.

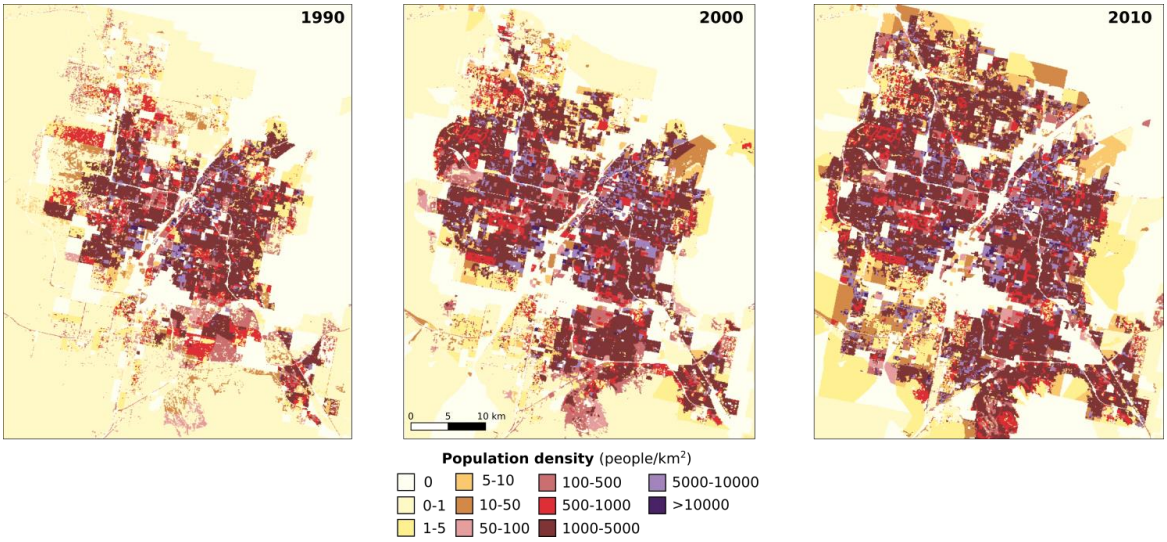


Fig. 1. Population dynamics change for Las Vegas, NV based on hi-res grids for 1990, 2000 and 2010.

4. Conclusion

Our project to provide open and convenient access to hi-res multi-year grids of US population is now completed. The grids are available for exploration and download from the GeoWeb application SocScape. Additional grids pertaining to racial diversity are also available through this application. The new grids are a significant improvement over our first generation grids. They are a valuable resource for anyone who wants to study population and racial dynamics in the US over the last two decades.

Acknowledgements

This work was supported by the University of Cincinnati Space Exploration Institute.

References

Bhaduri B, Bright E, Coleman P, Urban ML, 2007, Land-Scan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal,* 69(1-2): 103–117.

Dmowska A, Stepinski TF, 2014, High resolution dasymetric model of US demographics with application to spatial distribution of racial diversity. *Applied Geography*, 53: 417–426.

Dmowska A, Stepinski TF, 2016, A high resolution population grid for the conterminous United States: The 2010 edition. *Computers, Environment and Urban Systems*, in review.

Sperling J, 2012. The tyranny of census geography: Small-area data and neighborhood statistics. *Cityscape*, 14(2), 219–223.

Theobald D. M, 2014, Development and applications of a comprehensive land use classification and map for the US. *PloS One*, 9(4), e94628.

Context-sensitive spatiotemporal simulation model for movement

S. Dodge

Department of Geography, Environment, and Society, University of Minnesota
Email: sdodge@umn.edu

Abstract

This paper presents a context-sensitive spatiotemporal model to simulate movement trajectories. The model incorporates both the correlated random walk and time-geography theories to generate a more realistic trajectory of an agent within its environment.

1. Introduction

Movement is an essential form of temporal change that is an integral characteristic of dynamic entities (e.g. humans, animals, vehicles, diseases). It is the focus of research in a range of application domains such as transportation, movement ecology, environmental studies, and human health. Movement models help us to better understand the characteristics of movement, enable us to simulate movement and predict its patterns (Dodge 2016). Examples of existing movement models include the random walk and its variations (Codling *et al.* 2008, Technitis *et al.* 2015), time-geography (Miller 2005, Song and Miller 2014), and Brownian Bridge (Horne *et al.* 2007) models. These models either generate trajectories using a set of geometric movement parameters (turn angle, distance), or they identify a visitation probability surface for an agent considering its speed and time budget. Existing models often disregard the characteristics of the environment or the context within which the movement takes place. Simulation of movement in relation to its embedding context is an essential problem that is applied to generate trajectories to fill gaps in low-resolution tracking datasets, or to examine behavioral responses of moving agents to environmental changes. This paper introduces a context-sensitive spatiotemporal simulation model based on a correlated random walk with external biases and is controlled by time-geography constraints of the moving agent. The novelty of the model is that at each step the simulation is driven by *behavior* and the *contextual factors* (i.e. environment, geography) that influence the local movement of the agent. As a case study, this research uses GPS observations of a tiger to parameterize the model and to simulate the tiger's movement between actual GPS observations.

2. Movement Simulation

The overall goal is to generate a trajectory (a sequence of spatiotemporal points) from a start location and time $S(x_s, y_s, t_s)$ to an end location and time $E(x_e, y_e, t_e)$. The simulation uses a correlated random walk from S with an external bias to move towards E (i.e. global constraints). The local movement at each *step* is driven by *agent's behavior* and *contextual factors*. The model specifications are: (1) the maximum movement speed is determined by behavior (e.g. patrolling, hunting, foraging, biking), (2) the global movement path and speed are controlled by the actual time-budget to reach the end-point, and (3) the path is influenced by agent's local choices based on context (environmental drivers and spatial constraints, -e.g. general movement direction, slope preferences, trail network).

The simulation algorithm runs on regular time intervals defined by the user, named *step time*, to ensure the global movement occurs within the time-budget ($tb = t_e - t_s$). The maximum speed of the agent (V_{max}) is determined based on expert knowledge or derived

from GPS observations for the given behavior. At each simulation step, V_{max} is used to delimit a *possible terminal region (PTR)* as shown in Figure 1 (gray area). This is the area which the agent needs to move to by the end of the step to satisfy the time-budget and the global constraints of reaching the end-point. How the agent gets to that region is determined by the local choices it makes along the way. The model uses a raster (in this case a digital elevation model (DEM)) to integrate the influence of contextual factors (e.g. slope) on local movement choices (shown in Fig.1b).

As shown in Figure 1a, the step PTR (gray region) is calculated using the time-geography theory and V_{max} (Miller 2005) as the intersection of (1) the general *potential path area* (khaki ellipse) between the current point $I(x_i, y_i, t_i)$ and the end-point $E(x_e, y_e, t_e)$, (2) the farthest locations (d_{step}, red buffer) that can be reached at each step, and (3) the maximum possible distance (d_{left}) that can still remain to reach E and satisfy the time-budget (green buffer).

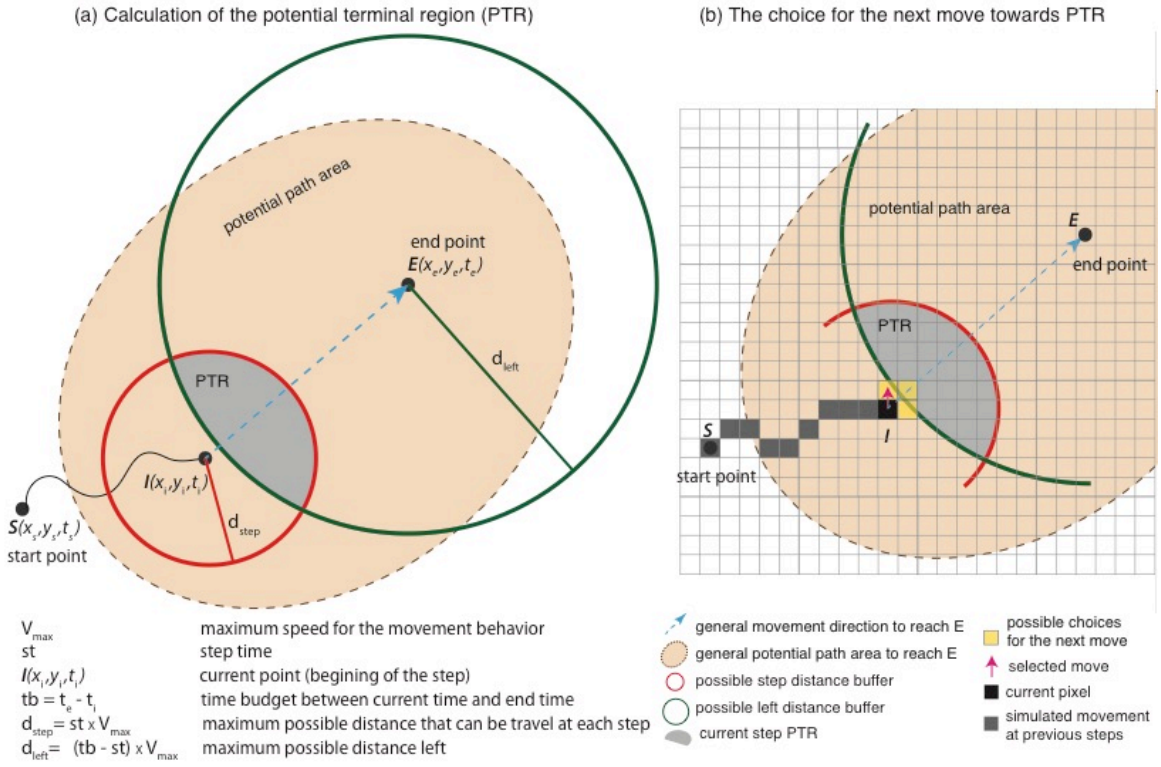


Figure 1. (a) Calculation of the potential terminal region (PTR) at each step, and (b) the choice for the next move towards PTR based on movement direction and context.

Following the calculation of the step PTR, the agent's movement proceeds from the current point I using a correlated random walk towards E. The deviation allowance from the general direction $\underset{IE}{\rightarrow}$ is drawn randomly from a normal distribution with a small δ (e.g. $X \sim N(0, 20°)$) to minimize backtracking. After the selection of movement direction, the associated pixel in that direction and its two neighboring pixels become possible choices for the next move (e.g. yellow pixels in Fig.1b). The move is then made based on the slopes of these three adjacent pixels calculated in the direction of movement. A random slope value is drawn from the χ^2 distribution of tiger slope use derived from actual GPS observations. From the three pixels the one with directional slope value closest to the random value is selected as the next move. The agent is moved to that pixel and the current simulation time is updated according to the speed and raster cell size. The simulation continues pixel-by-pixel until the agent reaches any location within the PTR. The simulation proceeds to the next step by calculating a new PTR from the terminal point of the previous step targeting E using the remaining time-budget. This process continues until the end-point E is reached at time t_e.

3. Results

The proposed model is implemented in Python using Numpy, GDAL, and Shapely libraries. The model is applied to actual observations of a tiger tracked over one year in Thailand Huai Kha Kaeng Wildlife Sanctuary with a sampling rate of 1 hour. The model was parameterized using the Kernel density plots of slope values used by the tiger and its speeds obtained from 4874 GPS observations (Fig.2). The maximum speed values for patrolling (1.8 km/h) and non-patrolling behaviors (0.7 km/h) are obtained using segmentation of the tiger trajectory.

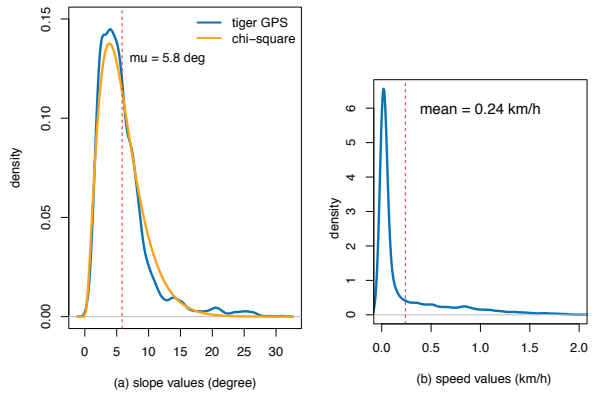


Figure 2. The Kernel density plots of slope values used by the tiger and tiger speed.

Figure 3 compares results of three simulations of 2-hour-long tiger trajectories with different behaviors: (a-b) non-patrolling and (c) patrolling, in different parts of the tiger's home-range. The simulations use two GPS observations, start-point (green) and end-point (red), and the DEM (30-meter) of the home-range. The simulation procedure (i.e. calculation of PTRs and local movement choices) at 10-minute *step times* is presented in Figure 3 (top row). Figure 3 (bottom rows) shows the resulted trajectories of three simulations. Although not used in the simulation, a control GPS mid-point (at hour 1) is marked (magenta) to test whether the simulated track hits the control point or not. Since the model follows a stochastic process, the three simulations result in different paths for each trajectory. The fact that the simulation often passes through or near the control point is promising.

4. Conclusions

This paper introduced a new context-sensitive spatiotemporal simulation model for movement. The model integrates behavior and contextual factors such as geography, spatial constraints, and environmental derivers of local movement choices in modeling trajectories. Although in this study only one environmental variable (i.e. slope) is considered, the model can be extended to include multiple contextual factors. The model can be used in a Monte Carlo approach to create a probability surface representing the probability of visitation of an area by the moving individual. And hence it can be compared to time-geography and Brownian Bridge models. Compared to similar approaches, the proposed simulation not only considers movement capacities of the moving individual and spatiotemporal constraints through time-geography, it also models the influence of the environment on local movement patterns. Future work will focus on the validation and extension of the model for simulating longer trajectories with different behavioral modes at multiple spatial and temporal scales.

Acknowledgements

The author thanks Achara Simcharoen (HKK Wildlife Sanctuary) and James L.D. Smith (University of Minnesota) for providing the tiger tracking dataset.

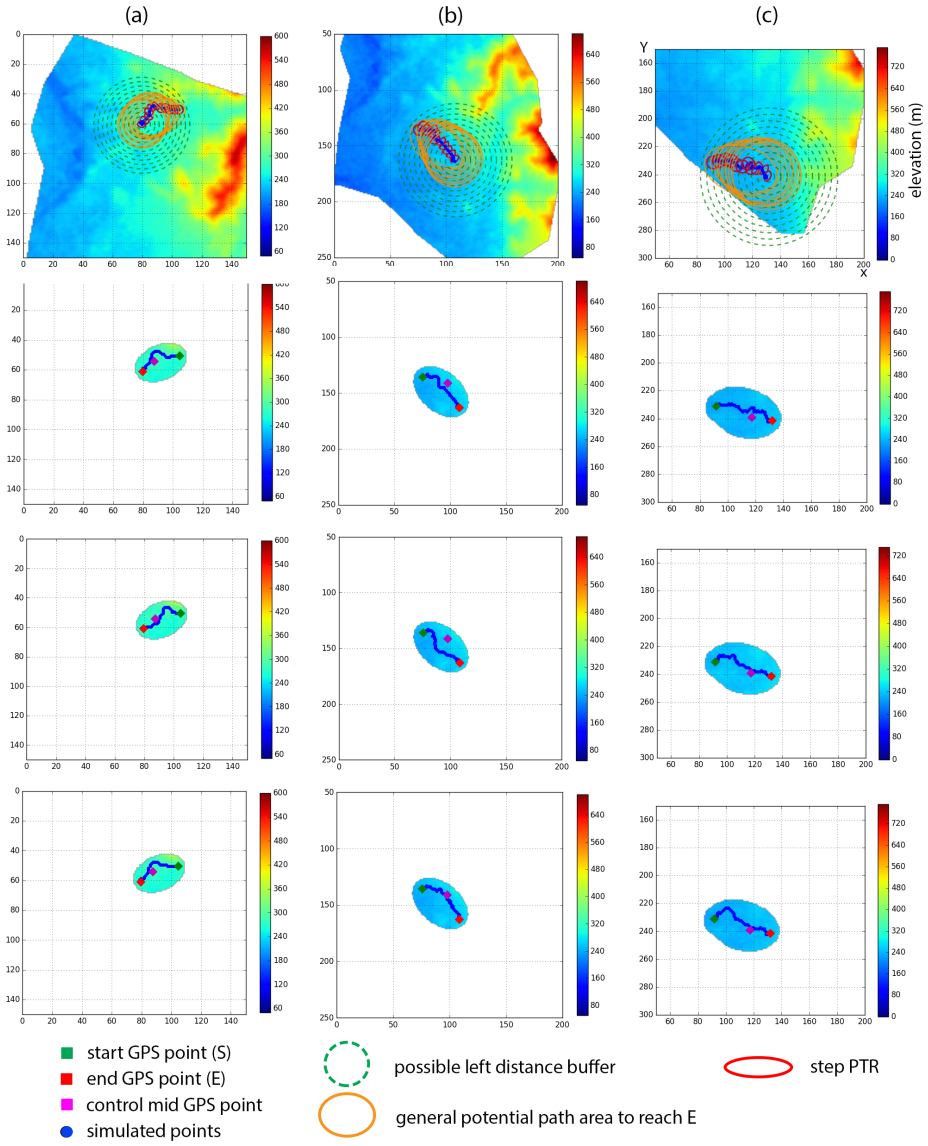


Figure 3. Three simulations for different start-end points and behaviors: the simulation process (top row), and resulted trajectories over the DEM of the potential path areas.

References

Codling, E. A., Plank, M.J., and Benhamou, S., 2008. Random walk models in biology. *Journal of The Royal Society Interface*, 5 (25), 813–834.

Dodge, S., 2016. From Observation to Prediction: The Trajectory of Movement Research in GIScience. *In*: H. Onsrud and W. Kuhn, eds. *Advancing Geographic Information Science: The Past and Next Twenty Years*. GSDI Association Press, 123 – 136.

Horne, J.S., Garton, E.O., Krone, S.M., and Lewis, J.S., 2007. Analyzing animal movements using Brownian bridges. *Ecology*, 88 (9), 2354–2363.

Miller, H.J., 2005. A Measurement Theory for Time Geography. *Geographical Analysis*, 37 (1), 17–45.

Song, Y. and Miller, H.J., 2014. Simulating visit probability distributions within planar space-time prisms. *International Journal of Geographical Information Science*, 28 (1), 104–125.

Technitis, G., Safi, K., and Weibel, R., 2015. From A to B, randomly: An endpoint-to-endpoint random trajectory generator for animal movement. *International Journal Geographical Information Science*, 29 (6), 912–934.

Spatial Preposition Use in Indoor Scene Descriptions

S. A. Doore, M.K. Beard, N.A. Giudice

[1]Spatial Informatics, School of Computing and Information Science, University of Maine
Email: {stacy.doore@maine.edu, beard@spatial,maine.edu, nicholas.giudice@maine.edu}

Abstract

In order to provide accurate automated scene description and navigation directions for indoor space, human beings need intelligent systems to provide an effective cognitive model. Information provided by the structure and use of spatial prepositions is critical to the development of accurate and effective cognitive models. Unfortunately, the use and choice of spatial prepositions in natural language is extremely varied, presenting difficulties for natural language systems attempting to provide descriptions of indoor scenes and wayfinding directions. The goal of the present study is to better understand how humans use spatial prepositions to communicate spatial relationships within virtual environment (VE) indoor scenes. A series of experiments investigates spatial preposition use and the influence scale, topology, orientation and distance within indoor scene descriptions and preliminary results are reported.

1. Introduction

Humans perceive and represent information differently for indoor spaces than they do for outdoor spaces (Guidice, Walton and Worboys 2010). Indoor spaces usually lack established distance and direction metrics, global landmarks, explicit route-networks, and cover a range of spatial scales from small rooms to large airports (Winter 2012). Our interest is in creating an indoor spatial description system to assist navigation in indoor environments. This research investigates natural language (NL) structures for describing objects and structural features within *vista scale* (Montello 1993) virtual environment (VE) indoor scene descriptions. We examine how spatial preposition choice may vary in different contextual settings and which spatial prepositions might yield more effective spatial representations with an increased ability to support spatial behaviors (e.g., object location and navigation).

This paper describes early work investigating how spatial factors such as room size, scene elements, and object/structure relationships impact spatial preposition use and semantics for indoor scene descriptions. Previous research has found these same variables of topology, scale, orientation and distance impact spatial preposition use in geographic and table top spatial settings. The current study focuses specifically on vista scale VE scenes (i.e., 10'by 12' and 20' by 30') that one might encounter in common built environments (e.g., home, business, or school size rooms).

2. Motivation

Consider two spatial expressions describing the same indoor scene:

Speaker 1: *"The bookcase is against the wall, beyond the table, just to your left."*

Speaker 2: *"The bookcase is to the left of the window, directly in front of you."*

In both cases, the speakers use the bookcase as the focus of the spatial description, however, the statements differ in how the bookcase is spatially situated in the scene. The first speaker uses the wall as the primary spatial landmark, whereas, the second speaker uses the window.

The spatial prepositions used by the speakers also differ. Although both expressions are semantically correct, the differences in the linguistic structure of the sentences will yield different cognitive maps for someone trying to mentally reconstruct the scene solely based on each individual description. For most people, scenes are perceived visually. However, in the absence of visual support, communicating a spatial description is an error prone process with a high probability of information uncertainty (e.g., vagueness, ambiguity, inaccuracy, etc.).

Humans construct cognitive maps, or allocentric, global representations of space, that are specialized to individual needs, sensory capacities, and tasks that support spatial inference as well as route planning and way-finding (Downs and Stea 1973). While many physical maps and verbal scene/route descriptions leave out or distort spatial detail, this tends to be the same information that is also omitted/distorted in cognitive maps, such as metric information about distance and direction (Tversky 2001). The goal of this study is to better define the types of descriptive spatial detail necessary to fill-in the perceptual cues supporting accurate cognitive map construction of indoor scenes.

3. Methods

A series of three experiments were conducted to investigate the contextual factors of spatial preposition use for a meets/touches relation within a highly controlled virtual environment (VE). The VE scenes allowed for a more controlled environment for isolating the specific scene elements of interest (topology, orientation, distance). An earlier study on this topic found no significant difference between real world and virtual world scenes descriptions (Kesavan and Giudice 2012). The experiments address the following questions in order to better define formal rules for a future indoor spatial description generation system:

- What are minimally descriptive but sufficiently effective NL descriptions for specifying spatial relations between entities within indoor environments when considering different types of context dependencies?
- What are critical factors that influence spatial prepositions choice/preference within different scale indoor environments (i.e., room-size, object-type, orientation, distance)?

One hypothesis of this study is that underspecified spatial prepositions (e.g., on, in, at, by) can be used effectively as spatial information communication short cuts in verbal indoor scene descriptions because for sighted users, unlike their blind or low BLV peers, a high level of verbal spatial detail is unnecessary to create effective cognitive maps. These simple spatial prepositions are the first to emerge during language acquisition between the ages of 1-5 and are overgeneralized to represent a variety of spatial relationships at this early language development stage (Clark 1973). We believe that simple spatial prepositions can be effectively overgeneralized in indoor scene descriptions in much the same manner as other types of spatial details are reduced in outdoor geographic settings (e.g., GPS route directions) in order to reduce the cognitive load of the description recipient in constructing a minimally descriptive but sufficiently effective cognitive map.

In the first experiment, participants were asked to respond to a series of prompts about objects and room structures in VE indoor scenes. The prompts required participants to provide missing spatial prepositions in both text and verbal response formats to describe the spatial configuration of the specified objects (e.g., desk, chair, bookcase) and room structures.(e.g., wall, door, window). Half of the participants were asked to create these short spatial descriptions for a hypothetical person who could see the VE scene themselves and the other half of participants were asked to create the short descriptions for a hypothetical BLV

user. This methodological approach has been used in a similar study investigating NL descriptions of spatial relations in outdoor geographic space (Klippel, Weaver and Robinson 2011).

In the second experiment, participants classified sets of five images of similar indoor scenes into 3 unlabeled groups or 4 pre-determined categories based on their evaluation of the most appropriate spatial preposition to represent the meets/touches or disjoint spatial relations in the VE scenes. This is a similar approach used in previous studies investigating NL representation of scene elements in geographic space (Klippel, Weaver and Robinson 2011). The final experiment asked participants to look at a VE scenes and then evaluate a spatial prepositions used for a meets/touches relation based on three scales: similarity, clarity, and preference (Schwering 2007).

4. Results

Preliminary results suggest a strong preference for the use of *on* for the meets/touches in both verbal and text response sets. Additional room structures were identified in the structured prompt responses, such as "corner" and the "middle" in the descriptions of object-structure relations within the VE scenes. These room structures worked as *containment* structures for objects when a clear *meets/touches* relation was not possible because of a clear disjoint relation with the object and room structure in question. We believe that these types of elementary spatial prepositions are more frequently used to describe object and room structure relations because, in most cases, more precise spatial information contained in more descriptive spatial prepositions (e.g., along, parallel, perpendicular, etc.) is not necessary for those who can access their vision as a primary sensory modality. A full analysis of the results is currently underway and will employ linguistic, machine learning, and cluster analysis techniques.

5. Conclusions

The results of this research will provide new information about the level of semantic precision necessary for describing indoor space in a manner robust enough to support spatial learning, cognitive map development, and spatial behaviors as well as real-time semantic scene searching. Practical applications of this research include more effective NL spatial language formalisms that can be used in the areas of assistive technologies, emergency response, location-based services and homeland security, which all require the accurate communication of fine scale indoor spatial knowledge.

Acknowledgements

This research was supported by NSF grant CDI-1028895.

References

Clark H H, 1973, Space, time, semantics and the child. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press. Coleman.

Downs R and Stea,D, 1973, *Image and the Environment*. Aldine, Chicago.

Guidice, N A, Walton, L A, and Worboys, M A, 2010, The informatics of indoor and outdoor space , (eds.) *ISA 2010.Proceedings of 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness*, ACM, New York, 47-53.

Kesavan S, Giudice N A, 2012, Indoor scene knowledge acquisition using a natural language interface. In Graf C Giudice, N Schmid F (eds) *SKALID 2012. Spatial Knowledge Acquisition with Limited Information Displays*, Kloster Seeon, Germany,1-6.

Klippel A, Weaver C, and Robinson, A C, 2011, Analyzing cognitive conceptualizations using interactive visual environments. *Cartography and Geographic Information Science*, 38,1, 52–68.
Montello D, 1993, Scale and multiple psychologies of space. In Frank A and Campari I (eds) *Spatial Information Theory: A Theoretical Basis for GIS. LNCS* 716, Springer-Verlag, New York, 312-321.
Schwering, A, 2007, Evaluation of a semantic similarity measure for natural language spatial relations. In Winter S Kuipers, B Duckham M and Kulik L (eds), *Spatial Information Theory. 9th International Conference, COSIT 2007*, Melbourne, Australia, Berlin: Springer, 116-132.
Tversky B, 2001, Spatial schemas in depictions. In Gattis M (ed), *Spatial Schemas and Abstract Thought.* MIT Press: Cambridge, 79-111.
Winter S, 2012, Indoor spatial information. *International Journal of 3-d Information Modeling*. 1,1. 25-42.

A comparative study of existing multi-scale maps: what content at which scale?

Marion Dumont[1], Guillaume Touya[1], Cécile Duchêne[1]

[1]Laboratoire COGIT, IGN, 73 avenue de Paris, 94165 Saint-Mandé Cedex, France
Email: {marion.dumont; guillaume.touya; cecile.duchene}@ign.fr

Abstract

This paper presents a comparative study of existing topographic multi-scale maps, regarding relations between display scale and level of abstraction (LoA) of the map content. The general trends in zoom levels distribution across scale and the original patterns in transitions between LoAs are especially highlighted.

1. Objectives

Multi-scale maps are displayed in mapping applications, i.e. websites where a multi-scale navigation in topographic maps is available. Each producer chooses the display scale and the map content for each zoom level. When users zoom in or out, they actually change the displayed zoom level in the multi-scale map.

In some multi-scale maps, the difference of content between two consecutive zoom levels can be strong, partly due to the change of scale. Mackaness (2007) explains that map scale also relates to a level of abstraction (LoA) of the map. It represents the amount of complexity of the map content: which geographic phenomena are represented, and with how much detail? Due to these changes, we believe that general users may have difficulties to recognize the depicted location or the different representations of a same object across zoom levels.

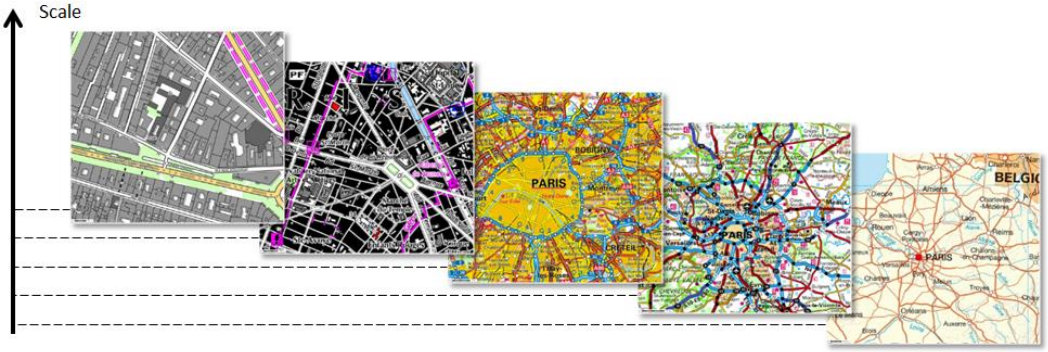


Figure 1. Zoom levels of this multi-scale map (IGN France) present large differences

To build knowledge from multi-scale maps specifications, we study sixteen existing multi-scale maps, provided by national mapping agencies, private companies or collaborative communities. In this paper, we study the correlation between zoom levels, display scale and level of abstraction of the map content, in general (section 2), then focusing on a particular geographic theme: the settlement areas (section 3).

2. How Zoom Levels, Display Scale and Map Content Are Related?

To compare the distribution of zoom levels across scale between multi-scale maps, we first need to define and measure the scale of each zoom level. Besides, most national mapping agencies build their multi-scale map from their topographic paper map series, where each map is designed for a specific printing scale. This map can then be displayed at one or more

zoom levels in the mapping application. We call "definition scale" the initial map scale and "display scale(s)" the scale(s) at which the map is displayed.

2.1 Display Scale

Some cartographic producers explicitly give the display scale in the mapping applications, or show a graphic scale bar. In this case, the display scales have been obtained by measuring the bar length. As this length varies according to the size and resolution of the display screen, we measured it on various screens to check the consistency of the obtained display scales. Although we found some variations, we considered it negligible according to the level of detail of the following analysis.

Comparing the display scales between multi-scale maps, we found that producers generally apply the Web Map Tile Service standard (WMTS). This standard defines a scale set, composed of twenty-one zoom levels, numbered from 0 for the 1: 100 scale to 20 for the 1: 500M scale. For each zoom level, we can thus compare the map content of different multi-scale maps, as their display scales are close (given the map projection approximations).

2.2 Definition Scale

When the definition scale was not mentioned in the mapping application, we obtained it by comparison with the map series of their producer. However, some multi-scale maps have not been built from map series (e.g. OpenStreetMap) and will not be considered in the following graph. Figure 2 represents the relation between definition and display scale of each zoom level (represented as a point) in considered multi-scale maps (differentiated by colour).

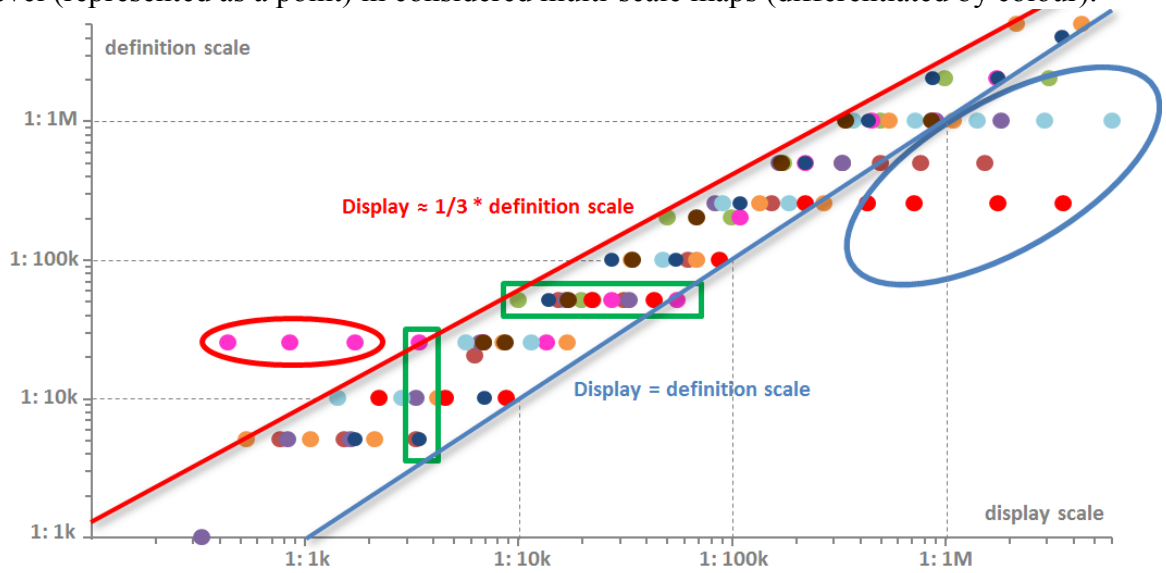


Figure 2. Relations between definition and display scales in considered multi-scale maps

Considering a given display scale (vertical green box) or a given definition scale (horizontal green box), we notice that producers use different relations between definition and display scales. This graph also confirms that many producers use a same map at several zoom levels (same coloured points on the same horizontal line). Multi-scale maps could be improved by adding new representations, specifically designed for these display scales.

We then observe that most zoom levels are concentrated between the two represented lines. According to the red line, most producers do not display a map until the display scale is equivalent to a third of its definition scale. Considering the blue line, most producers do not display a map at a display scale smaller than its definition scale. As the circled outliers present readability issues, we think that these two rules can be considered relevant.

3. Representation of Settlement Areas across Display Scales

The distribution of definition scales across zoom levels gives information about the variation of LoA across scale. However, map content at a same definition scale may differ between producers. For instance, at the 1: 50K scale, some producers represent the individual buildings, whereas others represent urban areas. To compare the representation of settlement areas between multi-scale maps, we define the following LoAs, illustrated from left to right on Figure 3: individual building, urban block, urban area and city point symbol.

Figure 3. Illustration of the four considered LoAs for settlement areas

As generalization operators may be used to refine the LoA of settlement areas, we also observed their use in each zoom level. We noticed four of them, which specifically deal with the LoA of map content: selection, simplification, aggregation and typification. Definitions and use cases of these operators can be found in Regnauld and McMaster (2007).

When two LoAs are present in a same zoom level, we also noticed if there are coexistent, i.e. representing different objects in different areas of the map (depending on the spatial context), or superimposed, i.e. simultaneously representing a same object (Figure 4).

Figure 4. Coexistent (left) and superimposed (right) representations

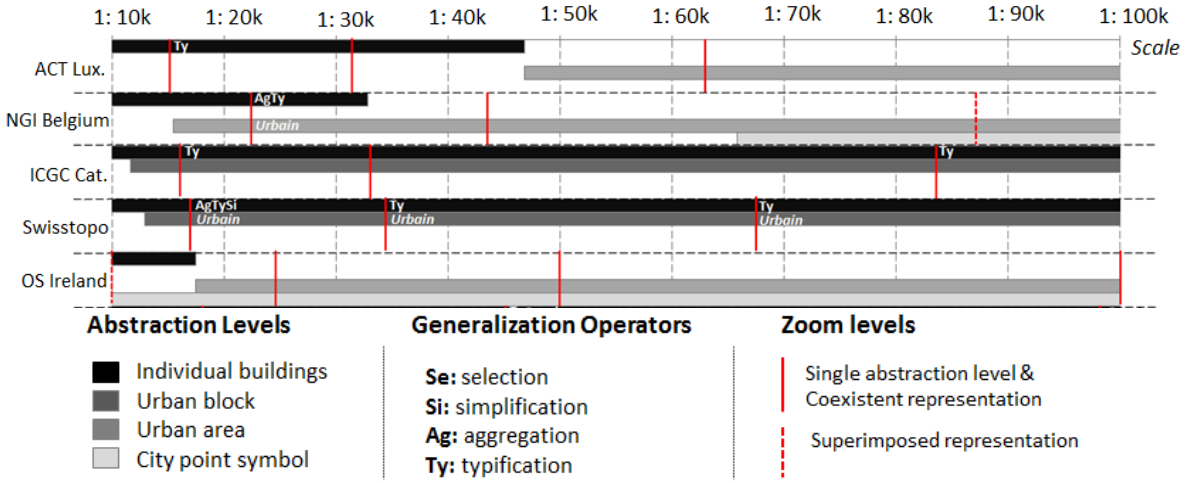


Figure 5. Extract of the representation synthesizing the surveyed information

Figure 5 is an extract of the representation synthesizing the surveyed information, inspired from the ScaleMaster tool (Brewer and Buttenfield, 2007). For each multi-scale map, the use of each LoA on a scale range is symbolized by a grey line. The different shades of grey distinguish the different LoAs. We also added on the graph if different LoAs are used in rural or urban contexts. For each zoom level (red line), coexistent or superimposed representations are identified. If generalization operators are used, their relative code is specified next to the resulting zoom level. Figure 5 shows that, each map producer applies its own variation of LoA across scale.

We analyzed the percentage of use of LoAs across scale, and found some general trends, which are represented in Figure 6. A scale range of common use (in red) could be observed for individual buildings and urban areas. This figure also confirms the use of coexistent and superimposed representations, but also the existence of different strategies used by map producers concerning the relations between LoAs and scales.

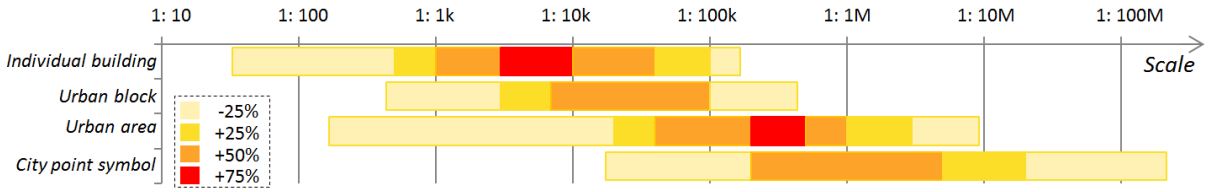


Figure 6. Use percentage of LoAs across scale in studied multi-scale maps

4. Conclusion and Perspectives

The study of zoom levels distribution across scale shows the common use of the WMTS standard. It also highlights rules about the relations between definition and display scale, ensuring the maps readability. Regarding the variation of LoA across scale, we highlight the heterogeneity of relations between LoAs and scales. We also discovered interesting patterns, as the superimposed representations, which could serve as intermediate representations between two LoAs and maybe help ease the navigation across scale.

We will thus now use the identified interesting representations and transitions between LoAs, to add intermediate representations in an existing multi-scale map. Then, we will evaluate their potential improvement for navigation across scale. Assuming that map visual complexity is a part of the problem, in an ongoing study we compare the variation of visual complexity in multi-scale maps, with visual clutter measures (Dumont *et al*. 2016).

Moreover, as reading a map is a human process, we also want to realize a user evaluation. We will measure user task performances, conducted on multi-scale maps with different intermediate representations, to identify the ones improving user navigation across scale.

Acknowledgements

This work is supported by the French National Research Agency, as part of the MapMuxing project [ANR-14-CE24-0011-01].

References

Brewer C, Buttenfield B, 2007, Framing Guidelines For Multi-Scale Map Design Using Databases At Multiple Resolutions, *Cartography and Geographic Information Science*, 34(1): 3-15

Dumont M, Touya G, Duchêne C, 2016, Assessing the Variation of Visual Complexity in Multi-Scale Maps with Clutter Measures, *Proceedings of the AGILE workshop on "Automated generalisation for on-demand mapping" and 19th ICA workshop on Generalisation and Multiple Representation*, Helsinki, Finland

Mackaness WA, 2007, Understanding Geographic Space. In: Mackaness W, Ruas A, Sarjakoski T (eds), *The Generalisation of Geographic Information: Models and Applications*. Elsevier

Regnauld N, McMaster R, 2007, A synoptic view of generalisation operators. In: Mackaness W, Ruas A, Sarjakoski T (eds), *The Generalisation of Geographic Information: Models and Applications*. Elsevier

"The Ridge Went North": Did the Observer Go as Well? Corpus-driven Investigation of Fictive Motion

E. Egorova[1,2], G. Boo[1], R.S. Purves[1]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
Email: {ekaterina.egorova, gianluca.boo, ross.purves}@geo.uzh.ch

[2] University Priority Research Programme Language and Space (URPP SpuR), University of Zurich, Zurich, Switzerland

Abstract

Fictive motion ("The ridge went north") can refer to both dynamic (observer is moving) and static (observer is visually scanning) scenes. Using a corpus of alpine narratives, we extract fictive motion constructions and compare those representing static and dynamic scenes. According to our findings, some of the verbs appear exclusively in static (or dynamic) scenes, while others can be found in both and thus require broader context for the correct annotation. The results can be seen as a step towards the automatic identification of the role of geographic objects in text.

1. Introduction

Text corpora are increasingly being recognised as a potential source of rich geographic information. However, extracting information from text requires understanding of the ways in which language encodes space (Talmy 2000) and the scope of spatial information reproduced in certain discourse. Thus, route descriptions have been shown to go well beyond straightforward references to displacement through the prototypical *go* and *turn* (Allen 2000; Denis 1997; Tversky and Lee 1991) and include instructions of positioning and inspection (Allen 2000; Moncla *et al.* 2015), exercising caution or remembering a certain geographic object (Egorova *et al.* 2015), as well as topological information (Denis 1997). An important task in automatic route extraction from text is thus differentiating whether a geographic object is introduced as an element of the actual path or in some other context (e.g. description of a vista at some point of a path). One approach to making this distinction uses verb semantics, where a verb of perception signals description of the geographic object as part of a scene, while a verb of motion indicates movement of the observer with regard to that object. (Moncla *et al.* 2015).

From this perspective, fictive motion (FM) is of central importance. It depicts "the form, orientation, or location of a spatially extended object in terms of a path over the object's extent" (Talmy 2000) and reflects the conceptual primacy of named objects and their configuration in physical space (Matlock and Bergmann 2014; Matsumoto 1996). Crucially, it can take one of the two forms (Langacker 2010; Matsumoto 1996). The first (static) construes the scene as static, observable from a specific point in space, its conceptual roots lying in visual, or mental, scanning by an observer along the feature ("The path rises quickly near the top." (Langacker 2010)). The second (dynamic) encodes the actual motion of the observer, where the series of immediate fields of view along their path are construed as a single entity moving through space itself ("The path is rising quickly as we climb." (Langacker 2010)). Distinguishing between these forms is important, since one encodes the movement of the observer, while the other describes the configuration of an object in space.

In this study, we explore FM in a corpus of alpine texts addressing the following questions:

- Which verbs are used in FM constructions in our corpus?

- Can we identify differences between the way in which static and dynamic scenes are encoded in FM?

2. Corpus and Methods

Our corpus contains 1'484 texts (6'356'455 words) of the digitized Alpine Journal from 1968 to 2008 (Bubenhofer *et al.* 2015).

As seeds for FM extraction, we compiled a list of 125 nouns from climbing and mountaineering glossaries[1]. Although FM is associated with linearly extended entities, limiting the list to such features seemed inadequate, since it is the FM expression itself that construes the feature as linearly extended ("A table runs along the wall." (Matlock and Bergmann 2014)). Verbs of motion occur with both "structural" (e.g. *ridge, gulley, crack*) and "functional" (e.g. *route, pitch, line*) entities, we thus included both (Klippel 2003).

Using CQPWeb (Hardie 2012) to query a POS-tagged corpus by a lemma and a part of speech, we retrieved nouns in our list followed by a verb in the past or present tense. Among the candidate phrases we identified instances of FM (based on the verb semantics) and annotated those as static or dynamic. Markers for static types include verbs of perception ("we could see that the ridge rose in 4 steps"), locatives referring to the observer-centric field of view ("to the left, the face plunged"), the scale of the scene ("the range runs 1400 km") and, often, a verb's present tense for topological information. For the dynamic types, they include explicit references to the observer ("the ledge led us"), or implicit references to their movement (difficulty: "the pitch went easily"; time: "the ridge went forever"; past tense of verbs: "the crack ran to the top").

3. Results

Our initial query returned 6'530 phrases, of which 1'057 were FM; 655 of these were dynamic and 402 static. The range of verbs is large (81), reflecting the rich inventory language has for encoding FM. The specifics of corpus is reflected in verticality-related verbs (*rise, fall, drop, ascend, plunge, arise, mount, plummet, sink, shoot up*) and verbs semantically rich in the geometry of path (*curve, snake, wind, curl, roll, sneak, swing, weave, zigzag*). Verb frequency is Zipfian with 44 occurring only once or twice.

In the following, we focus on the top quintile represented by the 22 most frequent verbs (Figure 1). Few verbs appear exclusively with static or dynamic scenes, which in some cases seems to be determined by their semantics. Thus, *soar* has a strong connotation appealing to the verticality of a feature and the sensation of seeing it; *stretch* implies vast horizontal extension and is used for communicating general spatial knowledge ("the ridge stretches 8 km north"). Among the dynamic constructions, *bring* implies the presence and displacement of the second object. The majority of verbs, however, demonstrate semantic compatibility with both types of constructions.

[1]https://en.wikipedia.org/wiki/Glossary_of_climbing_terms, http://climber.org/data/glossary.html, http://www.ukclimbing.com/articles/page.php?id=33, http://www.summitpost.org/glossary-summit-peak-etc/173401, http://www.sierradescents.com/glossary, http://www.mountainzone.com/glossary_a_l.html

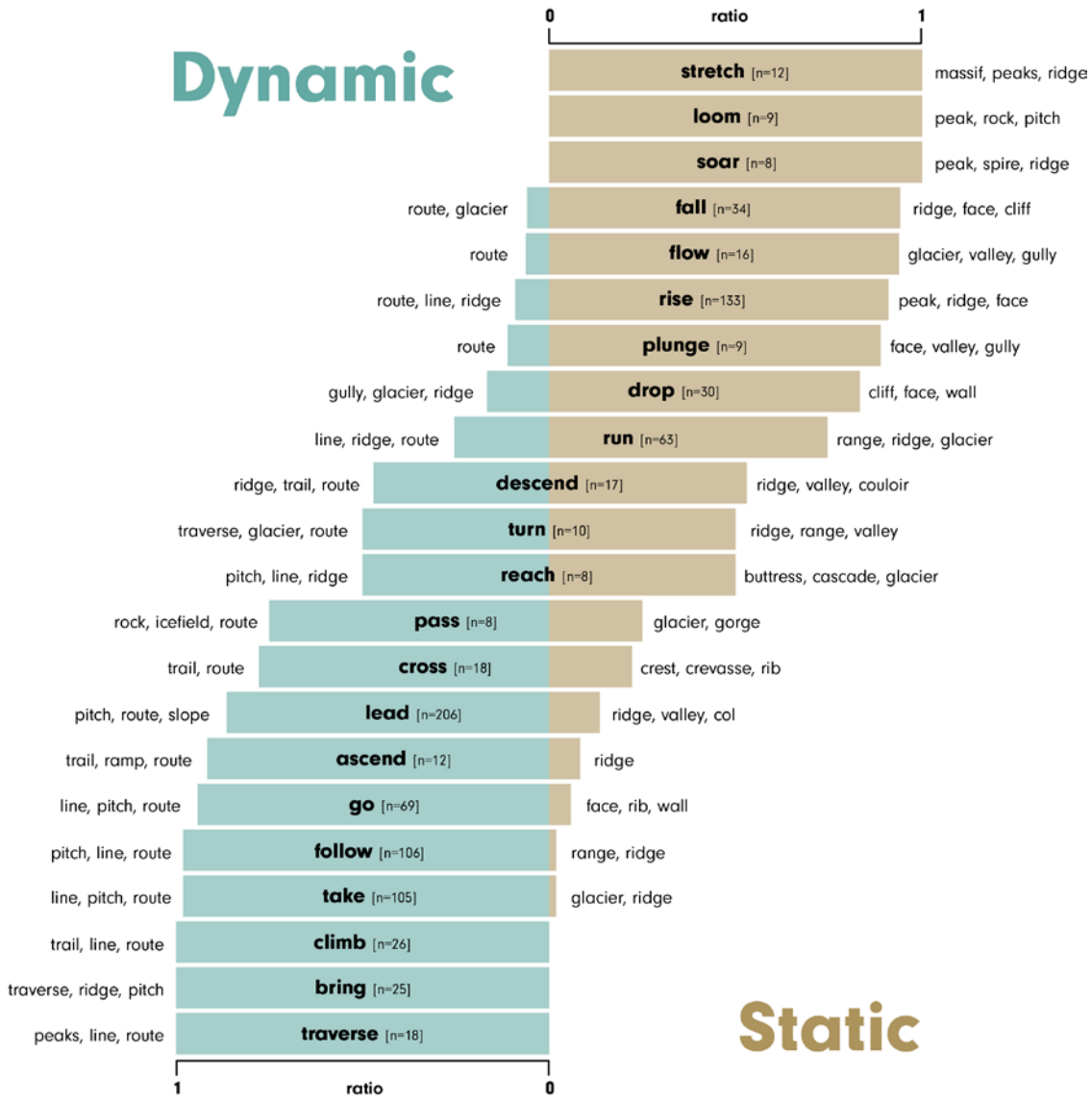


Figure 1. Ratio of static and dynamic uses of the 22 most frequent verbs and their three most frequent noun collocates.

When we explore frequent noun collocates with verbs the predominance of structural terms (e.g. *cliff, valley*) in static scenes becomes clear. In dynamic scenes, certain verbs clearly prefer functional collocates (e.g. *go, follow, take* with *line*, *pitch*, *route*), while others appear with both structural and functional terms. Thus, the same collocates (e.g. *ridge* and *descend*; *ridge* and *rise*) may encode both dynamic and static scenes and can only be annotated through surrounding context.

4. Conclusion and Outlook

Our corpus-driven exploration of FM in alpine narratives reveals a number of points important for further work. Using a set of nouns as seeds, we were able to extract a rich set of verbs used in FM. By developing a set of annotation rules linked to the context of collocates, it was possible to annotate these as either dynamic or static forms of FM. We found a tendency for some verbs to appear in dynamic (or static) scenes exclusively, for static scenes – to prefer structural entities,

which provides some indications as to possible features in automatic classification. However, some collocates remain ambiguous and contextual information (e.g. the markers we used in our annotation scheme) is required. In ongoing work we aim to evaluate the quality of our classification rules through inter-annotator agreement. Furthermore, we will systematically investigate the efficacy of the identified markers in the automatic classification of FM structures as part of our broader agenda for extracting semantically rich spatial information from text.

References

Allen GL, 2000, Principles and Practices for Communicating Route Knowledge. *Applied cognitive psychology*, 14(4):333–359.

Bubenhofer N, Volk M, Leuenberger F, Wüest D (eds), 2015, *Text+Berg-Korpus*. Institut für Computerlinguistik, Universität Zürich, Zürich, Switzerland.

Denis M, 1997, The description of routes: A cognitive approach to the production of spatial discourse. *Cahiers de psychologie cognitive*, 16(4):409–458.

Egorova E, Tenbrink T and Purves RS, 2015, Where Snow is a Landmark: Route Direction Elements in Alpine Contexts. In *Spatial Information Theory: 12th International Conference, COSIT 2015,* Santa Fe, NM, USA, 175–195.

Hardie A, 2012, CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3):380–409.

Klippel A, 2003, Wayfinding choremes. In *Spatial information theory: Foundations of geographic information science*. Springer, Berlin, Germany, 320–334.

Langacker RW, 2010, Dynamicity, Fictivity, and Scanning. In *Grounding Cognition: The Role of Perception and Action in Memory, Language and Thinking.* Cambridge University Press, Cambridge, UK, 164–197.

Matlock T and Bergmann T, 2014, Fictive Motion. In *Handbook of Cognitive Linguistics,* Walter de Gruyter, Berlin, Germany, 546–562.

Matsumoto Y, 1996, Subjective motion in English and Japanese. *Cognitive Linguistics*, 7(2):183–226.

Moncla L, Gaio M, Nogueras-Iso J and Mustière S, 2015, Reconstruction of Itineraries from Annotated Text with an Informed Spanning Tree Algorithm. *International Journal of Geographical Information Science*, 8816:1–24.

Talmy L, 2000, *Toward a Congitive Semantics*. The MIT press, Cambridge, MA, USA.

Tversky B and Lee PU, 1998, How Space Structures Language. *Spatial Cognition*. Springer, Berlin, Germany, 157–175.

Stress Supports Spatial Knowledge Acquisition during Wayfinding with Mobile Maps

P. Frei[1], K.-F. Richter[1], S. I. Fabrikant[1]

[1]Geography Department, University of Zurich
Winterthurerstrasse 190, 8057 Zurich, Switzerland
{patrice.frei|kai-florian.richter|sara.fabrikant}@geo.uzh.ch

Abstract

We detail a novel empirical approach with in-situ psycho-physiological measurements to assess how stress influences spatial knowledge acquisition, mobile map use, and wayfinding success when performing a navigation task in an unknown urban environment. We recorded pedestrians' navigation trajectories, mobile map interactions, eye movements, and galvanic skin responses, with varying stress conditions. Our results clearly indicate that stress supports navigators in forming good survey knowledge, possibly due to enhanced engagement. This seems to emerge from their goal-oriented interaction with the mobile map. Our study results contradict earlier findings, contribute to the on-going debate whether using mobile navigation systems are harmful for humans' capacity to acquire environmental knowledge during navigation, and highlight the important influence of people's individual spatial abilities and emotional states on their knowledge acquisition.

1. Introduction

Does spatial knowledge acquisition deteriorate when people rely on mobile navigation assistance, as prior studies suggest (e.g., Gardony et al. 2013)? If yes, does it relate to disengagement from the navigated environment (Leshed et al. 2008), and/or from the wayfinding decision-making process (Bakdash et al. 2008)? These questions motivated our study. As increasing empirical evidence suggests that navigation performance can be predicted by varying individual differences, for example, spatial abilities (Hegarty et al. 2002), personality traits, such as anxiety (Thoresen et al. 2016), or emotional states, such as stress (Wilkening and Fabrikant 2011), we wondered how individual differences might interact with spatial knowledge acquisition and mobile map use.

2. Methods

Thirtyfive members (f:2; m:33) of the Swiss Armed Forces International Command (SWISSINT) participated in our study. Before the experiment participants completed a demographic questionnaire, the Santa Barbara Sense of Direction Test (in German), the perspective-taking/spatial orientation test, and the building memory test. We divided participants into two ('stress' | 'control') groups by median-split on their perspective-taking test results, such that each group contained the same number of 'high' and 'low' performing participants.

Participants were asked to navigate from a start to an end location along five waypoints in a given sequence, in an environment unfamiliar to them. They were given a mobile map application running on a tablet computer, which displayed start and end locations, the waypoints, and participants' current (GPS) position, but no prescribed routes.

Participants wore a mobile eye tracker connected to a laptop carried in a backpack, and a wrist-band[1] that recorded various psycho-physiological signals (e.g., galvanic skin responses

[1] http://bodymonitor.de/smartband/

(GSR)). The stress group participants also wore in-ear headphones. Each participant filled in the short stress state questionnaire (SSSQ) before and after the experiment to assess changes in their perceived feelings and stress-related attitude. To induce stress, the stress group was given:

1. A time limit of 14 minutes (assessed in pilot tests), indicated by a countdown timer on the mobile map;
2. A work-related search and rescue scenario (fellow soldiers missing in action);
3. A random mix of annoying sounds and disturbing noises (e.g., shooting, loud music) through the headphones;
4. On screen dialog messages and/or vibrations at irregular intervals, to be removed by pressing 'OK'.

The control group had no time limit, but was told to reach the end point in about 15 minutes. They were asked to do a simple patrol walk while checking each waypoint in the same order as the 'stress' group.

After reaching the end point, participants responded to several questions to quantitatively assess their acquired spatial knowledge. Here, we only report results regarding survey knowledge, which we measured by asking participants to point back to the starting point and to each of the waypoints, and by having them estimate the distance to these points.

3. Results

Based on the SSSQ responses and the GSR measurements, the stress group appears to have been more stressed than the control group. Comparing their answers for the SSSQ before and after the navigation task, the distress score for the stress group shows a slight increase (μ=.05), while that of the control group decreased (μ= –.39). Similarly, GSRs of the stress group (μ= –188.972%) increased more than that of the control group (μ= –52.591%). However, GSR differences are not statistically significant, likely due to the large variance between participants.

The average direction estimation error shows no statistically significant difference between the two groups ('stress': 27.14º, 'control': 32.7º). However, the stress group estimated directions more consistently than the control group. In the former, errors are about the same for each point; in the latter average error for the first two estimates is over 50º, then dropping to that of the stress group for the third to fifth estimate (Figure 1).

For each participant, we calculated Fisher r-to-z transformed Pearson correlation coefficients between real-world and estimated distances for each endpoint-waypoint pair. This indicates participants' distance estimation consistency. There is no statistical difference between transformed mean correlation coefficients for the stress group (r=.588) and the control group (r=.70). Moderate coefficients show that the distance estimation task was hard.

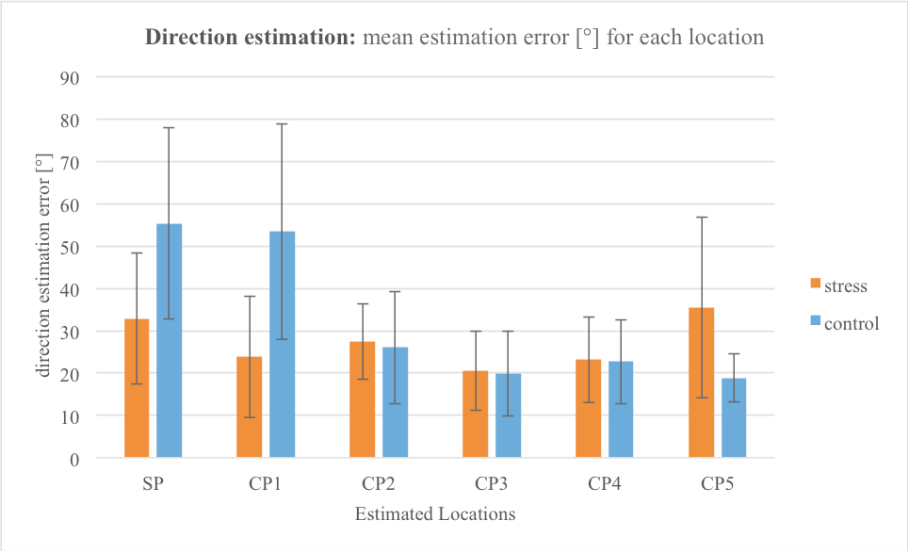


Figure 1: Mean direction estimation error of stress and control group for each location. Error bars show ±2 standard errors.

Based on their direction and distance estimations, we reconstructed participants' 'mental maps' of the environment using a bi-dimensional regression. It provides scaling and rotation factors to compare the true geographic point configuration with the pattern of the participants' estimated locations (Figure 2).

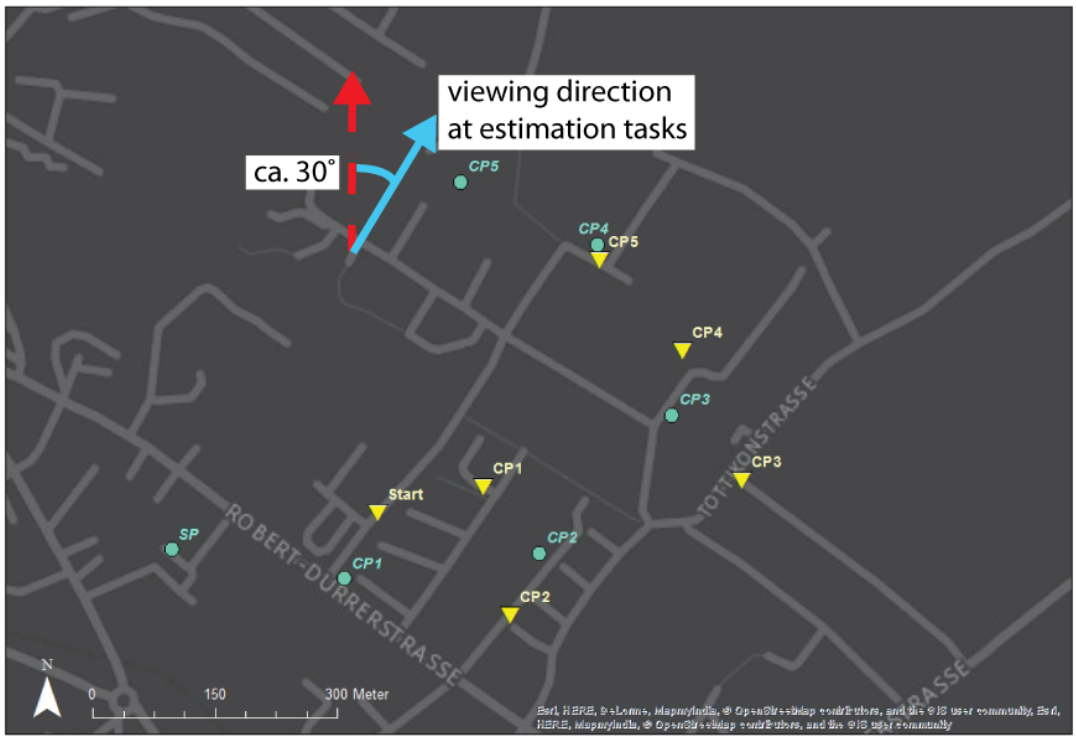


Figure 2: Geographic locations (yellow) and a participant's location estimations (blue; 'SP' corresponds to yellow 'Start'). The North arrow is in red, the participant's viewing direction during the estimation task in blue.

The variance explained by the regression model is slightly higher for the control group (R^2=.797) compared to the stress group (R^2=.764), but the predictions of the stress group are more consistent. The stress group shows less distance distortions (ϕ_x=1.183, ϕ_y=.909) in the scaling factor than the control group (ϕ_x=1.604, ϕ_y=1.331). For ϕ_y, this difference is

statistically significant. The rotation factor indicates by how much the actual and estimated locations need to be rotated to align. Rotation is larger for the stress group, but again more consistent, with a clear pattern for counter-clockwise rotation (μ=-36.91, sd=6.72º) while the control group shows no clear pattern (μ=-.56º, sd=10.3º).

As expected high-spatial participants perform better than low-spatial participants in all tasks, but the difference is only statistically significant for distance estimations and for R^2.

4. Discussion

We find qualitative differences in spatial knowledge acquisition due to stress, but not as expected by prior research. Participants acquire a good level of survey knowledge despite stress, and construct a consistent mental map that we deem better than those of the control group. This is evidenced by their consistency in the direction estimation task, the regression model's scaling factors being close to 1, and the rotation angle approximately corresponding to the angle that would align the street network with the cardinal directions (Figure 2).

The stress group spent more time looking at the mobile device, as qualitative assessments of the eye-tracking data suggests. Participant gazes in the stress group seem to cover a wider area of the map compared to those of the control group, and stressed participants seem to have processed more of the information presented on the map than control group participants. While the control group seems to have exhibited the well-observed passive consumption of information presented by the mobile device, it seems that the stress group perused the map effectively for planning ahead, as to navigate the waypoints most efficiently.

To conclude, our study contributes to the on-going debate of the (potentially detrimental) effects of mobile navigation assistance on our engagement with and knowledge acquisition from the environment. Contrary to most prior work, we find that there are conditions in which environmental learning takes place, even during navigation under stress. An individual user's ability and their emotional states as so often seem to have a significant impact on solving spatial tasks, which calls for further research identifying the influence of these important factors.

References

Bakdash, J.Z., Linkenauger, S.A., Proffitt, D., 2008. Comparing decision-making and control for learning a virtual environment: Backseat drivers learn where they are going. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2117-2121.

Gardony, A.L. et al., 2013. How navigational aids impair spatial memory: Evidence for divided attention. *Spatial Cognition & Computation* 13(4):319-350.

Hegarty, M. et al., 2002. Development of a self-report measure of environmental spatial ability. *Intelligence* 30(5):425-447.

Leshed, G. et al., 2008. In-Car GPS navigation: Engagement with and disengagement from the environment. In: *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM. 1675-1684.

Thoresen, J. C. et al., 2016. Not all anxious individuals get lost: Trait anxiety and mental rotation ability interact to explain performance in map-based route learning in men. *Neurobiology of Learning and Memory* 132:1-8.

Wilkening, J., and Fabrikant, S., 2011. How do decision time and realism affect map-based decision making? In: Egenhofer M.J. et al. (eds), *Spatial Information Theory,* Berlin: Springer. 1-19.

Geodemographic travel to work flows into London, UK.

C.Gale and D.Martin

Administrative Data Research Centre England, University of Southampton, Southampton, SO17 1BJ, UK.
E-mail: c.gale@soton.ac.uk; d.j.martin@soton.ac.uk

Abstract

We present a method for using geodemographic classifications to profile travel-to-work flows into and within London, UK. Geodemographics have been used in the past to profile these flows, but have focused on the residential locations of flow origins due to limitations in data availability and appropriate spatial units. The 2011 census in England and Wales introduced new spatial units designed specifically for workplace data, leading to creation of a classification of workplaces and workers. Combining this workplace classification with a residential based equivalent means the flows between each can be calculated. This innovative approach results in a two-way classification which can be interrogated to better understand and simplify the complexities of commuting into and within London, thus exploring travel to work within a global city at fine spatial resolution.

1. Introduction

Travel to work flows enhance our understanding of local labour markets, economic delivery, transport planning, daytime service delivery and general mobility within a population. Traditional sources of these data, such as national censuses, provide detailed outputs (Stillwell et al., 2010), however the level of complexity that accompanies such rich and large datasets makes any attempt to summarise, visualise or interpret the flows challenging. For example, 4.5 million travel to work flows were recorded into and within London in the 2011 census in England and Wales.

A solution to this problem is to reduce the level of complexity by using geodemographic classifications – summary indicators of the social, economic, demographic and built characteristics of small areas. However, conventional geodemographic classifications focus on residential populations, rather than those of the workplace, with the 2011 Area Classification for Output Areas (2011 OAC) (Gale et al., 2016), being one such example. Without an accompanying workplace classification, the scope for geodemographic analysis of travel to work flow data are limited. The creation of the Classification of Workplace Zones for England and Wales, or COWZ-EW (Cockings et al., 2015), using a new workplace geography of England and Wales (Martin et al., 2013) is therefore an important development. Combining the 2011 OAC and COWZ-EW to profile places of work and residence allows for a unique insight into the population flows into London.

2. Data and methods

The 2011 OAC (Gale et al., 2016) is a geodemographic classification of the UK created using 2011 census data for output areas (OAs). OAs are the smallest geographical units for which census residential statistics are available in the UK. The 2011 OAC is the most recent example of a residential based classification based on census data, with previous classifications created based on 2001 (Vickers and Rees, 2007), 1991 (Blake and Openshaw, 1994, 1995) and 1981 (Charlton et al., 1985) UK census outputs. Conversely, COWZ-EW takes advantage of a new type of small geographical units for workplace statistics from the latest census of England and Wales in 2011. These units, known as workplace zones (WZs) (Martin et al., 2013), allowed for a classification based on the characteristics of the workplace population at places of work

to be constructed for the first time in England and Wales (Cockings et al., 2015). The 2011 OAC and COWZ-EW share some input variables, such as employment types. However, these primarily relate to individuals and are required to characterise both residential and workplace locations and hence use different geographical referencing frames. Variables primarily focused on residential characteristics, for example, are absent from the workplace classification.

The 2011 OAC and COWZ-EW are both hierarchical classifications, with the 2011 OAC consisting of 8 Supergroups, 26 Groups and 76 Subgroups; while COWZ-EW consists of 7 Supergroups and 29 Groups. For consistency we use the designated nomenclature of each cluster in our analysis to provide descriptive shorthand labels for the different area types, although the analysis presented here is restricted to using only the Supergroup level of each classification. However, the methods used could also be applied to other tiers in each classification.

The aim of our analysis is not to provide a comprehensive analysis of travel to work patterns into and within London, but rather to demonstrate an alternative approach to established techniques like the use of Travel to Work Areas (TTWAs), often known internationally as Labour Market Areas, (Coombes and Bond, 2008; Coombes and ONS, 2015) to understand the complexities of population flows. To that end, we examine travel to work flows in terms of travel from the 2011 OAC (residential) to COWZ-EW (workplace). This is facilitated by the availability of 2011 census data for England and Wales detailing counts of persons travelling from their OA of residence to their WZ of primary employment. In total there are 26.6 million flows in England and Wales and 4.5 million flows into and within London. It would be interesting to extend this analysis by considering travel to work in terms of time rather than distance, but this is not asked in the census, nor is there information on the composition of multimodal journeys. Estimation of these extra flow dynamics using external data is however a potential extension to the research presented here.

Characterisation of journeys to work can be undertaken in terms of both where a journey originates (the 2011 OAC) and by its destination (the COWZ-EW). A 56-way classification of journeys to work can therefore be created based on the 8 2011 OAC and 7 COWZ-EW Supergroups. To demonstrate the potential of this new data combination we use the example of flows into and within London, UK. London's status as a special settlement within the UK (Petersen et al., 2011) and the net increase of 500,000 people aged between 16 and 74 in employment during the day make it a challenging study of real substantive interest.

3. Results

Figure 1 maps the 2011 OAC (left) and COWZ-EW (right) at the Supergroup level in south east England, with the boundary of the Greater London administrative area shown. While the detailed pattern of individual OAs and WZs is not visible at this level some general observations can be made. Firstly, the Supergroups labelled as rural in both classifications cover the majority of non-urban areas. Secondly, the other clusters in both classifications are essentially urban or suburban in nature, thereby highlighting the towns and cities of south east England. A comparison of the two classifications in London shows differences in the internal subdivision, with the 2011 OAC displaying a broadly concentric pattern of urban and suburban clusters. Conversely, COWZ-EW displays the 'Top Jobs' Supergroup is predominately found in central areas with outer areas being primarily classified as 'Metro Suburbs'.

Table 1 provides a summary of the commuting flows from 2011 OAC Supergroups to COWZ-EW Supergroups located in London. Each cell contains the number of people in employment travelling from all OAs in one 2011 OAC Supergroup located anywhere in England or Wales to all WZs in one COWZ-EW Supergroup found in London, a measure of how this differs from expected and the median distance travelled. Our table effectively provides a new geodemograhic classification of journeys to work. Values in the cells show the total flow

between each pair of groups (mean value 80,799), divergence from uniformity (in brackets) and the median distance travelled for each pairing. Divergence from uniformity is calculated using the observed minus the expected flow, divided by the expected flow, which is the flow that would be expected if residence and work locations were uniformly mixed. Values greater than 1.0 or lower than -0.5 have been highlighted in the table.

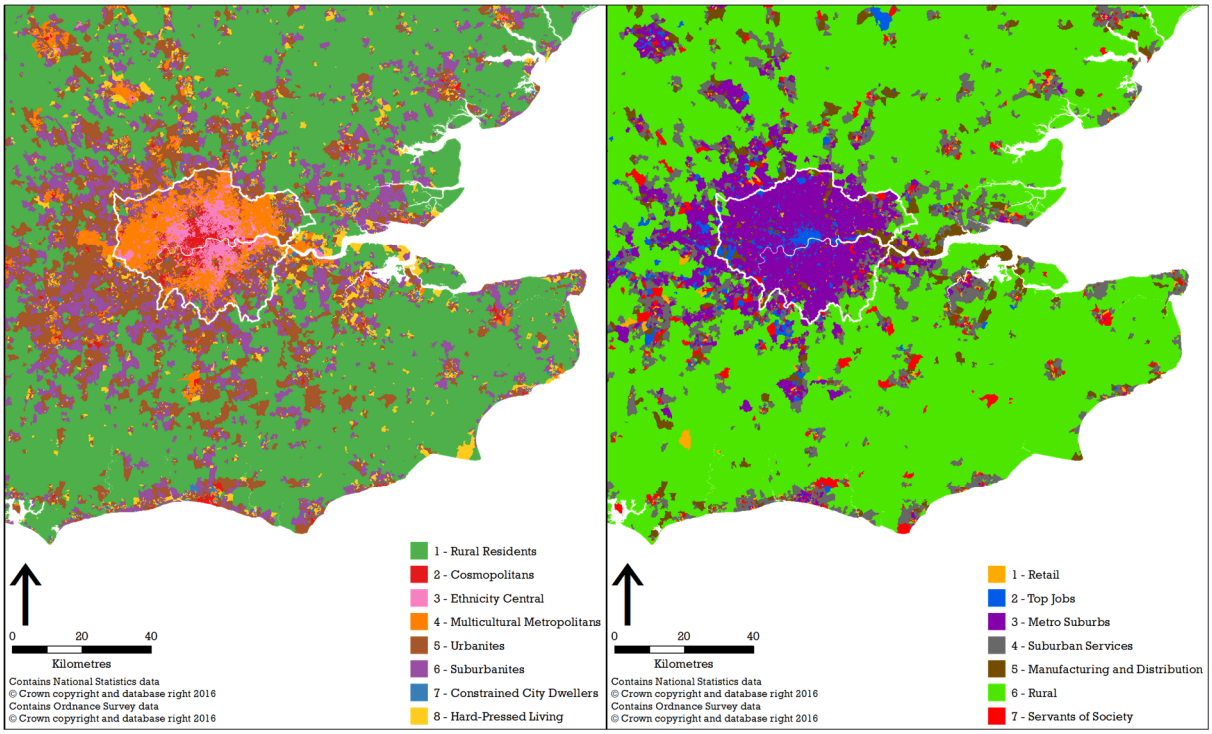


Figure 1. 2011 OAC Supergroups in south east England (left), COWZ-EW Supergroups in south east England (right).

Flows to 'Top Jobs' and 'Metro Suburbs' dominate journeys into and within London, accounting for 90% of all commutes; with the majority of these being close to the expected flow levels. There are notable exceptions to this, such as a larger than expected number of flows occurring from 'Rural Residents' to 'Top Jobs' (0.5), with a median distance travelled of 75.2km; an indication of the willingness and financial resources of some to commute large distances in order to live in rural areas outside of London. In general people appear to be willing further to travel to 'Top Jobs' than 'Metro Suburbs', as evidenced by the lower than excepted flow from 'Rural Residents' to 'Metro Suburbs' (-0.61) coupled with a median distance to travel of 62.3km. The larger distances travelled to reach 'Top Jobs' are likely to a combination of the attractiveness of the jobs in these areas and because the areas where 'Top Jobs' intersect for the most part only with the 2011 OAC Supergroups 'Cosmopolitans' and 'Ethnicity Central'.

An alternative method of analysing the flows between places of home and work is to map residential areas according to the destinations of their primary flows. Figure 2 maps OAs in southern England in terms of the COWZ-EW Supergroup destinations in London of their first, second and third largest commuting flows. The geographic extent in which people commute to London is shown in the first order map. A notable feature is that the majority of OAs across southern England contain individuals who commute into London, with the most likely destination being the 'Top Jobs' Supergroup. The second order map shows the reduced geographic area of flows into London with most travelling to 'Metro Suburbs', although a greater mix of COWZ-EW Supergroups is apparent. The third order map reveals a greater

complexity of more localised flows into and within London, with 'Retail' and 'Manufacturing and Distribution' being the most visible within London.

Table 1. Observed flows of 2011 OAC to COWZ-EW Supergroups in London: count of flows, (divergence from expected), median distance (km)

2011 OAC (rows) and COWZ-EW (columns)	1: Retail	2: Top Jobs	3: Metro Suburbs	4: Suburban Services	5: Manufacturing and Distribution	6: Rural	7: Servants of Society	Total
1: Rural Residents	**732 (-0.59) 32.8**	50,273 (0.5) 75.2	**10,228 (-0.61) 62.3**	332 (-0.41) 17.5	2,638 (0.05) 49.5	**994 (6.32) 0**	1,004 (-0.13) 42.1	66,201 (0.73)
2: Cosmopolitans	**6,508 (-0.62) 3.0**	434,008 (0.37) 6.0	174,394 (-0.3) 0.1	**551 (-0.90) 6.1**	**5,909 (-0.75) 8.1**	**131 (-0.9) 13.6**	**4,396 (-0.6) 4.6**	625,897 (-0.53)
3: Ethnicity Central	24,816 (-0.31) 4.2	675,442 (0.01) 6.3	578,072 (0.09) 0.9	**1,155 (-0.90) 8.8**	30,058 (-0.4) 5.3	**434 (-0.84) 13.0**	12,624 (-0.45) 4.7	1,322,601 (-0.4)
4: Multicultural Metropolitans	48,712 (0.31) 3.8	530,693 (-0.24) 11.8	689,129 (0.26) 1.3	**4,554 (-0.61) 3.6**	74,164 (0.42) 4.8	**1,216 (-0.57) 8.2**	25,542 (0.07) 3.8	1,374,010 (-0.05)
5: Urbanites	20,341 (0.24) 3.7	329,112 (0.07) 23.0	194,417 (-0.19) 3.5	**13,682 (1.66) 0.0**	26,321 (0.15) 11.1	**2,732 (1.2) 1.9**	17,551 (0.67) 4.9	604,156 (0.54)
6: Suburbanites	13,227 (0.31) 4.2	203,980 (0.08) 32.6	109,598 (-0.26) 5.5	**13,290 (3.2) 0.0**	17,855 (0.26) 17.1	**2,667 (2.49) 1.5**	11,592 (0.79) 5.7	372,209 (0.98)
7: Constrained City Dwellers	**3,272 (1.18) 4.3**	24,192 (-0.14) 29.4	18,897 (-0.15) 7.7	**1,632 (2.45) 0.7**	**5,075 (1.4) 11.2**	**357 (2.13) 4.9**	**2,194 (1.27) 5.7**	55,619 (1.16)
8: Hard-Pressed Living	4,759 (0.69) 5.6	49,832 (-0.06) 38.0	31,523 (-0.24) 14.4	**3,299 (2.73) 0.3**	**10,103 (1.55) 17.5**	**753 (2.53) 5.3**	**3,755 (1.08) 9.0**	104,024 (1.18)
Total	122,367	2,297,532	1,806,258	38,495	172,123	9,284	78,658	4,524,717

4. Conclusion

We have combined residential and workplace geodemographic classifications to provide a unique method of exploring the structure of travel-to-work flows. The example of flows into and within London provides an illustration of the different methods of investigation possible, such as exploring the characteristics of flows from home to work, identifying the level of connectivity between different clusters and the distances people are prepared to travel. This has been made possible by the release of origin and destination data from the 2011 census in England and Wales at a granular level as an open dataset. These data are however limited to only providing counts of OAs to WZs due to disclosure control. A potential future avenue of research on this topic would be through the use of microdata to explore the demographic and socio-economic characteristics of individuals. Linking flow data and individual characteristics of commuters with geodemographic classifications would provide a rich dataset that could further enhance our understanding of the complexities of travel-to-work flows.

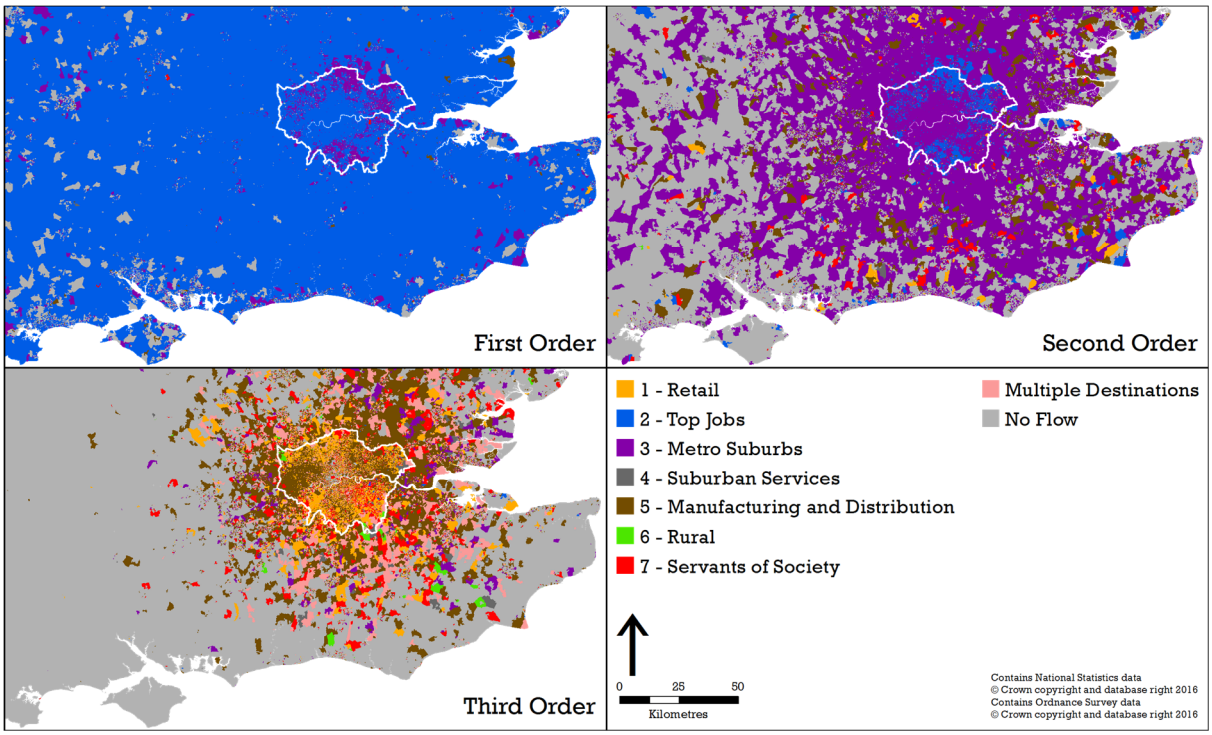


Figure 2. First, Second and Third order dominant COWZ-EW Supergroup destinations in London by output areas

Acknowledgements

The authors gratefully acknowledge the support of ESRC award ES/L007517/1.

References

Blake, M. and Openshaw, S. (1994) *GB Profiles: A User Guide*, Leeds, UK, School of Geography, University of Leeds.

Blake, M. and Openshaw, S. (1995) *Selecting variables for small area classifications of 1991 UK census data*, 95-2, Leeds, UK, School of Geography, University of Leeds, [online] Available from: http://www.geog.leeds.ac.uk/papers/95-2/ (Accessed 12 December 2011).

Charlton, M., Openshaw, S. and Wymer, C. (1985) 'Some new classifications of census enumeration districts in Britain: a poor man's ACORN', *Journal of Economic and Social Measurement*, 13(1), pp. 69–96.

Cockings, S., Martin, D. and Harfoot, A. (2015) *A Classification of Workplace Zones for England and Wales (COWZ-EW)*, Southampton, UK, University of Southampton, [online] Available from: http://cowz.geodata.soton.ac.uk/download/files/COWZ-EW_UserGuide.pdf (Accessed 31 August 2015).

Coombes, M. and Bond, S. (2008) *Travel-to-Work Areas: the 2007 review*, Office for National Statistics.

Coombes, M. and ONS (2015) 'Travel to Work Areas: Research undertaken with and for the Office for National Statistics', Office for National Statistics, [online] Available from: www.ncl.ac.uk/curds/publications/documents/RR2015-05.pdf (Accessed 19 March 2016).

Gale, C. G., Singleton, A. D., Bates, A. G. and Longley, P. A. (2016) 'Creating the 2011 Area Classification for Output Areas (2011 OAC)', *Journal of Spatial Information Science*, 12, pp. 1–27.

Martin, D., Cockings, S. and Harfoot, A. (2013) 'Development of a geographical framework for census workplace data', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2), pp. 585–602.

Petersen, J., Gibin, M., Longley, P., Mateos, P., Atkinson, P. and Ashby, D. (2011) 'Geodemographics as a tool for targeting neighbourhoods in public health campaigns', *Journal of Geographical Systems*, 13(2), pp. 173–192.

Stillwell, J., Duke-Williams, O. and Dennett, A. (eds.) (2010) *Technologies for Migration and Commuting Analysis: Spatial Interaction Data Applications*, Hershey, PA, IGI Global.

Vickers, D. W. and Rees, P. H. (2007) 'Creating the UK National Statistics 2001 output area classification', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2), pp. 379–403.

Exploring the Structure of Contact Networks at Multiple Scales

P. Gao and L. Bian

Department of Geography, University at Buffalo, The State University of New York, Buffalo, NY 14261, United States

Email: {pgao3, lbian}@buffalo.edu

Abstract

This study examines the scale effects on the network structure and spatial structure of contact networks. Regular grids with different cell sizes are used to divide one observed and two reference random networks into multiple scales. Four metrics are used to represent the two structures. Results show that the network structure of the observed network is sensitive to scale changes at fine scales. In comparison, the clustered spatial structure is scale independent.

1. Introduction

Contact networks play a critical role in disease dispersion, as repeatedly stressed in reports on the most dangerous communicable diseases, such as SARS, H1N1, and Ebola (Ferguson *et al*. 2005). Understanding the properties of contact networks helps gain insights into how to prevent and control the dispersion of these diseases (Keeling and Eames 2005).

Disease dispersion is inherently a spatial process, while scale is involved in all spatial phenomena (Wu and Wu 2013). There are rarely studies addressing the scale effects on network properties. Contact networks have both network structure and spatial structure (Barthélemy 2011), while the spatial structure has received attention only in recent years (Bian 2013).

This study aims to evaluate (1) the scale effects on the network structure and spatial structure of contact networks, and (2) the ranges of scale at which contact networks are scale dependent. To achieve these goals, three networks, one observed and two randomly structured, are partitioned into multiple scales using regular grids. The properties of the resultant networks represented by four metrics are compared across scales.

2. Contact Network

This study uses a contact network previously constructed for a metropolitan community in the Northeastern US (Bian *et al.* 2012). Each individual belongs to a family and most belong to a workplace. There are two types of contact relationships between individuals, those between family members and those between co-workers. The network represents individuals as nodes and the contact relationships as edges, resulting in a total of 64,726 nodes and 194,683 edges. The two types of relationships are treated as family and co-worker edges, respectively.

To examine the scale effects, the contact network is projected into space. Nodes are projected according to their home locations. A family edge is within a home location with a zero physical distance. The co-worker edges are between two different home locations and have various physical distances.

3. Methods

3.1 Four Metrics

Three metrics are used to represent the network structure, including the relative size of the largest component *S*, clustering coefficient *cc*, and relative average path length *l*. A component is a cluster of nodes within a network. All nodes are directly or indirectly (through a chain of edges) connected to all other nodes within the cluster but disconnected with nodes in other clusters. *S* is the ratio of the number of nodes in the largest component to the total number of nodes in the network (Newman 2010). A greater *S* indicates a more cohesive network.

The clustering coefficient of a given node is the number of edges between its neighboring nodes divided by the number of all possible edges between the neighboring nodes (Newman 2010). *cc* of the network is the average clustering coefficient over all individual nodes. A higher *cc* means a stronger locally clustered structure.

The path length is the number of consecutive edges between a pair of nodes. *l* is the average length of all shortest paths divided by the shortest path length between the pair of nodes that are most apart in the network (Newman 2010). A shorter *l* implies a more efficiently connected network.

The statistical distribution of edge distance (*Dist*) measures the spatial structure of the network (Barthélemy 2011). A negatively skewed distribution indicates the dominance of short edges, thus a spatially clustered structure, while a normal distribution indicates a spatially scattered structure.

3.2 Three Networks

In addition to the observed network, a random-node network and a random-edge network are simulated to serve as references to the observed network. The two random networks keep the identical number of nodes and edges and identical degree distributions as the observed network.

The random-node network keeps the network structure of the observed network, but alters the spatial structure by randomizing node locations. The random-edge network keeps the spatial structure of the observed network, but alters the network structure by randomly shuffling edges between nodes. Both random networks are generated 1,000 times. The average of the metrics is used to represent the properties of the random networks for the subsequent analysis.

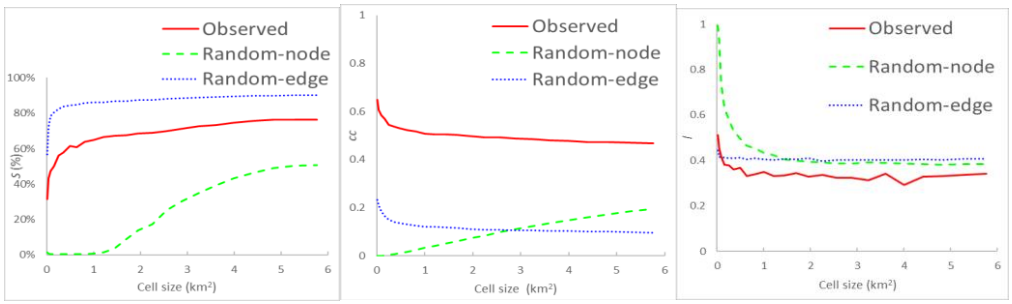
3.3 Division of Networks

The study area is divided into 24 levels of regular grids ranging from 100m x 100m to 2400m x 2400m using 100m increments. During the division, those edges across boundaries of grid cells are eliminated, while those within the cells are kept. In this way, each cell contains its own network, called a ‘unit network’. Properties of the unit networks are calculated using the four metrics and compared across scales and between the three networks.

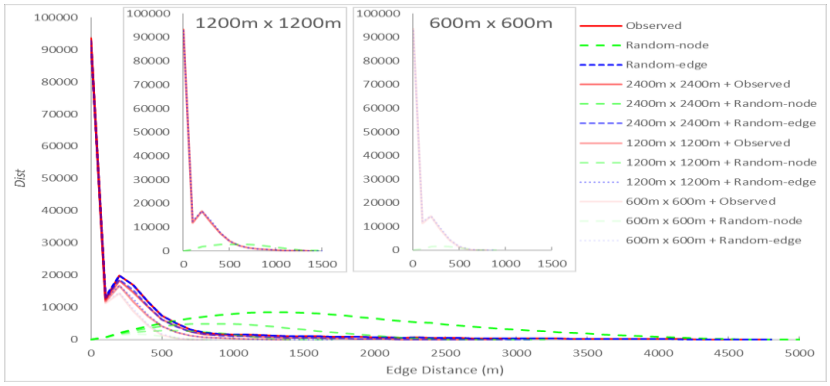
4. Results and Discussion

4.1 Observed Network

The three network structure metrics (*S*, *cc*, and *l*) are plotted against the 24 cell sizes in Figures 1a-c. The spatial structure metric *Dist* for cell sizes of 600m x 600m, 1200m x 1200m, and 2400m x 2400m, are shown in Figure 1d. The three network metrics of the observed network are scale dependent, as their values vary with scale (Figures 1a-c). All three metrics show a characteristic scale of 0.6 km^2. At scales finer than 0.6 km^2, the value of the relative size of the largest component is low and that of the clustering coefficient and relative average path length are high. This indicates the unit networks are fragmented, locally clustered, inefficient, and consequently robust against disease dispersion. Beyond the characteristic scale, the values of the three metrics level off, behaving independently of scale change. The network structure at these coarser scales is more cohesive (higher *S*), remains clustered (similar *cc*), and is more efficient (lower *l'*), thus are more vulnerable to disease dispersion.



(a) (b) (c)



(d)

Figure 1. The four metrics. (a) The relative size of the largest component, (b) the clustering coefficient, (c) the relative average path length, and (d) the statistical distribution of edge distance (including that of the three original networks).

Unlike the network structure metrics, the spatial structure metric, *Dist*, appears to be scale independent. Although absolute quantities of edge distance change, the response of *Dist* to the scale change is invariant. It peaks at 0m that reflects the large number of 0-distance family

edges that are not affected by the division. The second peak between 0-800m is caused by a large supply of short co-worker edges. The 800m diminishing point is equivalent to the characteristic scale of 0.6km^2. Such a short-edge dominant pattern indicates that the unit networks are highly clustered in space. Such networks facilitate short distance disease dispersion and lead to an epidemic surge in small areas.

4.2 Discussion and Conclusions

The two random networks help reveal the unique properties of the observed contact network. In terms of network structure, the observed network is sensitive to scale change at fine scales. In terms of spatial structure, the observed network is scale independent at all scales. Compared to the random-edge network that has a random network structure, the observed network tends to be 'dense', thus more cohesive, clustered, and efficient. Compared to the random-node network that has a random spatial structure, the observed network is spatially clustered.

The network structure and spatial structures are closely related. Spatially, the co-worker edges are mostly short (<800m). This can be attributed to urban dwellers' preference of being in close proximity to schools and workplaces (Ben-Akiva and Lerman 1985). Those who live close to each other tend to go to the same schools or workplaces. In other words, the closer ones are more connected. Findings of this research help deploy intervention strategies to spatially targeted areas.

Acknowledgements

Research reported in this publication was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM108731. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Barthélemy, M. (2011). Spatial networks. *Physics Reports, 499*, 1-101

Ben-Akiva, M.E., & Lerman, S.R. (1985). *Discrete choice analysis: theory and application to travel demand*. Cambridge, MA: MIT press

Bian, L. (2013). Spatial approaches to modeling dispersion of communicable diseases–A review. *Transactions in GIS, 17*, 1-17

Bian, L., Huang, Y., Mao, L., Lim, E., Lee, G., Yang, Y., Cohen, M., & Wilson, D. (2012). Modeling individual vulnerability to communicable diseases: A framework and design. *Annals of the Association of American Geographers, 102*, 1016-1025

Ferguson, N.M., Cummings, D.A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., & Burke, D.S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature, 437*, 209-214

Keeling, M.J., & Eames, K.T. (2005). Networks and epidemic models. *Journal of the Royal Society Interface, 2*, 295-307

Newman, M.E.J. (2010). *Networks: An introduction*. Oxford: Oxford University Press

Wu, J., & Wu, T. (2013). Ecological resilience as a foundation for urban design and sustainability. *Resilience in Ecology and Urban Design* (pp. 211-229). New York: Springer

Identifying Local Spatiotemporal Autocorrelation Patterns of Taxi Pick-ups and Drop-offs

Song Gao[1], Rui Zhu[1], Gengchen Mai[1]

[1]STKO Lab, Department of Geography, University of California, Santa Barbara, USA
Contact Email: sgao@geog.ucsb.edu

Abstract

Analyzing spatiotemporal autocorrelation would be helpful to understand the underlying dynamic patterns in space and time simultaneously. In this work, we aim to extend the conventional spatial autocorrelation statistics to a more general framework considering both spatial and temporal dimensions. Specifically, we focus on the spatiotemporal version of Getis-Ord's G^*. The proposed indicator STG^* can quantify the local association of adjacent features in space and time. As a proof of concept, the proposed method is then applied in a large-scale GPS-enabled taxi dataset to identify local spatiotemporal autocorrelation patterns of taxi pick-ups and drop-offs in New York City.

1. Introduction

Nowadays, large-scale spatiotemporal data (e.g., taxi trajectories, phone call records, social media posts) become available, which provide rich information to support research on human behaviors, transportation, urban landscape, and human-environment interactions. However, discovering patterns hidden in large-scale spatiotemporal datasets is challenging and thus attracts a lot of attention from the GIScience community (Hardisty and Klippel 2010; Demšar and Virrantaus 2010; Scholz and Lu 2014; Claramunt and Stewart 2015).

Spatial autocorrelation statistics, like *Moran's I* and *Getis-Ord's G* are commonly designed for identifying spatial autocorrelation patterns (Fischer and Getis 2009). However, there is a gap in building corresponding measurements for spatiotemporal autocorrelations. For example, although human movements and activities may vary over time across different places, the observed activity hotspots and movement flow might exhibit a pattern of spatial dependence. Also, ignoring the temporal dimension would not be sufficient to discover underlying spatiotemporal dynamics. Therefore, our work aims to contribute to extend the conventional local spatial autocorrelation statistics to include both spatial and temporal dimensions. As a proof of concept, the proposed method is then applied in a large-scale GPS-enabled taxi dataset to identify local autocorrelation patterns of taxi pick-up points (PUPs) and drop-off points (DOPs) in New York City.

2. Methodology

Spatial autocorrelation measurements can be divided into two categories: *global* and *local* indices. Classic global indices of spatial autocorrelation include *Moran's I, Geary's C*, and *Getis-Ord's General G*, while local indices of spatial association (LISA) can be established by transforming the global indices into corresponding local measurements (Anselin 1995). Spatiotemporal autocorrelation concept refers to the relationship between some variable observed in each of space-time settings and the association with its neighbors. In a previous work, Gao (2015) proposed three global spatiotemporal autocorrelation indices but didn't describe how to decompose them into local versions. As an initial trial, this work focuses on extending *Getis-Ord's* G^* (Equation 1) (Getis and Ord 1992) by adding temporal indexes into

the formula, and we name this local spatiotemporal autocorrelation measure as *Spatiotemporal Getis-Ord's G** (*STG**) (Equation 2).

$$G_i^* = \frac{\sum_j w_{ij} x_j}{\sum_j x_j}, \; j \text{ may equal to } i, \quad (1)$$

where w_{ij} indicates the spatial weight between location i and j and x_j is the attribute value at location j.

$$STG_i^* = \frac{\sum_s \sum_t w_{ist} x_{st}}{\sum_s \sum_t x_{st}} \quad (2)$$

where w_{ist} is the extended weighting matrix regarding both spatial and temporal dimensions and x_{st} is the attribute value at space s and time t.

The STG^*_i quantifies the spatiotemporal concentration of adjacent features associated with the target *i*, and works as an indicator for measuring local association in space and time simultaneously. To conceptualize the spatiotemporal neighbors, we implement a 3D-cube framework as shown in Figure 1, where each voxel consists of a geographic coordinate *S(x,y)* and a timestamp *T(t)*. The adjacency can be defined as the first-degree of "Queen" type, in which there are 26 spatiotemporal neighbors for a target voxel in the center cube. The weight w_{ist} for them in calculating STG^*_i is 1 otherwise is 0.

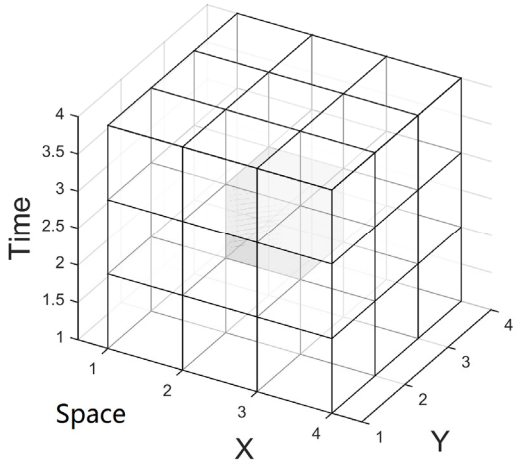


Figure 1. The 3D-cube visualization of spatiotemporal neighbors.

Furthermore, to statistically test the significance of the concentration of either high or low attribute values surrounding the target voxel, the Z-score of STG^*_i, as illustrated in Equation 3, is calculated. Thus, if a tested z-score is significantly different from the expected mean, the target feature would be a hot-spot (z-score > 0) or a cold-spot (z-score < 0). Note that both hot-spots and cold-spots only represent positive spatiotemporal autocorrelation, i.e., where each voxel has a similar value to its neighbors for the same variable.

$$Z_{G_i^*} = \frac{\sum_s \sum_t w_{ist} x_{st} - \bar{X} \sum_s \sum_t w_{ist}}{S \sqrt{\frac{\left[n \sum_s \sum_t w_{ist}^2 - \left(\sum_s \sum_t w_{ist} \right)^2 \right]}{n-1}}} \quad (3)$$

Where *n* is equal to the total number of voxels, and:

$$\bar{X} = \frac{\sum_s \sum_t x_{st}}{n} \text{ and } S = \sqrt{\frac{\sum_s \sum_t x_{st}^2}{n} - \bar{X}^2}$$

3. Case Study: Taxi Drop-offs and Pick-ups in New York City

3.1 Data and Processing

Taxi pick-up and drop-off locations in cities can reveal human movement patterns and thus playing an important role in urban informatics and transportation management. The data used in this study is downloaded from the NYC Taxi and Limousine Commission trip GPS records[1]. We extract one-week trips in five boroughs (*Manhattan, Brooklyn, Queens, The Bronx, and Staten Island*) of NYC from Jan. 3 to Jan. 9, 2015. As shown in Figure 2, by applying exploratory spatial autocorrelation analysis for the whole time period, we can find that three regions are significant "hot-spots" (*Manhattan, JFK International Airport and LaGuardia Airport*) for both PUPs and DOPs. In order to further identify fine-scale local autocorrelation patterns, we spatially filter the original data to include only trips generated in *Manhattan* and there exist 2,548,952 PUPs and 2,462,199 DOPs in total. Figure 3 shows their temporal variations in different hours.

In order to further conduct local spatiotemporal autocorrelation analysis, we need to aggregate those points (PUPs or DOPs) into introduced space-time-cube structure One research question is how to find appropriate bin sizes in both spatial and temporal dimensions for defining neighbors. After calculating the nearest-neighbor distance for each point and the time difference for each pair, we found that spatial proximity is related to temporal closeness. Therefore, we suggest a strategy to find optimal bin sizes for defining spatiotemporal neighbors: Firstly, we spatially aggregate those points into regular grids or administrative polygons; the city block is taken in this study and the spatial bin can be set to one quarter of the average city-block size (about 520 meters) in *Manhattan*. Secondly, the average time difference of temporal adjacent points in each block is calculated and the median (or mean) of average time differences across all city blocks can be used as the temporal bin. Finally, the space-time cubes are constructed with a 130-meter spatial distance and about 20-minute temporal interval in this study. Figure 4 shows the visualization of aggregated PUPs in space-time cubes, in which the attribute value represents the count of PUPs in each voxel.

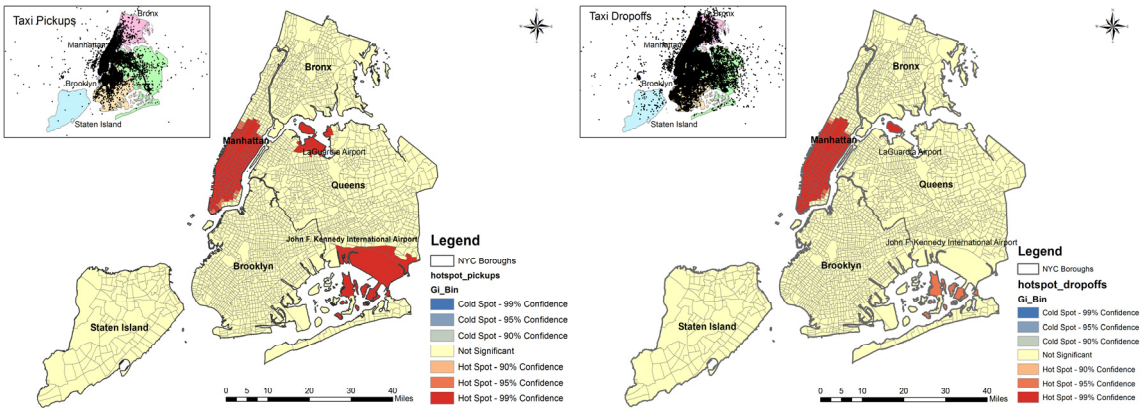


Figure 2. The spatial distributions of PUPs and DOPs and spatial autocorrelation results in NYC.

[1] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

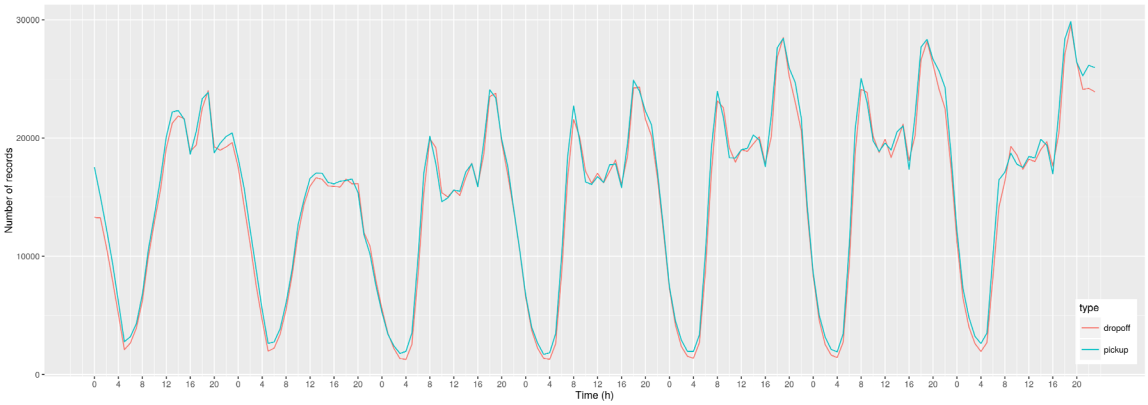


Figure 3. The temporal variations of PUPs and DOPs in different hours in Manhattan.

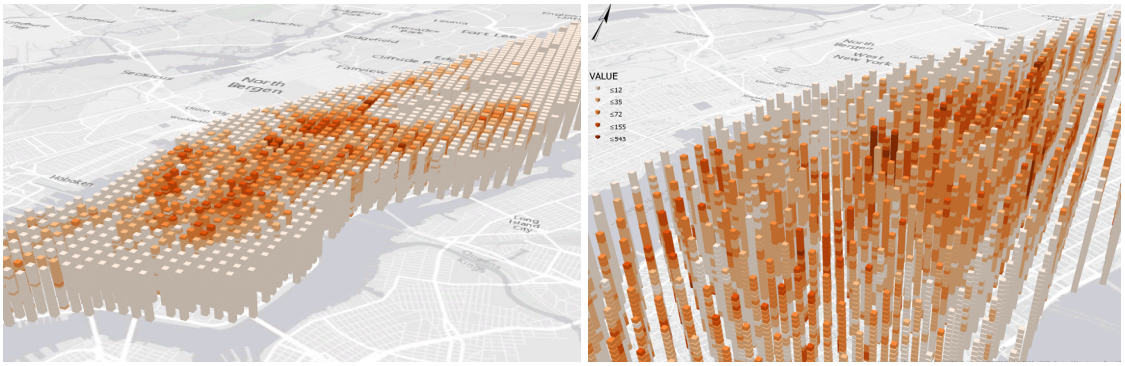


Figure 4. The spatiotemporal visualization of PUPs in Manhattan.

3.2 Results

By applying the proposed local spatiotemporal autocorrelation method, we calculate the STG^* statistic of PUPs (and DOPs) for each voxel and the corresponding z-score. Figure 5 shows different confidence levels (90%, 95%, and 99%) of spatiotemporal "hot-spots" (red color: a large statistic value exists and its spatiotemporal neighbors also have large values) and "cold-spots" (blue color: a small statistic value exists and small values for its spatiotemporal neighbors) for PUPs in *Manhattan*. We find statistically significant local spatiotemporal hotspot clusters in the southern part and cold-spot clusters in the northern part of *Manhattan*. Interestingly, those regions are spatially divided by the *Central Park*. Such spatiotemporal pattern of taxi trips usually links to the mixture land-use structure and human home-to-job activities, which has also been identified by other studies (Liu et al. 2012; Liu et al. 2015).

Table 1. The top ranked local hotspots for taxi pick-ups and drop-offs in Manhattan.

Rank	Pick-ups	Drop-offs
1	LocationID: 9536 Time: 1/8/2015 8:00-8:20 AM	LocationID: 9536 Time: 1/8/2015 8:00-8:20 PM
2	LocationID: 9536 Time: 1/4/2015 10:00-10:20 PM	LocationID: 7725 Time: 1/8/2015 8:20-8:40 AM
3	LocationID: 9535 Time: 1/4/2015 10:00-10:20 PM	LocationID: 7725 Time: 1/8/2015 8:40-9:00 AM
4	LocationID: 9535 Time: 1/4/2015 10:20-10:40 PM	LocationID: 7357 Time: 1/8/2015 9:20-9:40 AM
5	LocationID: 9536 Time: 1/4/2015 10:20-10:40 PM	LocationID: 7725 Time: 1/8/2015 8:00-8:20 AM

In addition, we can zoom to specific voxels in the space-time cubes and compare their local STG^* values. Table 1 shows the top 5 hotspots of taxi PUPs and DOPs ranked by their z-scores of STG^*. It proves the existence of local spatiotemporal autocorrelation patterns.

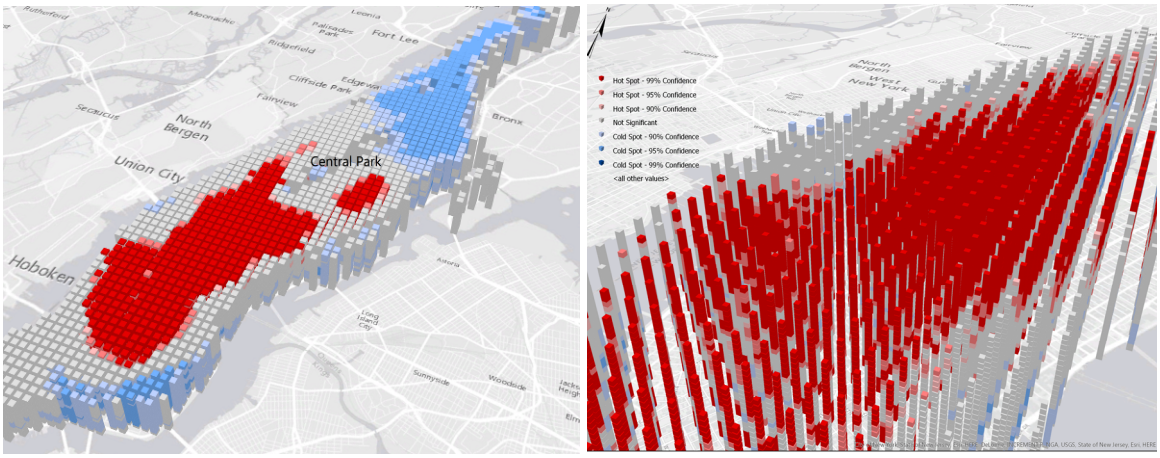


Figure 5. The visualization of spatiotemporal hot-spots and cold-spots in Manhattan.

4. Conclusions and Future Work

In this research, we extend the spatial association statistic *Getis-Ord's* G^* to the local spatiotemporal autocorrelation indicator STG^* which takes the adjacency in both space and time into consideration. The space-time-cube structure has been constructed to support spatiotemporal point pattern analysis and visualization. By performing the proposed method in a large-scale taxi trips, we find that STG^* can sufficiently identify local spatiotemporal autocorrelation patterns of taxi pick-ups and drop-offs in *Manhattan*. The proposed method can also be applied in other event data with spatiotemporal tags and thus has a broad impact.

In future work, more complex spatiotemporal weighting matrix rather than binary ones and its impact on autocorrelation structure will be studied. More empirical studies in other cities will also be conducted to find underlying association between space and time.

References

Anselin L, 1995, Local indicators of spatial association—LISA. *Geographical analysis*, 27(2), 93-115.

Claramunt C and Stewart K, 2015, Special issue on spatio-temporal theories and models for environmental, urban and social sciences: where do we stand?, *Spatial Cognition & Computation*, 15(2), 61-67.

Demšar U and Virrantaus K, 2010, Space–time density of trajectories: exploring spatio-temporal patterns in movement data. *International Journal of Geographical Information Science*, 24(10), 1527-1542.

Fischer MM and Getis A (eds.), 2009, *Handbook of applied spatial analysis: software tools, methods and applications*. Springer Science & Business Media.

Gao S, 2015, Spatio-temporal analytics for exploring human mobility patterns and urban dynamics in the mobile age. *Spatial Cognition & Computation*, 15(2), 86-114.

Getis A and Ord JK, 1992, The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3), 189-206.

Hardisty F and Klippel A, 2010, Analysing spatio-temporal autocorrelation with LISTA-Viz. *International Journal of Geographical Information Science*, 24(10), 1515-1526.

Liu X, Gong L, Gong Y and Liu Y, 2015, Revealing travel patterns and city structure with taxi trip data. *Journal of Transport Geography*, 43, 78-90.

Liu Y, Wang F, Xiao Y and Gao S, 2012, Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning*, 106(1), 73-87.

Scholz RW and Lu Y, 2014, Detection of dynamic activity patterns at a collective level from large-volume trajectory data. *International Journal of Geographical Information Science*, 28(5), 946-963.

What are the Probabilities of Land-Use Transitions? The Answer Depends on the Classification Method

Yulia Grinblat[1, 2], Michael Gilichinsky[3] and Itzhak Benenson[1]

[1]Department of Geography and Human Environment, Tel Aviv University, Tel-Aviv
Email: juliagri@post.tau.ac.il, bennya@post.tau.ac.il

[2]The Porter School of Environmental Studies, Tel Aviv University, Tel-Aviv

[3]Elbit Systems, Rehovot
Email: gilichinsky@gmail.com

Abstract

Based on four time intervals within a 36-year period, we construct Land-Use/Land-Cover (LULC) maps and estimate the transition probabilities between six LULC states: built-up, agriculture, green and open spaces, transportation, and water surfaces. The LULC maps and transition probabilities matrices (TPM) were built based on the manual classification of high-resolution aerial photos and multispectral Landsat images for the same years.

We considered the maps and TPM constructed from the aerial photos as a control, and compared them to those constructed from the Landsat images classified with several methods: mean-shift segmentation followed by Random Forest classification methods, and three pixel-based methods of classification: K-means, ISODATA, and maximum likelihood. For each classification the TPM were compared to the TPM constructed from the aerial photos.

The goodness of fit of all maps obtained with the pixel-based methods was insufficient for estimating the LULC TPM. The LULC map obtained with the object-based classification method fit well to that based on the aerial photos, but the estimates of TMP were qualitatively different from those constructed from the aerial photos.

This article raises doubts regarding the adequacy of Landsat data and standard classification methods for establishing LULC CA model rules, and calls for the careful reexamination of the entire land-use CA framework.

1. Introduction

Conceptual simplicity and the ability of explicit representation of landscapes and their changes make Cellular Automata (CA) a standard tool for simulating urban and regional land-use dynamics. Typically, the CA models focus on estimating the rules of the LULC changes and analysis of the simulation results. However, the models put aside the uncertainty of the LULC maps that are used for establishing the transition rules.

The major source of data for the CA modeling is Remote Sensing (RS) multispectral imagery classified for establishing LULC dynamics. It is often reported that the CA models are quite successful in predicting LULC, with the high overall fit (80-90%) between the real LULC and model outputs. This is indeed true when the validation is based on comparing the *entire modeled area*. However, as far as initial area is excluded from the comparison, the spatial fit between the predicted and real *changes* drops down (Hagen-Zanker *et al.* 2005; Pontius and Petrova 2010).

A hierarchy of reasons of limited capacity of the CA models for predicting LULC changes can be proposed: (1) CA framework as a whole is insufficient for predict

LULC dynamics; (2) The CA framework works, but wrong CA rules are chosen; (3) The CA framework works, the rules are properly established, but the data chosen for estimating parameters of the rules do not represented the real of the LULC changes. In this paper we deal with the latter and investigate the adequacy of the RS data for calibration and validation of the CA models.

2. Testing the adequacy of classifications methods

The adequacy of the RS classification for representing LULC *changes* remains on the margin of the CA modeling studies. The modeling studies carelessly exploit simplest methods of the RS images classification, take their outputs for granted, and focus on model calibration. This may evidently result in *inadequate transition rules* regardless of the calibration methods.

LANDSAT imagery is a common choice of RS data and we investigated the adequacy of different methods of their classification for establishing CA model rules.

1.1 LULC Transition Probability Matrix

The background of the CA model is Transition Probability Matrix (TPM) $\{p_{ij}\}$ - a set of probabilities, per time unit, of transition $S_i \rightarrow S_j$ between the states S_i and S_j of the LULC CA. Our study compares TPMs estimated based on the LANDSAT maps obtained by the different classification methods to the TPM that is estimated based on the manual interpretation of high-resolution aerial photos of the same area.

1.2 Experimental area and Remote Sensing Data

The experimental area is the 15x6 km transect that starts in the center of the city of Netanya, Israel, and extends to surrounding agriculture areas. The period of comparison 1972 – 2008 (36 years) is divided into 4 intervals of 6 - 11 years, depending on availability of the LANDSAT images and aerial photos. Based on the manual interpretation of the high-resolution aerial photos, we have constructed the maps of Netanya LULC dynamics of six LULC states: built-up areas (BU), roads (RD), agricultural (AG) and vegetation (VG) areas, open spaces (OS) and water surfaces (WA). In this short paper, we present the results aggregated into two states only – Built-up (BU) and the NB-state that aggregates the rest five LULC states.

1.3 Classification Methods

In parallel to the manual interpretation, four pixel-based methods and one object-based method were applied for classifying these LULC on the LANDSAT imagery. All exploited pixel-based methods are traditional first choice of a CA modeler: K-means, ISODATA, Maximum Likelihood (ML) and hybrid classification. The object-based method we apply is two-staged: mean-shift clustering segmentation is followed by a Random Forest classification.

2. The results

The fit between the LANDSAT-based maps and the map that is based on manual classification varies depending on the method. Segmentation and maximum likelihood methods represent better results of LANDSAT classification (Figure 1). All the rest methods showed low values of accuracy.

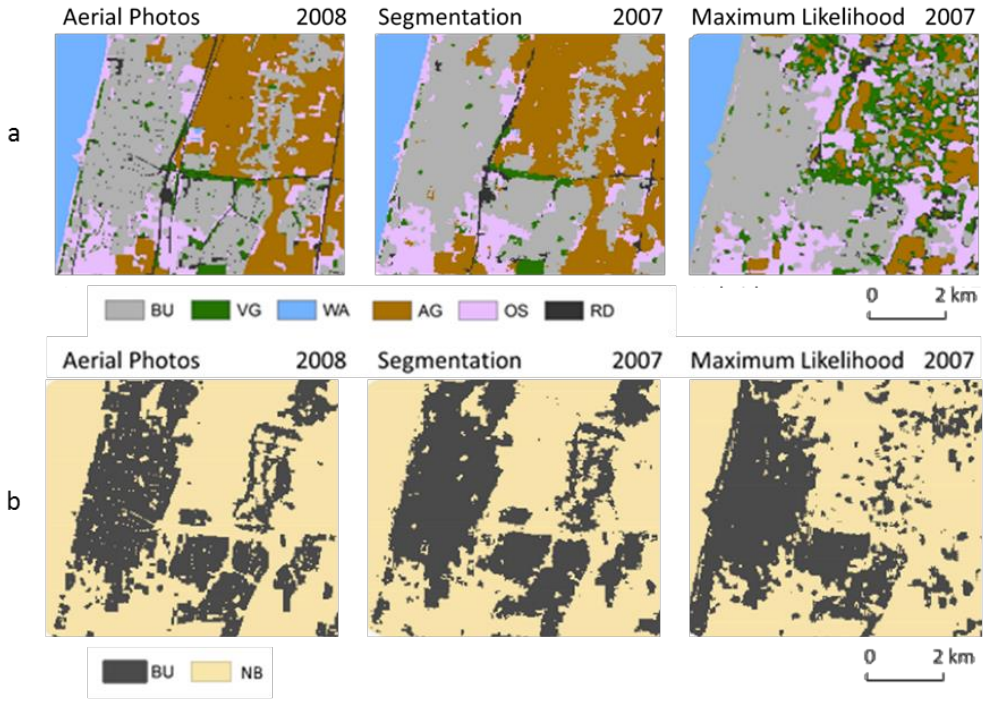


Figure 1. Land-use maps of 2007/2008 for three of six investigated methods, part of the study area: (a) all six LULC states and (b) aggregated BU and NB states

For each observation period we constructed TPM normalized to the 10-year period and compared them to the TPM constructed based on the manual interpretation (Table 1). Due to limited space, the TPMs are presented for the LULC uses aggregated into two classes: BU - built-up areas; NB – non-built-up areas, which include agricultural and vegetated areas, open spaces, water surfaces and roads.

Table 1. Average TPM for the probabilities normalized by the 10 year period

		LULC at year t + 10											
		Aerial Photos		Segmentation		ML		Hybrid		K-means		Isodata	
LULC at year t		BU	NB	BU	NB	BU	NB	BU	NB	BU	NB	BU	NB
	BU	**0.98**	0.02	**0.81**	0.20	**0.58**	0.42	**0.42**	0.58	**0.45**	0.55	**0.42**	0.58
	NB	0.05	**0.95**	0.08	**0.92**	0.11	**0.89**	0.08	**0.92**	0.16	**0.84**	0.19	**0.81**

As can be seen from Table 1, for the presented period, the TPMs obtained with the ML and Segmentation methods are *qualitatively* and *quantitatively different* from the TPM estimated based on the aerial photos. Most important, in reality, LULC states are changing in time essentially less frequently than it is obtained based on the RS images classified with the ML method; for example, in reality, the probability of the BU→BU and NB→NB transitions per 10 years are close to 1, while according to the ML map these probabilities are 0.58 and 0.89. The fit is even worse for the rest of the pixel-based methods.

The TPM obtained with the Segmentation method fits to the TPM for the aerial photos better than the TPM of the pixel-based methods, but is yet essentially biased.

We thus conclude that none of the maps obtained, based on the LANDSAT images, with the help of the popular pixel-based classification methods can be

exploited for establishing CA transition rules. Object-based method provided better, but yet insufficiently precise estimates.

We call for the revision of approach to the CA calibration and validation. An open depository of high-resolution, carefully validated, long-term series of the land-use/cover maps that reflect different types of LULC dynamics, and represent different types of land planning systems for different periods of population growth and economic development should be established. Instead of establishing a new database for every new CA model, one has to use these data series for calibration and validation of her/his new model. Only then, the model can be applied to the new dataset which, as we have demonstrated, must be constructed with the great care.

References

Hagen-Zanker, A., J. van Loon, A. Maas, B. Straatman, T. de Nijs, and G. Engelen. 2005. Measuring performance of land use models: An evaluation framework for the calibration and validation of integrated land use models featuring Cellular Automata. Paper read at 14th European Colloquium on Quantitative Geography, September 9-13, at Tomar, Portugal.

Pontius Jr, R. G., and S. H. Petrova. 2010. Assessing a predictive model of land change using uncertain data. *Environmental Modelling & Software* 25 (3):299-309.

Semantically Refining the Groundwater Markup Language (GWML2) with the Help of a Reference Ontology

Torsten Hahmann[1], Shirly Stephen[1], Boyan Brodaric[2]

[1]School of Computing and Information Science, University of Maine, Orono, ME 04469, USA
Email: {torsten.hahmann | shirly.stephen}@maine.edu

[2]Geological Survey of Canada, Ottawa, Canada K1A0E9
Email: boyan.brodaric@canada.ca

Abstract

Reference ontologies are intended to aid domain ontology design, identify gaps and inconsistencies in representations of domain information, and facilitate data interoperability. The application of a reference ontology to the water domain is untested. We present findings from using a first-order logic reference ontology for the water domain, the Hydro Foundational Ontology (HyFO), to identify and remedy semantic gaps and inconsistencies in the Groundwater Markup Language (GWML2), a data model for groundwater information with less detailed formal semantics. We express GWML2 as a logical extension of HyFO, thereby improving GWML2's compatibility with other hydro data models. We derive general desiderata for a "good" domain reference ontology in the geosciences and discuss the benefits one can expect from their use for the ontological analysis of geoscience data models.

1. Introduction

Effective water management requires exchanging and integrating information about the location, quantity, and flow of water throughout the water cycle. The information is typically stored in multiple data stores based on different data structures, terminologies, and light-weight ontologies, subsequently summarily referred to as *data models*. Knowledge integration and querying across these data stores requires interoperability between their representations at the syntactic, schematic, and semantic (comprising differences in terminology and definitions) levels. To prepare for automated integration of geoscience knowledge across these levels, we explore the use of a reference ontology (Noy 2004) as a tool for increasing semantic precision and coherence in geoscience data models. We specifically test this idea within the hydro domain by using the Hydro Foundational Ontology (HyFO), a reference ontology for the hydro domain developed since 2011 (see e.g., Hahmann & Brodaric 2012, 2013; Brodaric & Hahmann 2014), to semantically analyze the Groundwater Markup Language (GWML2) (Boisvert & Brodaric 2012; Brodaric 2015) as one example of a hydro data model developed by domain scientists. The result is an improved version of GWML2 with (1) increased semantic precision through axiomatic constraints (i.e., constraints expressed in a logical language such as first-order logic) and definitions based on well-defined reference terms from HyFO, (2) a stratified formalization that separates concepts based on how broadly they apply (across geosciences, to the entire water domain, or only to groundwater), and (3) a completed taxonomy that fills gaps and renames classes to better reflect their position within the stratification.

The domain reference ontology HyFO is not yet another standard that restricts fairly generic scientific terms (classes and relations), such as *geologic unit*, *water body*, or *aquifer*, to a single interpretation. Instead, it provides a neutral but concise language for describing in a machine-interpretable format how terms are used in existing data models. Such formal descriptions subsequently allow data models to be semantically compared and integrated in a largely automated fashion for integrated querying and knowledge discovery as envisioned in semantic e-science (Brodaric and Gahegan 2010).

2. Background and Related Work

State-of-the-Art Semantic Representations for the Hydro Domain A number of data models have emerged that standardize water data syntactically and, to some extent, semantically. However, they are fragmented in that they describe only disconnected subareas of the hydro domain, such as groundwater storage and flow (e.g., GWML2 (Boisvert & Brodaric 2012; Brodaric 2015), INSPIRE Geology (INSPIRE 2013)), surface hydrography and connectivity (e.g., USGS's NHDPlusV2, INSPIRE Hydrography (INSPIRE 2009), HY_Features (Dornblut & Atkinson 2013)), water quality (e.g, WaterML2) or stream geometry (e.g., RiverML). Moreover, the meaning of classes in the existing data models are described only via subclass relationships, via generic UML associations, and via free-text descriptions, which are insufficient for machine-interpretability and incompatible across standards. This is especially problematic for central scientific terms that are used almost universally in all hydro data models, such as *geologic unit*, *water body*, *aquifer*, or *channel*. Other concepts central to modeling water storage on the Earth, such as *container spaces* and *voids*, are omitted altogether or alluded to only vaguely.

Existing Approaches to Semantic Integration Existing *ontology mapping* and *alignment* techniques as surveyed, e.g., in (Kalfoglou & Schorlemmer 2003), aim to find similarities, equivalences, and subsumption relations between the contents of ontologies. These largely automated techniques are limited in ways that prevents their use for integrating the existing hydro data models: the ontologies must (1) be specified in a language from the OWL family (Noy 2004), (2) be already syntactically and schematically integrated, and (3) be of similar scope (i.e., describe the same part of the domain). Most problematically, automatic techniques to semantically integrate different data models or ontologies can do so only to the extent to which the semantics are already specified in a machine-interpretable way. The lack of formal specifications of the semantics of the existing hydro data models requires encoding them manually. In our work presented here, we manually construct a machine-interpretable version of GWML2 by expressing its semantics using HyFO's rigorous axiomatization.

3. Approach Using a Domain Reference Ontology

Nature of a Domain Reference Ontology Generally, a domain reference ontology can support semantic integration by providing a formal language that provides a set of *neutral, formal terms* for concisely identifying, via axiomatic mappings, the differences and nuances in the interpretations of similar concepts across data models. The set of formal terms should be small, but each term should have tightly restricted semantics. These formal terms are not meant to capture a single, agreed-up meaning of inherently complex scientific terms, but instead serve as a machine-interpretable language to precisely describe the differences between alternative interpretations of scientific terms by providing fine-grained semantic targets to map to.

Thus, a domain reference ontology must: (1) *identify a core set of formal domain concepts*, and (2) *tightly constrain and relate them axiomatically* in sufficient detail. The second aspect requires specification in an expressive[1] machine-interpretable language, to ensure that the formal terms are interpreted unambiguously, and to permit automation of verification and subsequent integration among data models. A reference ontology must further (3) *cover the entire domain of interest* (e.g., the hydro domain) *broadly*, meaning it should omit concepts that are only relevant in specific subdomains or applications. To ensure that the formal terms are well-distinguished from one another, the reference ontology is (4) ideally *grounded in an upper ontology* that provides the philosophical underpinning for the distinctions between the different kinds of objects and processes relevant to the domain.

[1] Even the ontology languages from the OWL family have proven insufficiently expressive for the purpose of a reference ontology, preferable are first-order or higher-order logics.

The Hydro Foundational Ontology (HyFO) In prior work, we have laid the foundation for the *Hydro Foundational Ontology* (HyFO) (Hahmann & Brodaric 2012; 2013; Brodaric & Hahmann 2014) as a reference ontology for the hydro domain. HyFO is the result of formal ontological analysis of concepts and relations that play a role in water storage on and below the Earth's surface by geo-ontology experts, rigorously formalizing them in full first-order logic as specialization of the DOLCE upper ontology (Masolo et al. 2003). HyFO identifies four key concepts that form the *Hydro Ontological Square* (Brodaric & Hahmann 2014): (1) a physical container such as a rock formation; (2) a physical void such as a depression in the ground surface, or microscopic pores in the container's matter where water can be stored, (3) a body of water located in the void (and contained by the container), (4) and the rock and water matter that constitute the container and water body. These concepts are interrelated by the relations of *containment*, *constitution*, and *hosting a void*. As result of the presented work, we propose to add *hydro rock body* as a fifth concept that represents a complex physical object that consists of (i) a container body constituted by solid (e.g., rock) matter, (ii) a void hosted therein, and (iii) a body of water that is located in the void.

4. Results and Discussion

Stratified version of GWML2 The first-order logical axiomatization of GWML2 developed here adds semantic precision and clarity to GWML2's core concepts obtained from the GWML2 conceptual schema and accompanying textual descriptions. It results in a merged ontology that treats GWML2 as a consistent logical extension of HyFO and DOLCE. At the core, it consists of a refined and stratified taxonomy spanning four layers of increasing specificity (Fig. 1): (0) DOLCE concepts; (1) generic geological concepts (*geologic unit*, *earth material*, *fluid body*) that transcend the water domain; (2) hydro concepts that span surface and subsurface water (e.g., *hydro rock body*, *water body*, *hydro void*); and (3) groundwater specific concepts (e.g., *aquifer*, *well*, *subsurface water body*, *hydrogeo void*). This layering ensures that GWML2's groundwater concepts consistently specialize HyFO concepts, with HyFO also being able to anchor surface water concepts and thus being shareable across hydro ontologies.

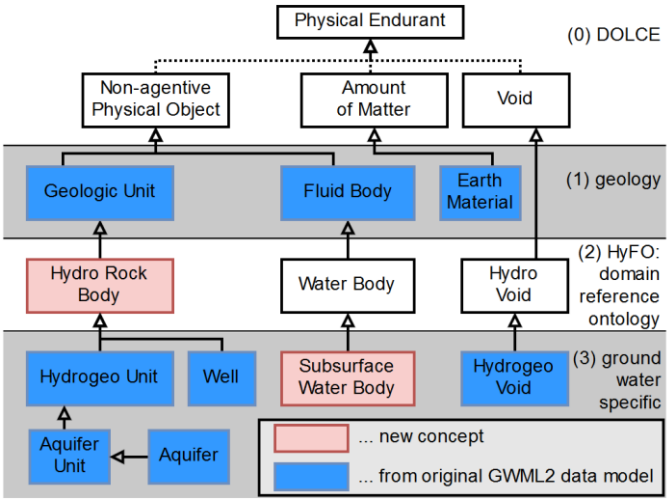


Fig. 1 Excerpt of the subclass hierarchy from the obtained stratified version of GWML2, with groundwater concepts (bottom layer) extending HyFO, general geology concepts and DOLCE.

Ontological analysis of GWML2 The resulting revised and refined GWML2 ontology reduces barriers to interoperability with other hydro data models. A more concrete contribution is our detailed ontological analysis that clarifies what kind of objects GWML2 terms refer to, fleshing out their spatial, spatio-temporal, material, physical, and ontological characteristics and the relationships (e.g., physical containment, constitution, or spatial parthood) between them. Some represent 3D physical objects (*geologic unit*, *hydrogeo void*, *aquifer*) constituted partly or wholly of solid and/or fluid matter, others denote 2D surfaces (e.g., *water table*), and others purely spatial abstractions (e.g., *monitoring site*). In our ontological analysis we particularly examined borderline cases – more unusual *geologic units*, *aquifers*, *wells*, or *fluid bodies* – that deviate from the typical textbook schemata of water storage in order to test the general applicability of the proposed axioms. We thereby avoid including axiomatic constraints that are true

in typical "textbook" schemata about water storage but that are not true in a more unusual situations and thus should not be encoded as part of the ontology's axiomatization.

The resulting ontology increases the semantic precision of GWML2 primarily by logically defining groundwater specific concepts (e.g., *subsurface water body* and *hydrogeo void* – the spaces where subsurface water bodies can be located) using a combination of HyFO's general hydro concepts (e.g., *water body* and *hydro void* – all spaces where water can be located) and spatio-physical relations. Adding precise definitions and classes that are missing from GWML2 (e.g., *container solid body* and *water matter*) also semantically connects classes (e.g., *well*) that were isolated in GWML2's original model. In addition, our analysis reorganizes GWML2 classes, moving those that are applicable beyond the groundwater domain to layers higher up in the taxonomy (e.g., *fluid body to* the general geology layer or *basin* to the HyFO layer). We specialize these concepts at more specific layer, for example, *water body* and *subsurface water body* are introduced as specializations of *fluid body* at the hydro and the groundwater layer.

Summary The following contributions are made: (1) Stratifying GWML2 classes, for a cleaner and more precise organization, and improved reusability and interoperability with other hydro ontologies. (2) Analyzing key GWML2 classes and proposing related axioms to add clarity, rigor, and detail. (3) Identifying a number of revisions to GWML2's conceptual schema, to better reflect its domain. (4) Recognizing *hydro rock body* as an important concept missing from the Hydro Ontological Square. (5) Completing initial tests that demonstrates HyFO's potential as a reference ontology for the water domain. More generally, our work exemplifies how a data model or lightweight ontology benefits from grounding in a deeply axiomatized reference domain ontology. Such grounding makes explicit subtle semantic differences between ontologies within a domain and thus enhances their semantic interoperability.

Acknowledgements

The authors gratefully acknowledge support for this work obtained from the Maine Economic Improvement Fund and the Groundwater Program of Natural Resources Canada.

References

Boisvert, E. and Brodaric, B. (2012) GroundWater Markup Language (GWML) – enabling groundwater data interoperability in spatial data infrastructures. Journal of Hydroinformatics, 14(1):93–107.

Brodaric, B. (2015): GroundWaterML2 – GW2IE Final Report. Technical Report Open Geospatial Consortium Engineering Report 15-082, version 2.1

Brodaric, B. and Gahegan, M. (2010) Ontology use for semantic e-science. Semantic Web Journal, 1(1–2):149–153.

Brodaric, B. and Hahmann, T. (2014) Towards a foundational hydro ontology for water data interoperability. In Proc. of the 11th Int. Conference on Hydroinformatics (HIC-2014).

Dornblut, I., Atkinson, R. (2013) HY_Features: a geographic information model for the hydrology domain. Technical Report GRDC 43r1, Global Runoff Data Centre, November 2013.

Hahmann, T. and Brodaric, B. (2012) The void in hydro ontology. In Conf. on Formal Ontology in Inf. Systems (FOIS-12), IOS Press, 45–58.

Hahmann, T. and Brodaric, B. (2013) Kinds of full physical containment. In: Tenbrink, T., Stell, J., Galton, A., Wood, Z. (Eds.) Conference on Spatial Information Theory (COSIT-13). Springer, 397-417.

INSPIRE Thematic Working Group Geology (2013) D2.8.II.4 INSPIRE Data Specification on Geology – Draft Guidelines, v3.0 rc3. Tech. report, INSPIRE, 369pp.

INSPIRE Thematic Working Group Hydrography (2009) D2.8.I.8 INSPIRE Data Specification on Hydrography – Guidelines, v3.0. Tech. report, INSPIRE, 175pp.

Kalfoglou, Y. and Schorlemmer, M. (2013) Ontology mapping: the state of the art. Knowledge Eng. Review, 18(1):1–31, 2003.

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A. (2003) Wonderweb Deliverable D18 – Ontology Library (Final Report). Technical report, CRN-ISTC, Trento, 349 pp.

Noy, N. F. Semantic integration: A survey of ontology-based approaches. SIGMOD Record, 33(4):65–70, 2004.

Deriving Locational Reference through Implicit Information Retrieval

T. Hervey, W. Kuhn

Department of Geography and Center for Spatial Studies,
University of California, Santa Barbara, CA 93106
Email: {thomas.hervey; kuhn} @geog.ucsb.edu

Abstract

The often fragmented process of online spatial data retrieval remains a barrier to domain scientists interested in spatial analysis. Although there is a wealth of hidden spatial information online, scientists without prior experience querying web APIs (Application Programming Interface) or scraping web documents cannot extract this potentially valuable implicit information across a growing number of sources. In an attempt to broaden the spectrum of exploitable spatial data sources, this paper proposes an extensible, locational reference deriving model that shifts extraction and encoding logic from the user to a preprocessing mediation layer. To implement this, we develop a user interface that: collects data through web APIs and scrapers, determines locational reference as geometries, and re-encodes the data as explicit spatial information, usable with spatial analysis tools, such as those in R or ArcGIS.

1. Introduction

GIS advancements have produced a growingly complex general-purpose toolbox rather than functionality tailored to domain-specific questions. Frequently, domain scientists including Green (2015: 717) highlight the salient lagging data and tool limitations associated with GIS. As Kuhn (2012: 2267) notes, it is essential to rethink the fundamentals of spatial information while promoting clarity that cuts across technical boundaries and broadens spatial literacy for non-experts. Contributing to the work by Kuhn and Ballatore (2015: 219) and Vahedi *et al*. (2016) to design an intuitive GIS language for question-driven spatial studies, we focus on bridging the gaps between data discovery and spatial analysis tools by broadening the spectrum of exploitable spatial data sources.

Compared to the vast amount of implicit spatial data (hidden location attributes often in the form of metadata, auxiliary place names, and geotagged attributes (Heinzle and Sester 2003: 335)), there remains a relatively limited quantity of online explicit spatial data (georeferenced geometry-based features (Brisaboa *et al.* 2011: 358)). When available, explicit data are typically served from a limited number of administrative portals or require intensive energy and time from a user searching, exporting, encoding and cleaning before being usable (Munson 2013: 65). These preprocessing requirements limit the feasibility of question-driven spatial analysis (Vahedi *et al*. 2016) and force domain scientists to base their studies on data availability.

It is clear that implicit spatial data is an attractive alternative. However, as the numerous research challenges associated with GIR (geographic information retrieval) suggest (Jones and Purves 2008: 219), current methods do not provide adequate solutions to navigate, gather or utilize the mass of heterogeneous implicit spatial data spread across the array of online repositories. Custom-constructed web API requests and scrapers can help retrieve and process this unpublished information. Yet, without technical expertise to build new or use existing

modules, this information remains unobtainable. This begs the question, why so few GIR studies have focused on aiding implicit spatial data retrieval.

To explore this idea, we propose an extensible locational reference deriving model that shifts extraction and encoding logic from the user to preprocessing software. In the following sections, we describe previous work and provide a prototype architecture with applications to test the plausibility of a model that adopts the core concepts of spatial information (Kuhn 2012: 2267).

2. Previous Work & Limitations

There have been many notable efforts to address the difficulties of GIR, but few have focused on the extraction and encoding of implicit spatial data. INSPIRE[1], NSDI[2], and SPIRIT (Jones *et al.* 2004: 125) for example, are spatial data infrastructure and search engine projects that focus on standardizing and indexing web documents rather than broadening access to new faster-growing data repositories (Jones *et al.* 2004: 125; Brisaboa *et al.* 2011: 358).

Google's Fusion Tables[3] have simplified data integration and ArcGIS Online has grown online spatial catalogs through user sharing and publishing. Yet, these tools do not aid in data extraction from external sources. Enterprise products like Crimson Hexagon[4] and Temboo[5] respectively provide users with spatial social media analytics and code snippets to ease web API querying. But these solutions provide neither data source extensibility, nor recognize multiple location types. Furthermore, studies on linked data, the semantic web, and semantic gazetteers (Cardoso *et al.* 2016: 389; Gao *et al.* 2014) propose a traversable data-centric organization of the web (Kuhn *et al.* 2014: 173; Wiegand and García 2007: 355). But until a semantic model broadly leverages new and existing data, significant user effort will remain necessary for collecting and preprocessing. The following architecture aims to address these issues and bridge the gap between discovered data and analysis tools.

3. Architecture

The proposed model's implementation will be accessible via web interface and GIS plugins. It is not an independent search engine, but rather works as mediation layer once a desirable data source has been found and a user supplies the text parameters noted in Table 1. It does not aim to replace, but rather leverage, existing extraction and encoding methods as well as supply locational reference deriving logic not present in current systems. Illustrated in Figure 1, the mediation layer divides preprocessing (extracting and encoding) into four sequential tasks: extraction, context building, geometry matching, and encoding.

Table 1. User interface and plugin input parameters

Mandatory	Optional
Extraction source (*e.g.* URL or API endpoint)	Basic aggregation (*e.g.* state-level)
Source type (*e.g.* Twitter tweets or hashtags)	Basic conditionals (*e.g.* exclude attributes)
Output geometry type (*e.g.* point, polygon)	Error handling (*e.g.* toponym conflicts)
Output file format (*e.g.* **.**shp, **.**kml, .geoJSON)	

[1] http://inspire.ec.europa.eu/

[2] http://www.fgdc.gov/nsdi/

[3] https://support.google.com/fusiontables/answer/2571232?hl=en

[4] http://www.crimsonhexagon.com/

[5] https://temboo.com/

The Extraction task makes up the majority of the codebase, which will expand as interest in new data sources emerge. It leverages parameterized HTTP requests constructed by web APIs (for database retrieval) and scrapers (for document text retrieval). The result will often be a structured or semi-structured tree of attributes (*e.g.* .json, .xml) presented to the user for accuracy feedback.

The second task searches the data for potential location characteristics in the form of raw geometry or toponyms using a combination of text matching and natural language processing paired with gazetteer lookups (*e.g.* GeoIP, GeoNames, Getty Thesaurus of Geographic Names). When applicable, user feedback will be necessary for place name disambiguation, scoping, and aggregation.

The third task handles geometry pairing and the fourth, output file formatting. Once location characteristics have been determined, associated geometries are retrieved from repositories (*e.g.* TIGER/Line, Open Street Map) and paired with the data. Additionally, rudimentary spatial extent and relationship logic is supported (using ESRI's Java geometry API, GDAL, etc.), such that a *within* operation could aggregate geotagged Flickr photo points to a bounding county polygon. Finally, conversion tools (*e.g.* GDAL ogr2ogr, Python pyshp), encode the data in an array of formats (*e.g. .shp, .kml*)., resulting in a spatially explicit data product usable in analysis.

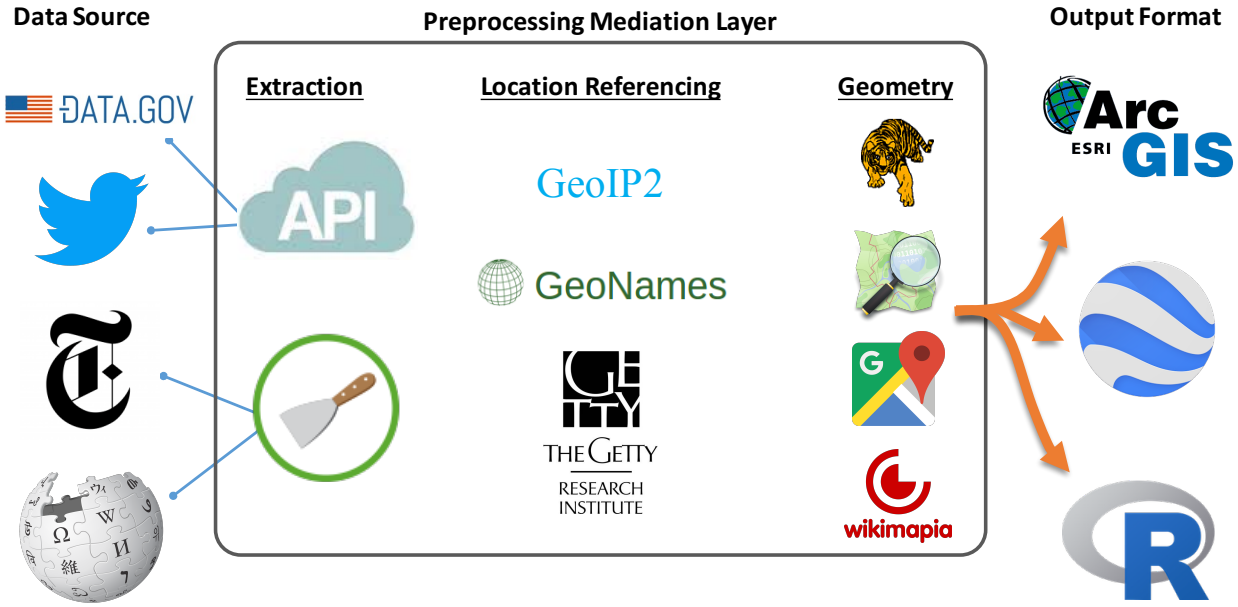


Figure 1. Examples of leveraged services between implicit data and analysis tools

4. Application

To articulate the model's potential, we present working and hypothetical examples where domain scientists encounter data retrieval problems. In our working example, a social scientist wants to examine the spatial interaction between crime incidents and income across Seattle, USA. She has statistical experience with R and has imported a block-group median household income shapefile, but cannot find explicitly spatial crime data. She has, however, found a well formatted crime statistics table on Seattle's web portal, and wants to extract the *time*, *crime type*, and *charge* attributes by city-block. Traditionally, to integrate this information, she would have to download and import the table, for example, into ArcGIS, remove unwanted attributes, download a block-group shapefile, aggregate the crime points, and handle geocoding anomalies. Alternatively, the previous steps are simplified by supplying the model's web interface with the

table's API endpoint, specifying which attributes to keep at which aggregation level, and specifying polygon shapefile as the output. On the back end, the software retrieves the data as JSON objects, determines street address as its location reference, geocodes and aggregates these points to block-group geometry served by *data.gov*, and produces a shapefile using ogr2ogr.

Our hypothetical example adds complexity when a user requires data from social media. A political scientist wanting to understand spatial voting trends between Californian counties during a presidential primary, decides to use trending Twitter hashtags as predictors. Without programming experience, he is limited to Twitter's simple query building wizards that returns unwieldy and spatially unreferenced tweet data. Instead, by providing twitter API credentials, hashtags of interest, and California counties as text parameters to the interface, the model can filter by hashtag, extract user's location information (through geotags or home locations) and aggregate tweet count to the county level, producing a .kml file from OpenStreetMap geometry.

5. Conclusions

We have presented a locational reference deriving model and associated prototype preprocessing layer that has the potential to promote critical spatial thinking by expanding data source options. Currently, the model's integrity is limited by the credibility of its data retrieval sources, and limited to handling vector data. Therefore, future research will investigate integrating new data sources, analyzing feedback to promote VGI (volunteered geographic information) supported gazetteers, as well as integrating data credibility metrics.

References

Brisaboa N, Luaces M, and Seco D, 2011, New Discovery Methodologies in GIS. *Geographic Information Systems*, 358–376.

Cardoso S, Amanqui F, Serique K, dos Santos J, and Moreira D, 2016, SWI: A Semantic Web Interactive Gazetteer to Support Linked Open Data. *Future Generation Computer Systems*, 54:389–398.

Gao S, Li L, Li W, Janowicz K, and Zhang Y, 2014, Constructing Gazetteers from Volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*.

Green T, 2015, Places of Inequality, Places of Possibility: Mapping "Opportunity in Geography" Across Urban School-Communities. *Urban Review*, 47(4), 717–741.

Heinzle F and Sester M, 2003, Derivation of Implicit Information from Spatial Data Sets with Data Mining. *Cartography*, 35(4)335–340.

Jones C and Purves R, 2008, Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones C, Abdelmoty A, Finch D, Fu G, and Vaid S, 2004, The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Geographic Information Science, Proceedings,* 3234:125-139.

Kuhn W and Ballatore A, 2015, Designing a Language for Spatial Computing. *AGILE 2015, Lecture Notes in Geoinformation and Cartography*, 219–234.

Kuhn W, Kauppinen T, and Janowicz K, 2014, Linked Data- A Paradigm Shift for Geographic Information Science. *Proceedings of The 8th International Conference on Geographic Information Science*, Vienna, AUT, 173–186.

Kuhn W, 2012, Core Concepts of Spatial Information for Transdisciplinary Research. *International Journal of Geographical Information Science*, 26(12):2267–2276.

Munson M, 2012, A study on the Importance of and Time Spent on Different Modeling Steps. *ACM SIGKDD Explorations Newsletter*, 13(2):65–71.

Vahedi B, Kuhn W, and Ballatore A, Question Based Spatial Computing - A Case Study. *AGILE 2016 (in press).*

Wiegand N and García, C, 2007, A Task-Based Ontology Approach to Automate Geospatial Data Retrieval. *Transactions in GIS*, 11(3):355–376.

Hybrid Indexing for Parallel Analysis of Spatiotemporal Point Patterns

Alexander Hohl[1], Irene Casas[2], Eric M. Delmelle[1], Wenwu Tang[1]

[1] University of North Carolina at Charlotte, Department of Geography and Earth Sciences and Center for Applied GIScience, Charlotte, NC, 28223, USA. Email: {ahohl; Eric.Delmelle; WenwuTang}@uncc.edu
[2] Louisiana Tech University, Department of Social Sciences, Ruston, LA, 71272, USA. Email: icasas@latech.edu

Abstract

High-performance parallel computing outperforms desktop workstations for computationally demanding problem solving. Domain decomposition and spatial indexing are widely used to accelerate spatial search. A single index method for spatiotemporal data processing lacks retrieval efficiency for massive computation. Combining multiple indexing methods to a hybrid spatiotemporal index holds potential for addressing this data retrieval challenge. We perform adaptive octree decomposition of the spatiotemporal domain and build local *k-d* trees to accelerate nearest neighbour search for space-time kernel density estimation (STKDE). Our parallel implementation reaches substantial speedup compared to sequential processing. The hybrid index outperforms octree decomposition alone, especially at lower-levels of parallelization. This approach facilitates finer-scale computation, enabling us to discover patterns that would be hidden otherwise.

1. Introduction

Many algorithms for analyzing spatiotemporal data are computationally intensive because they often rely on extensive nearest-neighbor (NN) search. Spatial indexing methods, including R-trees and quadtrees, have long been used to accelerate spatial queries (Lu and Ooi 1993). Further, parallel computing offers the capacity for efficient processing of intensive analyses. Efficient algorithms for range queries and k nearest neighbor (kNN) queries have been developed within parallel communication models, like Hadoop (Agarwal *et al*. 2016). Hering (2013) showed that performance of in-memory *k-d* trees is best for intermediate number of dimensions (6-13). In addition, kNN search was implemented using graphics processing units (GPU) (see Liang *et al*. 2010; Sismanis *et al*. 2012), where distance calculation and sorting are parallelized. Alternatively, the strategy of spatial domain decomposition is used for distributing the resulting subdomains to multiple computing resources for concurrent processing (Wilkinson and Allen 1999). However, including time in geographic models complicates requirements for spatial indexing, resulting in low retrieval efficiency (Gu 2011). Merging multiple indexing methods to form hybrid spatiotemporal indices was recently proposed by Azri *et al*. (2013). In this study, we perform octree-based recursive decomposition of the space-time domain for parallel computation of STKDE. Using *k-d* tree indexing within octree leaf nodes (Liu *et al*. 2008) we accelerate kNN search for STKDE.

2. Methods

2.1 Data

We used a dataset of dengue fever cases in Cali, Colombia for years 2011 and 2012. Each of the 11,168 records holds x- and y-coordinates and a timestamp (Delmelle *et al*. 2013). The rectangular envelope of the dataset spans 15,000m * 20,000m * 727days.

2.2 Space-Time Kernel Density Estimation

STKDE results in a 3D volume where each voxel (volumetric pixel) is assigned a density estimate based on the surrounding datapoints. Specifically:

$$\hat{f}(x,y,t) = \frac{1}{nh_s^2h_t}\sum_i i(d_i < h_s, t_i < h_t)k_s\left(\frac{x-x_i}{h_s}, \frac{y-y_i}{h_s}\right)k_t\left(\frac{t-t_i}{h_t}\right) \quad (1)$$

Density $\hat{f}(x,y,t)$ of voxel s is estimated based on neighboring data points (x_i, y_i, t_i), which are weighted using the spatial and temporal Epanechnikov kernel functions, k_s and k_t (Epanechnikov 1969). Spatial and temporal distances between voxel and data point are given by d_i and t_i. If they are smaller than the spatial (h_s) and temporal bandwidths (h_t) respectively, the indicator function $i(d_i < h_s; t_i < h_t)$ equals 1, otherwise 0. We chose a combination of large bandwidths (h_s=2500m and h_t=14days), and a discretization level of 100m*100m*1 day, resulting in 16,442,664 voxels for NN search.

2.3 Spatiotemporal Domain Decomposition and Indexing

Accelerating NN search for STKDE, we develop a hybrid decomposition and indexing method. Figure 1 illustrates octree decomposition and *k-d* tree indexing (2D was used for illustration purpose). The black lines symbolize octree decomposition, creating subdomains of similar computational intensity (CI). The decomposition algorithm generates 8 cuboid subdomains by dividing each of the three axes into two equal parts. Decomposition continues recursively until no subdomain contains more points than the specified threshold (50 here), given the ratio between subdomain volume and combined subdomain and buffer volume stays above 0.1 (preventing unnecessarily small subdomains). Avoiding edge effects in STKDE, we implement buffers equal to h_s and h_t around all subdomains. For details, see Hohl *et al*. (2015). For each octree leaf node, we build a local *k-d* tree on the containing points (red lines in Figure 1). *K-d* tree is a binary tree structure for arranging points in k-dimensional space (Bentley 1975). It allows for efficient retrieval, and has been widely used for NN search (Azri *et al*. 2013). We use *k-d* tree index to accelerate the kNN search for each voxel for STKDE (blue crosses in Figure 1), resulting in massive queries. We quantify CI of each subdomain as a function of number of points and number of voxels within. To balance workloads, we distribute subdomains by equalizing cumulative CI among CPUs. Using computing time (T) and speedup (S) (Wilkinson and Allen 1999) as metrics, we compare two approaches: 1) octree decomposition only, 2) hybrid octree decomposition with *k-d* tree indexing. We compare impact of problem size between the two approaches by computing above metrics for sub-datasets (by resampling the dengue fever dataset to 1000, 2000, 3000, … , 11000 points), while fixing the number of processors at 100.

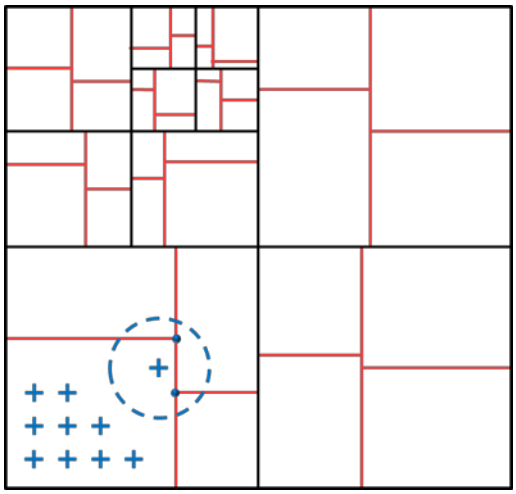


Figure 1. Spatiotemporal domain decomposition

3. Results and Discussion

Using approach 1, sequential computing time (T_s) is 40,297s. (Figure 2). When utilizing 200 CPUs in parallel, execution time (T) drops to 244s. and speedup is 165. Using approach 2, T_s is almost cut in half (22,170s.). For 200 CPUs, T is not much lower than in approach 1 (191s.), whereas speedup is lower (122). Therefore, the parallel algorithm scales better for bigger computations (approach 1) than for smaller ones (approach 2), where we used *k-d* tree indexing to reduce complexity of NN search.

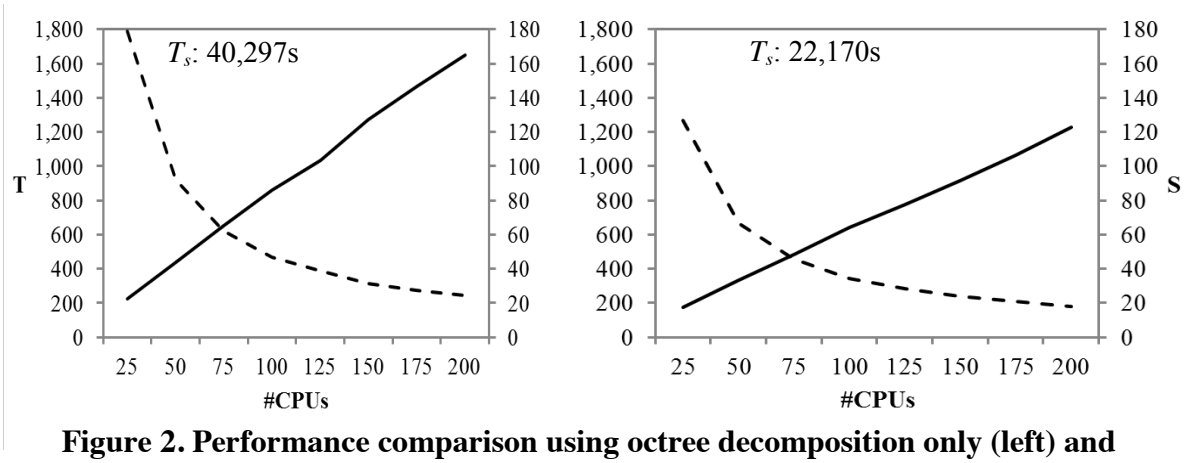


Figure 2. Performance comparison using octree decomposition only (left) and hybrid indexing (right) for 25-200 CPUs. (Speedup S: —, Execution time *T*: ---). Sequential time: T_s.

The impact of problem size confirms above findings (Figure 3): for approach 1, T_s increases from 7,795s. (using 1,000 points) to 40,060s. (11,000 points). Utilizing 100 CPUs, T linearly increases from 210 to 457s. and the increase in speedup flattens out at around 4,000 points. Using hybrid indexing, T_s increases from 1,620s. (1,000 points) to 21,511s. (11,000 points), T linearly increases from 58s. to 332s. while speedup flattens out at 5,000 points.

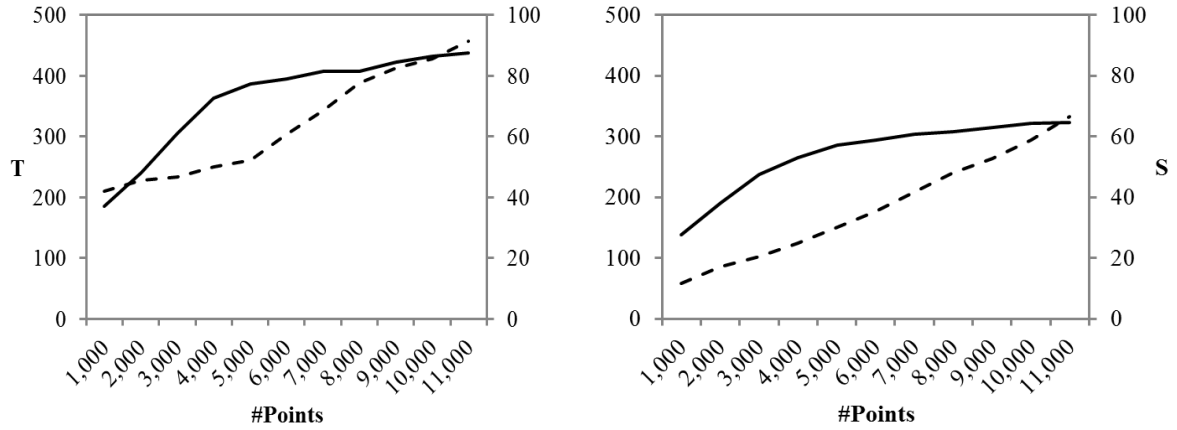


Figure 3. Performance comparison using octree decomposition only (left) and hybrid indexing (right) for 100 CPUs and 1,000 – 11,000 points. (Speedup S: —, Execution time *T*: ---).

4. Conclusions

Our parallel implementations of STKDE reach significant speedup. Both approaches dramatically reduce computational effort for analyzing space-time patterns. As the octree

decomposition only approach (approach 1) performed similarly to hybrid indexing (approach 2) for high number of CPUs, we conclude that larger computations harness better high-performance parallel computing power. Because STKDE necessitates a huge amount of spatiotemporal queries due to the discretization level, the computation is truly massive. Future experiments include bigger datasets and higher thresholds for octree decomposition, further exploring the utility of hybrid indexing in parallel spatial computing.

References

Agarwal, P. K., K. Fox, K. Munagala, and A. Nath. 2016. Parallel Algorithms for Constructing Range and Nearest-Neighbor Searching Data Structures. In Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems, 429-440.

Azri, S., U. Ujang, F. Anton, D. Mioc, and A. A. Rahman. 2013. Review of Spatial Indexing Techniques for Large Urban Data Management. In International Symposium & Exhibition on Geoinformation (ISG).

Bentley, J. L. 1975. Multidimensional binary search trees used for associative searching. *Communications of the ACM* 18 (9):509-517.

Delmelle, E., I. Casas, J. H. Rojas, and A. Varela. 2013. Spatio-temporal patterns of Dengue Fever in Cali, Colombia. *International Journal of Applied Geospatial Research (IJAGR)* 4 (4):58-75.

Epanechnikov, V. A. 1969. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications* 14 (1):153-158.

Gu, W. W., Jishui; Shi, Hao; Liu Yongshan. 2011. Research on a Hybrid Spatial Index Structure. *Journal of Computational Information Systems* 7 (11):3972-3978.

Hering, T. 2013. Parallel Execution of kNN-Queries on in-memory KD Trees. In BTW Workshops. 257 - 266.

Hohl, A., E. Delmelle, and W. Tang. 2015. Spatiotemporal domain decomposition for massive parallel computation of space-time kernel density. *ISPRS Annals of the Photographer, Remote Sensing and Spatial Information Sciences* 2(4): 7.

Liang, S., Y. Liu, C. Wang, and L. Jian. 2010. Design and evaluation of a parallel k-nearest neighbor algorithm on CUDA-enabled GPU. In 2010 IEEE 2nd Symposium on Web Society.

Liu, H., Z. Huang, Q. Zhan, and P. Lin. 2008. A database approach to very large LiDAR data management. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, China* 37 (B1):463-468.

Lu, H., and B. C. Ooi. 1993. Spatial indexing: Past and future. *IEEE Data Eng. Bull.* 16 (3):16-21.

Sismanis, N., N. Pitsianis, and X. Sun. 2012. Parallel search of k-nearest neighbors with synchronous operations. In IEEE Conference on High Performance Extreme Computing (HPEC), 2012.

Wilkinson, B., and M. Allen. 1999. *Parallel programming*: Prentice hall New Jersey.

Tweet Geolocation Error Estimation

E. Holbrook[1], G. Kaur[1], J. Bond[1], J. Imbriani[1], C. E. Grant[1], and E. O. Nsoesie[2]

[1]University of Oklahoma, School of Computer Science
Email: {erik; cgrant; gkaur; jared.t.bond-1; joshimbriani}@ou.edu

[2]University of Washington, Institute for Health Metrics and Evaluation
Email: en22@uw.edu

Abstract

Tweet location is important for researchers who study real-time human activity. However, few studies have examined the reliability of social media user-supplied location and description information, and most who do use highly disparate measurements of accuracy. We examined the accuracy of predicting Tweet origin locations based on these features, and found an average accuracy of 1941 km. We created a machine learning regressor to evaluate the predictive accuracy of the textual content of these fields, and obtained an average accuracy of 256 km. In a dataset of 325788 tweets over eight days, we obtained city-level accuracy for approximately 29% of users based only on their location field. We describe a new method of measuring location accuracy.

1. Introduction

With the rise of micro-blogging services and publicly available social media posts, the problem of location identification has become increasingly important. Accurately assigning a geolocation to a tweet is extremely useful to a broad range of applications, like tracking the spread of disease through social media (Paul and Dredze, 2011). However, few users report their true location online. Only around 1% of tweets contain accurate location information on where the tweet was created, as retrieved from the Twitter API. (Culotta, 2014).

Previous methods use machine learning models trained on the textual content of the tweet and the user's past tweets to identify city-level location information (see § 2). These methods are heavily dependent on heuristics and the granularity of the training data: any extension of the techniques would require intense restructuring of the classifiers. Investigation into a different geographic region would require new data, retraining the classifiers, and a complete rework of the location types to be identified. Furthermore, similar methods require large amounts of data, e.g. all available tweets from each user or a large social network graph. These methods are often time- and cost-prohibitive: access to a large portion of the Twitter stream is expensive.[1] In contrast, we attempt to predict geolocations given only a single tweet.

In this paper, we examine the reliability and accuracy of predicting tweet origin locations based on meta-content of the of the tweet itself, specifically, the user-supplied 'location' and 'description' fields, as well as propose a lightweight system for identifying the location of a user.

[1]http://readwrite.com/2010/11/17/twitter_to_sell_50_of_all_tweets_for_360kyear_thro/

2. Related Work

The Twitter public API allows for programmatic access to a subset of all public tweets, and with open-source libraries like Tweepy.org, developers can investigate social trends. Identifying the locations of these trends has been examined from several different angles (Compton *et al.*, 2014; Zhang and Gelernter, 2014; Priedhorsky *et al.*, 2014). The common focus is predicting the home location of a Twitter user based on their past tweets.[2]

Compton *et al.* (2014) sought to identify the home location of users through a large social network consisting of approximately 110 million users and reciprocated mentions between users. Home locations were estimated by the strength of connections between users with known locations. They reported city-level accuracy for around 90% of the users.

Priedhorsky *et al.* (2014) sought to combine social features with tweet content features, and to produce a prediction with a quantitative error estimation in terms of distance. To this end, they proposed a location-estimation method based on Gaussian mixture models, and presented several variations of their algorithm for this estimation, which ranged in accuracy from approximately 1700 to 8000 kilometers.

Each of these methods requires a large amount of operational overhead, i.e. a very large number of tweets from each user and other data obtained from the Twitter API, or a complicated prediction mechanism. There is also a lack of a uniform, straight-forward error measurement: each defines it in a different way, making direct method comparisons between methods difficult.

Our approach addresses these issues in several ways. First, we attempt to identify the origin location of a specific tweet, that is, the location of the user when they composed and delivered the tweet as opposed to the main home location of the user. This definition simplifies API requests and data scale requirements compared to other researchers. Secondly, we focus specifically on the meta-features of the tweet, namely the location and self-description, as opposed to the textual content of the tweet. Finally, we build a fast estimator. Our priority is to quickly and simply locate the origin of a tweet with an accuracy in term of distance to the true location.

3. Methods

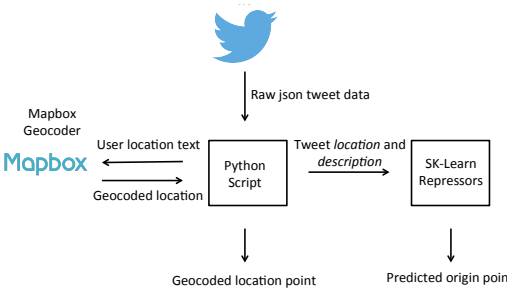
Our dataset consists of tweets collected from April 11 to April 19, 2016. The Twitter API was used to identify users in Oklahoma who supplied location information with their tweets. 325788 tweets were collected during this period. A python program sent the unaltered text in the user 'location' fields to the Mapbox geocoder[3]. In parallel, two Scikit-learn (Pedregosa *et al.*, 2011) SVR regressors were trained on the 'location' and 'user description' fields of the tweets.

4. Experiments

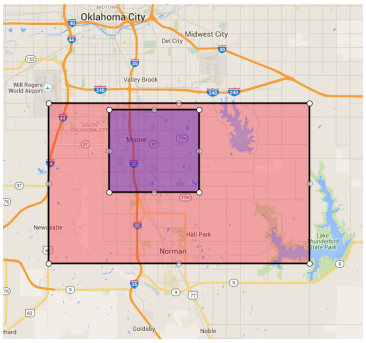
First, we determined what percentage of users in our data set reported a location which could be resolved (not necessarily correctly) with the Mapbox API. We found that approximately 56% of users' locations could be resolved. Next, we examine the accuracy of these locations as compared to the actual origin. The location field is geocoded with the Mapbox geocoder API. If a location is identified, its coordinates are determined via the API. Finally, we built two estimators with the Scikit-learn library to estimate the origin of the tweet based on the meta data to predict the latitude and longitude, respectively. The user's 'location' and 'user description' fields were tokenized into uni- and bi-grams, and each regressor was trained independently on the first 90% of the tweet

[2]Twitter Geolocation `https://dev.twitter.com/overview/terms/geo-developer-guidelines`.

[3]`https://www.mapbox.com/api-documentation/#geocoding`



(a) High level system architecture.



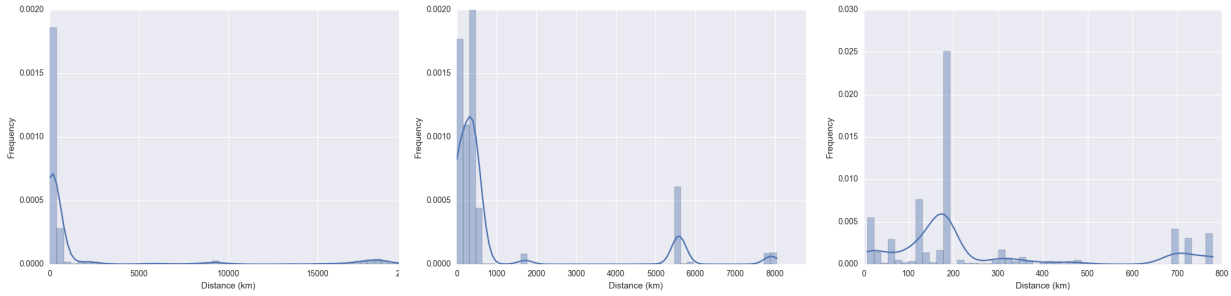
(b) Example tweet Mapbox geocoder and Twitter API bounding box.

Figure 1: Figure 1a is the high level architecture. In Figure 1b the outer, red box represents Mapbox geocoder results. The inner purple box corresponds to the Twitter API location data.

dataset, leaving the remaining 10% for testing.

Since both the geocoder and the Twitter API describe users' locations as bounding boxes, we require an error calculation method that can describe an average distance, and also the potential range of that distance to deal with the problem of overlap. We model the distance between the actual and predicted regions as the average distance between any two random points selected respectively from the two region. We assume that any point within the regions could be the origin point with an equal probability. Therefore, the two bounding boxes are simply two uniform probability distributions, whose mean distance is simply the distance between the two centroids: $\mu_{AB} = \mu_{center_A} - \mu_{center_B}$.

To address the problem of overlap and containment, we propose examining the standard deviation of the mean of μ_{AB} : $\sigma_{AB} = \sqrt{\sigma_A^2 + \sigma_B^2}$, where σ^2 is in the latitudinal or longitudinal direction. If two regions are centered exactly on the same point, the mean distance will be zero, but this calculation reveals how much the two regions differ in size.



(a) Total error. **(b)** Standard deviations. **(c)** Regressor prediction error.

Figure 2: Distribution of error from geocoding.

5. Discussion and Summary

This method of error calculation relies on two assumptions. First, the tweet could have originated from any point with uniform probability. Second, both regions are rectangles, whose respective

sides are parallel. The first holds for relatively small regions (i.e. U.S. state-sized areas) which are far from the poles. The second assumption is satisfied by the fact that the Twitter API returns bounding boxes a few kilometers in size.

We found almost 30% of the tweets with *location* field data can be resolved to city-level accuracy by simply geocoding the text with no alterations. Our mean accuracy was approximately 1941 km, though the long tail distribution in Figure 2a suggests that future work on filtering could be applied to improve this result greatly.

Hecht *et al.* (2011) reported that approximately 34% of users had either non-geographic information or simply nothing entered in the 'location' field. Our results suggest that automated techniques could eliminate these from the dataset and thus improve overall prediction accuracy. We found that locations with a mean distance greater than 500 km had sarcastic or nonsensical locations, for instance: `Im out here` and `most likely school`. These locations were still resolved by Mapbox to physical locations.

The results from the two regressors are intended to demonstrate the usefulness and validity of the error estimation, and to give context to the geocoding results. Two regressors yield a prediction accuracy of 256 km, which is expressed in terms of a mean distance from the actual origin location. Our method yields a reliable prediction error estimation for all points. Classification here, while useful for the correctly classified tweets, is useless for the incorrectly classified data points.

In this work, we made several steps toward location prediction of social media users. We have examined the accuracy and reliability of geocoding users' location information as a way of predicting tweet origin locations. We have also demonstrated the potential utility of this information through machine learning on the textual content. Finally, we have established a robust and straightforward method for accuracy measurement in terms of the distance between the predicted location and the ground-truth location of tweets. In the future, we will engineer several techniques to improve accuracy and evaluate the results on a larger data set.

6. Acknowledgements

Thanks to Le Gruenwald for her helpful comments and the Robert Wood Johnson Foundation for their generous support.

References

Compton, R., Jurgens, D., and Allen, D. (2014). Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 393–401. IEEE.

Culotta, A. (2014). Estimating county health statistics with twitter. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 1335–1344, New York, NY, USA. ACM.

Hecht, B., Hong, L., Suh, B., and Chi, E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 237–246. ACM.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *ICWSM*, 20:265–272.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830.

Priedhorsky, R., Culotta, A., and Del Valle, S. Y. (2014). Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1523–1536. ACM.

Zhang, W. and Gelernter, J. (2014). Geocoding location expressions in twitter messages: A preference learning method. *Journal of Spatial Information Science*, 2014(9):37–70.

Understanding the Mapping Sequence of Online Volunteers in Disaster Response

Yingjie Hu and Krzysztof Janowicz

{yingjiehu,jano}@geog.ucsb.edu
STKO Lab, Department of Geography, University of California, Santa Barbara, USA

Abstract

In recent years, online volunteers have been actively participating in disaster response, thanks to the advancement of information technologies and the support from humanitarian organizations. One important way in which online volunteers contribute to disaster response is by mapping the affected area based on remote sensing imagery. Such online mapping generates up-to-date geographic information which can provide valuable support for the decision making of emergency responders. Typically, the area affected by an disaster is divided into a number of cells using a grid-based tessellation and each volunteer can select one cell to start the mapping. While this approach coordinates the efforts from many online volunteers, it is unclear in which sequence these grid cells have been mapped. This sequence is important because it determines when the geographic information within a particular cell will become available to emergency responders, which in turn can directly influence the efficiency of rescue tasks and other relief efforts. In this work, we study three online mapping projects which were deployed and utilized in 2015 Nepal, 2016 Ecuador, and 2016 Japan earthquakes to gain insights into the mapping sequences performed by online volunteers.

Keywords: Disaster response, crisis mapping, volunteered geographic information

1 Introduction

In recent years, online volunteers have been actively involved in disaster response. On the one side, information and communication technologies allow volunteers to contribute to disaster relief without having to be physically present at the affected areas. On the other side, humanitarian communities, such as Standby Task Force (Meier, 2012a) and Crisis Mapper (Shanley et al., 2013), play an important role in bringing together online volunteers and coordinating their efforts. With the support from technologies and humanitarian organizations, volunteers have made important contributions to 2010 Haiti earthquake (Zook et al., 2010), 2012 Hurricane Sandy in the U.S. (Meier, 2012b), 2013 Typhoon Haiyan in the Philippines (Humanitarian OpenStreetMap Team, 2013), and the 2015 Nepal Earthquake (Hu and Janowicz, 2015).

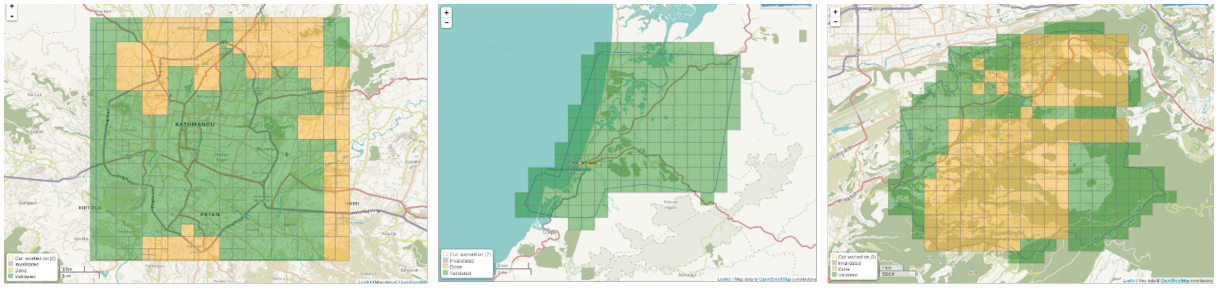
One important way in which volunteers contribute to disaster response is by mapping the affected areas based on remote sensing images. During the online mapping process, volunteers digitize geographic features which may be missing from the previous maps, as well as update the existing geographic data to reflect the current status (e.g., a road may be blocked after an earthquake). This process generates up-to-date geographic information which can provide valuable support for the decision making of emergency responders. While these volunteer-contributed data may not be of highest quality, they generally satisfy the needs of disaster response (Goodchild and Glennon, 2010).

Since many volunteers may be participating in online mapping at the same time, humanitarian organizations often divide the affected area into cells using a grid-based tessellation. Each online volunteer can then select a grid cell to start the mapping task. While this approach helps avoid editing conflicts and duplications, there is a lack of understanding on the sequence in which online volunteers map the grid cells. Such a sequence is important because it directly determines the time when the geographic information within a particular cell will become available. From our observation, the mapping time difference between two neighboring grid cells can be 4 days and sometimes even longer. In disaster response, the first 72 hours after a disaster have been widely considered as the critical period for rescue tasks. After this period, the survival rate drops dramatically (Fiedrich et al., 2000; Comfort et al., 2004; Ochoa and Santos, 2015). Thus, if the grid cells that contain the critical information for disaster response are mapped first, more people can be potentially saved. Intuitively, selecting cells at random is not an ideal solution, since population, transportation infrastructure, and potential rescue routes should be taken into account.

Consequently, the research question that we would like to address here is *whether online volunteers are mapping grid cells in a near-random order or whether they are following certain strategies?* To answer this question, we examine the volunteer mapping sequences in three different projects, namely the 2015 Nepal, 2016 Ecuador, and 2016 Japan earthquake mapping campaigns. In the following sections, we first describe the studied mapping projects and datasets, then present the analysis methods, and finally discuss the results and conclusions.

2 Studied Online Mapping Projects

The three online mapping projects studied in this research are depicted in Fig. 1. The Humanitarian OpenStreetMap Team (https://hotosm.org) is the major organization that provides the mapping platform for online volunteers. As shown in Fig. 1, yellow cells are those that have been mapped by volunteers, and green cells are those that have been both mapped and validated (mapping and validation are performed by different volunteers).



(a) Kathmandu, Nepal (b) Pedernales, Ecuador (c) Kumamoto, Japan

Figure 1: Three online mapping projects studied in this work.

Table 1 summarizes information about the three studied projects. It can be seen that different projects have different numbers of grid cells and participated volunteers. The large number of participants in the Kathmandu project might be explained by the big impact of the 2015 Nepal earthquake: more than 8,000 people died and 21,000 were injured. The cell sizes are predefined by the humanitarian organizations, and can vary among different projects as well as within the same project. The time difference between the earliest and latest mapped cells also varies among the three projects, and in the case of Kumamoto, Japan, a decision maker may need to wait about 76 hours until he/she can obtain the geographic information within the last mapped cell.

Table 1: Information about the three studied projects.

	Kathmandu, Nepal	Pedernales, Ecuador	Kumamoto, Japan
Number of Cells	208	186	340
Varied Cell Sizes	Yes	Yes	No
Number of Volunteers	321	85	52
Earliest Finish Time	2015-04-27 15:20:43	2016-04-23 13:40:13	2016-04-16 21:41:46
Latest Finish Time	2015-04-30 10:57:21	2016-04-25 21:29:25	2016-04-20 01:54:14

3 Analysis Method

To analyze the mapping sequences performed by online volunteers, we retrieved the timestamps when the mapping of the cells has been finished. To examine whether the volunteers are already following certain patterns, we made use of two additional datasets, population and road networks. 2014 LandScan data have been used for population, and road data from OpenStreetMap have been retrieved for the road network. We reproduced the same grid cells based on the three studied projects, and aggregated the population and road length to each grid cell. We then ranked all the grid cells based on the mapping finishing time (from earliest to latest), total population within a grid (from highest to lowest), and total road length within a grid (from highest to lowest).

With the three generated ranks, we examined whether the mapping sequence performed by online volunteers is related to population or road networks using Spearman's correlation coefficient:

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)} \tag{1}$$

where d_i is the difference between two ranks (e.g., the rank based on the mapping time and the rank based on the population), and n is the total number of cells (e.g., 208 for the Kathmandu case).

4 Results and Discussions

The Spearman's correlation coefficients derived from the experiments are summarized in Table 2. It can be seen that the mapping sequences performed by volunteers in the Kathmandu and Kumamoto projects show a *weak to moderate* and significant correlations with the total population and road length. No significant correlations are observed for the Pedernales project, possibly because this project involves a number of cells which cover only the ocean (thereby no population and roads) but which were mapped first by volunteers. To better understand the different rankings for each of the three mapping projects, we also visualize the results as Fig. 2 to 4.

Table 2: Result of the correlation analysis.

	Kathmandu, Nepal	Pedernales, Ecuador	Kumamoto, Japan
Correlation with Population	0.45 ($p < 0.001$)	-0.05 ($p = 0.521$)	0.48 ($p < 0.001$)
Correlation with roads	0.46 ($p < 0.001$)	0.07 ($p = 0.369$)	0.26 ($p < 0.001$)

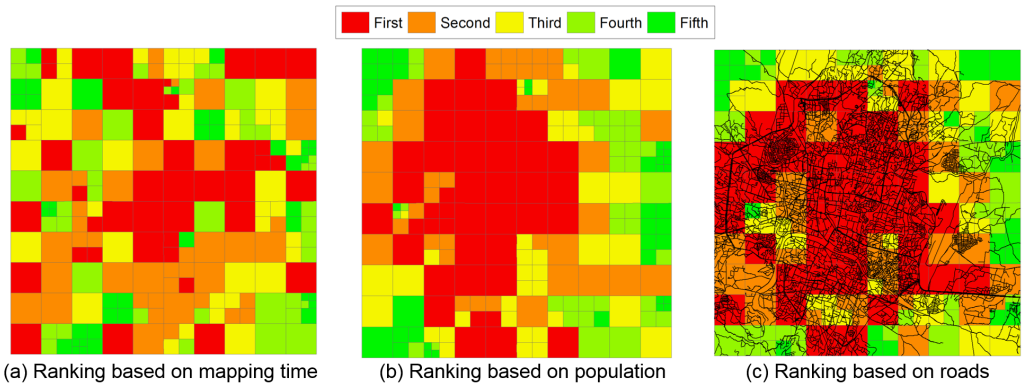


Figure 2: Different rankings for the Kathmandu mapping project.

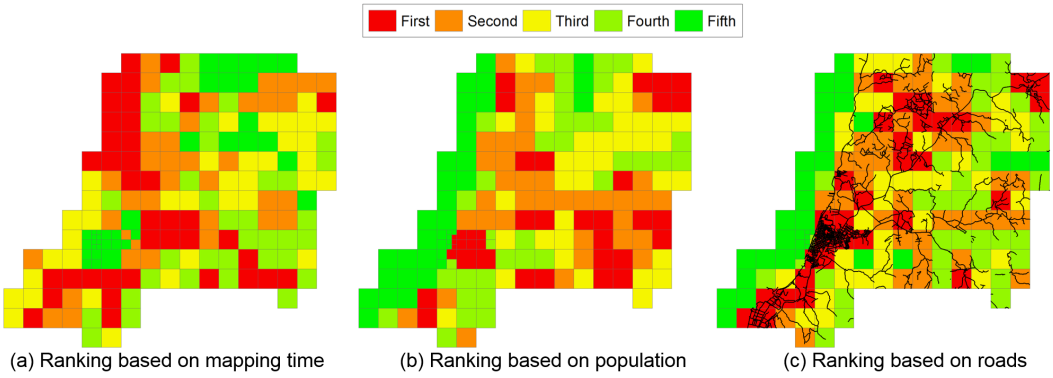


Figure 3: Different rankings for the Pedernales mapping project.

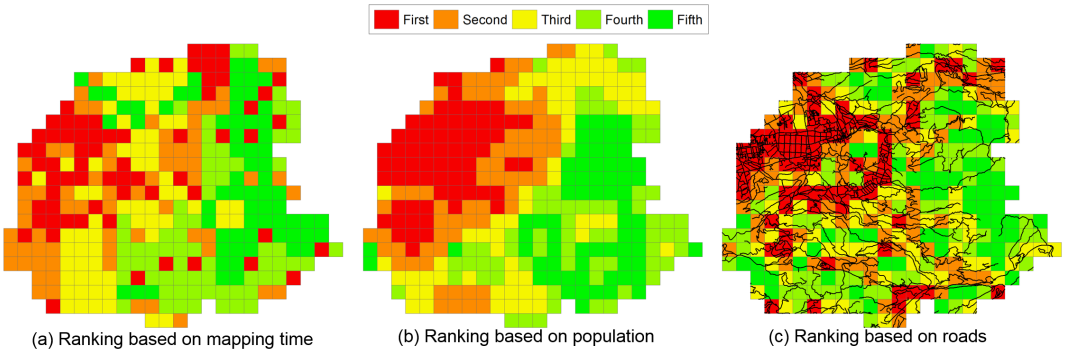


Figure 4: Different rankings for the Kumamoto mapping project.

One can see that the mapping sequence in the Kumamoto project is comparatively similar to the population distribution, whereas the mapping sequence in the Kathmandu shows a weak similarity with

population and road network density. In all three projects, it seems there are some grid cells which were randomly mapped at an early stage (the cells in red color). Overall, we do not observe strong correlations between the mapping finishing time and the population and road networks, indicating room for improvement.

5 Conclusions and Future Work

In this work, we studied the mapping sequences performed by online volunteers in disaster response. Understanding such sequences is important because they determine when the geographic information within a cell will become available to emergency responders. Such time difference can be more than 72 hours, which are beyond the golden period for rescue tasks. Three projects from different countries and different disaster events have been utilized for this study. We described these projects based on the number of grid cells, number of participated volunteers, cell sizes, and earliest and latest mapping finished times. Spearman's correlation analysis has been computed between the mapping sequences and the population distribution as well as the road networks. The results indicate no strong correlation between the mapping sequences and the two additional datasets, thereby pointing to improvements that could be implemented to make disaster mapping more efficient.

Here we have studied three mapping projects (although in total 458 online volunteers were involved in these three projects) and correlated them with two types of additional datasets. More work is needed to understand disaster mapping projects in general and their dynamics over time and study areas. Nevertheless, it is possible that online volunteers may have randomly picked grid cells to start the mapping tasks. Due to the spatial heterogeneity, different grid cells may contain geographic information that is of different importances to the emergency responders. If the grid cells containing more critical information can be mapped first, such information will become available to emergency responders in an earlier stage, which can help potentially save more people. Therefore, we believe that there is a need to prioritize the grid cells based on the values of the contained geographic information, and provide online volunteers with a general guidance for mapping.

References

L. K. Comfort, K. Ko, and A. Zagorecki. Coordination in rapidly evolving disaster response systems the role of information. *American Behavioral Scientist*, 48(3):295–313, 2004.

F. Fiedrich, F. Gehbauer, and U. Rickers. Optimized resource allocation for emergency response after earthquake disasters. *Safety Science*, 35(1):41–57, 2000.

M. F. Goodchild and J. A. Glennon. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3):231–241, 2010.

Y. Hu and K. Janowicz. Prioritizing road network connectivity information for disaster response. In *Proceedings of the 1st Workshop on Emergency Management*, EM-GIS 2015, pages 1–4, New York, NY, USA, 2015. ACM.

O. Humanitarian OpenStreetMap Team. Openstreetmap and yolanda: A report from manila, 2013. URL `https://hotosm.org/updates/2013-12-05 _openstreetmap_and_yolanda_a_report_from_manila`.

P. Meier. Crisis mapping in action: How open source software and global volunteer networks are changing the world, one map at a time. *Journal of Map & Geography Libraries*, 8(2):89–100, 2012a.

P. Meier. What was novel about social media use during hurricane sandy, 2012b. URL `http://irevolution.net/2012/10/31/hurricane-sandy/`.

S. F. Ochoa and R. Santos. Human-centric wireless sensor networks to improve information availability during urban search and rescue activities. *Information Fusion*, 22:71–84, 2015.

L. Shanley, R. Burns, Z. Bastian, and E. Robson. Tweeting up a storm: the promise and perils of crisis mapping. *Available at SSRN 2464599*, 2013.

M. Zook, M. Graham, T. Shelton, and S. Gorman. Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Medical & Health Policy*, 2(2):7–33, 2010.

Travel Pattern Analysis from Trajectories Based on Hierarchical Classification of Stays

Ryo Inoue, Motohide Tsukahara

Graduate School of Information Sciences, Tohoku University, 6-6-06 Aramaki Aoba, Aoba, Sendai 980-8579, Japan
Email: {rinoue, tsukahara}@plan.civil.tohoku.ac.jp

Abstract

It has recently become possible to utilize a large amount of detailed trajectories for travel pattern analysis. However, methods proposed in this area have limitations in applying the data taken over a wide area. To overcome these limitations, this paper proposes extraction of travel patterns through hierarchical classifications of stays and travel patterns based on the Huffman coding algorithm. The results of experiments conducted using the proposed method on trajectories in Okinawa, Japan confirm its feasibility for analyzing travel patterns.

1. Introduction

A huge amount of trajectory data has become available with the widespread use of positioning technology. One of its applications is tourist activity analysis, with research focused on extracting typical travel patterns actively underway.

Previous studies in this area first detect and code stays, then analyze travel patterns from stay sequences. Many approaches have been proposed for the latter process. Zheng *et al.* (2007) analyzed transition probabilities between sites, Giannotti *et al.* (2007) mined sequential patterns, and Shoval and Isaacson (2007) and Shoval *et al.* (2015) used sequence alignment. However, the former process has not been sufficiently investigated. Most of the studies conducted simply judged stays by the preset area classification. Giannotti *et al.* (2007) and Zhang *et al.* (2009) extracted stays via density-based analyses; however, no threshold setting criteria were presented.

Previous analyses are limited especially when the study area is wide. However, because places of interest and arrival and departure times vary, the number of travel patterns is significant. This results in difficulty finding similar patterns. Thus, adjusting the resolution of the analysis to reduce patterns is essential.

This study focuses on "stays," which each consists of a visited place and its arrival and departure time. The analysis resolution of a stay may vary from site- to region-basis in space and from minute- to day-basis in time. Analysis methods whose results change according to resolution settings are useless as they cause difficulty interpreting results. We believe that hierarchical classification of travel patterns to spatio-temporal resolution settings is key to solving the problem. Thus, this paper proposes classification of travel patterns based on Huffman coding of stays, and reports on tests of its applicability using trajectory data obtained in Okinawa, Japan.

2. Hierarchical Classification of Stays and Travel Patterns

Huffman coding outputs compact code with average code length close to Shannon entropy: the average information contained in the data. It constructs a binary tree by repeating the aggregation

of the two least-frequent data elements. The algorithm can classify stays hierarchically; however, only their frequencies are considered, their similarities are disregarded. The stay classification approach proposed in this paper permits grouping stays according to similarity.

2.1 Stay Classification

Let d_{ij} denote the Euclidian distance between the locations of stays i and j, at_i and dt_i denote the arrival and departure times of stay i, respectively, and α denote a weight parameter between spatial and time difference. The proximity of stays is defined by

$$p_{ij} = \alpha d_{ij} + (1-\alpha)\{|at_i - at_j| + |dt_i - dt_j|\}. \ (0 \le \alpha \le 1) \quad (1)$$

By utilizing the link setting condition of relative neighborhood graph, stays are adjudged to have similarity if the following conditions are satisfied:

$$p_{ij} \le \max\{p_{ik}, p_{jk}\} \ \ \forall k \ne i, j. \quad (2)$$

Through the classification of stays, the stay pair with the smallest proximity is grouped first if multiple pairs have the same frequencies, and the similarity of stay groups is evaluated by their elements.

Every node on the tree represents a class, and nodes on a path from the root to leaf nodes represent classes from low to high resolutions.

The classifications are dependent on α. We propose to select α that minimizes the average code length, as it outputs the most compact classification of stays. The average code length in this method is larger than that of Huffman coding; the difference indicates additional information by considering stay similarity.

2.3 Travel Pattern Classification

We classify travel patterns based on stay classification by dividing stay classes individually. Figure 1 indicates an initial state in which the nodes with bold lines signify active classes. Figure 1(a) illustrates a state in which the root node of the stay classification tree is divided into classes #0 and #1, and Figure 1(b) shows a travel pattern classification at this stage: classes with stay #0, #1, and both stays. Figure 2 shows the classification in the next stage. Class #1 that covers a wider range is divided, and travel patterns are hierarchically classified into seven groups.

3. Application

3.1 Trajectory Data, Stay Detection, and Minimum Resolution Settings

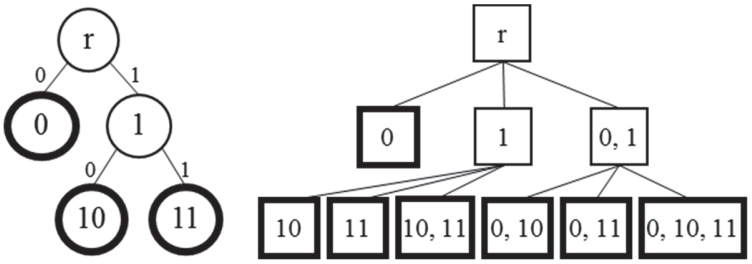
The proposed method was applied to rental car trajectories in Okinawa, Japan. The trajectories consisted of positions at one to five-second intervals observed by onboard GPS devices of 614 tourist groups from August 29 to December 1, 2014 in three to five-day trips (Figure 3).

Stays are defined as states where the tourists stayed in a 100-meter radius circle for more than 15 minutes, and their locations are defined as the centroids of trajectories. Trajectories with less than four stays per day were excluded from farther analysis: 4,167 stays and 823 daily sequences of stays were extracted.

The minimum resolution for analysis was set as follows. Neighbor stay locations were aggregated to analyze visits to the same site by different tourists; the set of locations within 300

(a) Stays. **(b) Travel patterns.**

Figure 1. Classifications when stays are classified in two.



(a) Stays. **(b) Travel patterns.**

Figure 2. Classifications when stay #1 is divided into #10 and #11.

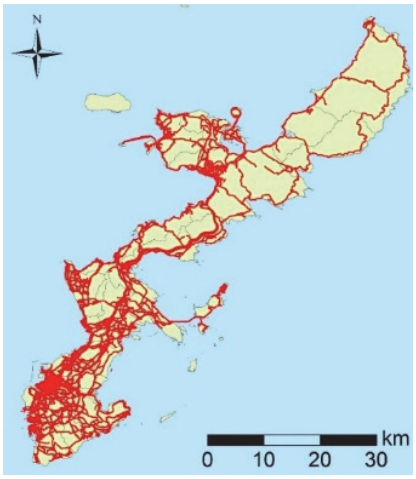


Figure 3. Positions in trajectories.

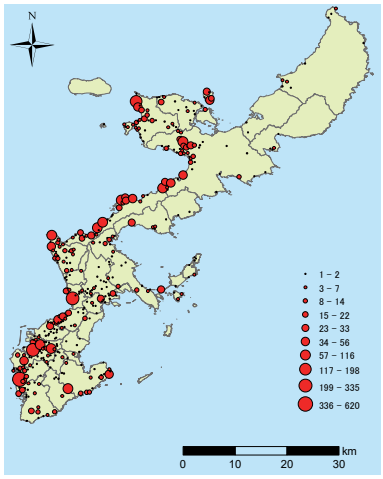


Figure 4. Stay locations and frequencies.

meters were aggregated. Figure 4 shows 399 locations; circles represent locations and their sizes represent visit frequencies. The temporal resolution was set at 10 minutes. On aggregating the same stays, the total number of stays returned was 3,714.

3.2 Hierarchical Classification of Stays and Travel Patterns

α was set to 0.55 to minimize the average code length, by the search within the range from zero to one at 0.05 intervals. When stays were classified in 28 classes, for example, 823 travel sequences classified into 689 patterns, travel patterns became rich in variety. Figure 5 shows the stay classification results and Table 1 shows sample classes. The characteristics of classes are apparent from their locations and arrival and departure times. Further, classes allocated in close proximity to each other on the tree have high similarity. Table 2 shows sub-classes of class “0011” in 156 classifications, a more detailed classification.

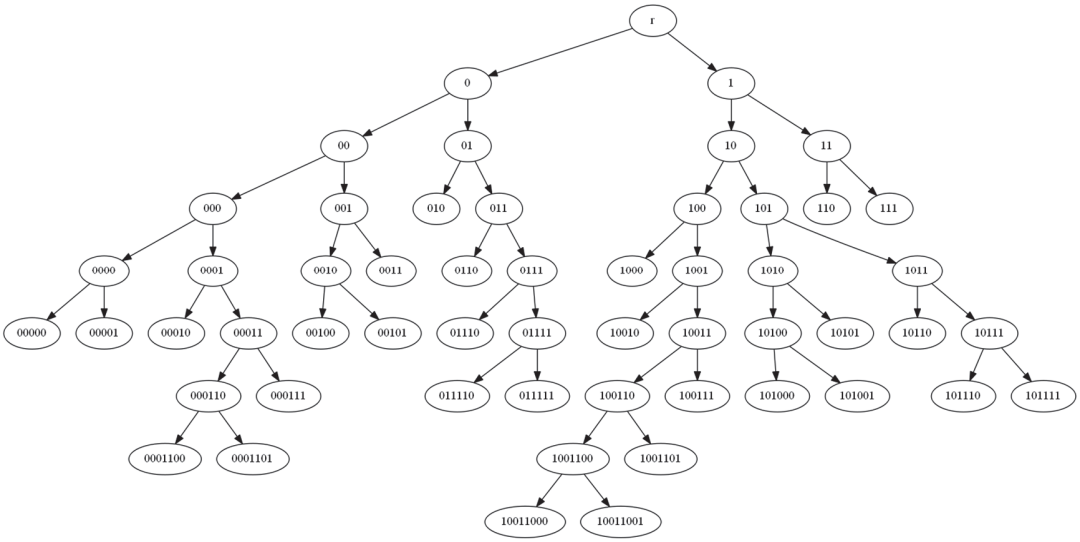


Figure 5. Binary tree of 28 classifications.

Table 1. Stay classes of 28 classifications.

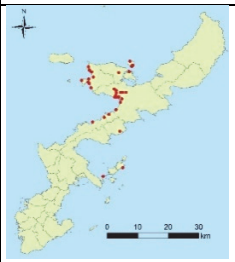
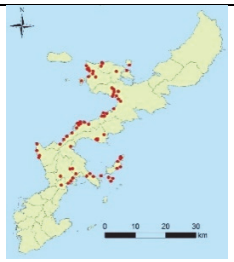
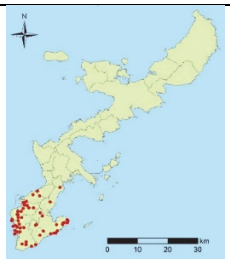
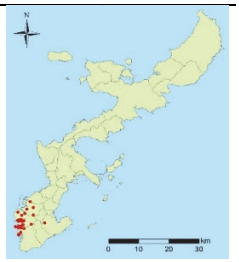
Class	00100	0011	010	0110
Locations				
Frequencies	101	215	414	229
Arrival time: Mean (*SD*)	10:57 (3:24)	16:06 (1:30)	13:48 (2:27)	11:18 (0:50)
Departure time: Mean (*SD*)	12:26 (2:23)	17:05 (1:40)	15:16 (2:20)	12:02 (0:52)

Table 2. Stay classes of 156 classifications.

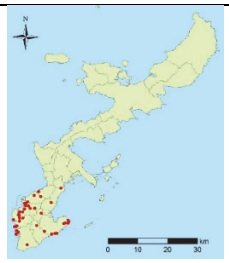
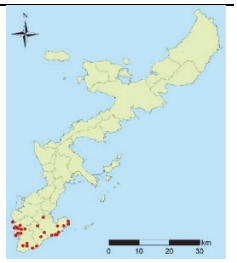
Class	0100	0101
Locations		
Frequencies	210	204
Arrival time: Mean (*SD*)	15:50 (1:28)	11:41 (1:09)
Departure time: Mean (*SD*)	17:18 (1:13)	13:10 (0:55)

Table 3. Most frequent travel patterns at stay classifications in 28 classes.

Travel patterns	Frequencies
010, 0110, 111	8
010, 111	6
0110, 111	5
010, 110, 111	5
010, 0110	5

Table 4. Most frequent travel patterns at stay classifications in 156 classes.

Travel patterns	Frequencies
0110, 1111	3
0101, 0110	3
0100, 0101, 0110	2
0100, 0101, 111011	2
0100, 0101, 111001	2

Tables 3 and 4 show several travel pattern classes, which correspond to the 28 and 156 stay classifications, respectively. Classes {“010,” “0110”} are divided into {“0101,” “0110”} and {“0100,” “0101,” “0110”} as stay “010” is divided into “0100” and “0101.” This confirms that the proposed method can classify travel patterns hierarchically; however, as the travel patterns are diverse, the frequent pattern extraction requires more trajectories.

4. Results

This paper focused on stay classification in travel pattern analyses, and proposed a travel pattern analysis approach from trajectories based on Huffman coding. Experimental results confirm that the proposed method can classify travel patterns; however, the available data were insufficient to discover typical travel patterns. Thus, further analysis with more trajectories is necessary to confirm its effectiveness.

Acknowledgements

Trajectories were provided by the Japan Road Traffic Information Center. This study was supported by JSPS KAKENHI (grants 24241053 and 25249069).

References

Giannotti F, Nanni M, Pinelli F, and Pedreschi D, 2007, Trajectory pattern mining. In: *Proceedings of the 13th ACM SIGKD International Conference on Knowledge Discovery and Data Mining*, 330–339.

Shoval N and Isaacson M, 2007, Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers*, 97:282–297.

Shoval N, McKercher B, Birenboim A, and Ng E, 2015, The application of a sequence alignment method to the creation of typologies of tourist activity in time and space, *Environment and Planning B*, 42(1):76–94.

Zheng Y, Zheng L, Xie X, and Wei M, 2009, Mining interesting locations and travel sequence from GPS trajectories. In: *Proceedings of the International Conference on World Wide Web*, 791–800.

Similarity Measures for Network Time Prisms

Young Jaegal, Harvey J. Miller

Department of Geography, The Ohio State University,
1036 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, U.S.A
Email: {jaegal.1; miller.81}@osu.edu

Abstract

Space-time paths and prisms are time geographic concepts delimiting the actual and potential mobility pattern of an object in space and time, respectively. While there is a range of similarity measures for space-time paths, it is only recently that researchers started to develop similarity measures for the space-time prism (STP). This paper proposes a new methodology for measuring the similarity between network time prisms (NTP), the extension of STP to a moving object on a transportation network. A temporal sweeping method reduces the dimensionality of NTPs to a temporal profile curve that summarizes properties of the NTP subnetwork at moments in time. Then, the existing path similarity measures can measure resemblance between the temporal curves for NTP. We demonstrate the method using selected network summary measures derived from graph theory.

1. Introduction

Time geography provides a powerful spatiotemporal framework for both observed and potential movement pattern of moving objects (Hägerstrand, 1970). For observed movement data, space-time paths capture the actual traces of an object in space with respect to time. For the potential mobility, *space-time prisms* (STPs) delimit the geographic limit of movement of an object with respect to time. Researchers have developed a variety of path similarity measures for quantifying the resemblance between two space-time paths (Long and Nelson, 2013; Ranacher and Tzavella, 2014; Yuan and Raubal, 2014). Previous studies have utilized these measures for various purposes including comparing, clustering and aggregating movement trajectories. However, it is only recently that researchers started paying attention to similarity measures for the STP.

Miller, Raubal and Jaegal (2016) develop a temporal sweep approach for measuring STP similarity. The basic idea is to measure selected properties of a STP at moments in time, reducing the dimensionality of a STP to a profile curve summarizing changes in the selected property over time. This paper extends the temporal sweep method for measuring STP similarity to the case of *network time prisms* (NTPs). We also provide a demonstration of the method using an empirical transportation network.

2. Background

2.1 Network time prism

The NTP is an extension of STP that models the potential mobility of an object moving within a transportation network. Since individual movement in the real world usually occurs within networks, the NTP can provide a more realistic model of accessibility than a planar STP. Analytically, the NTP between locations and times (x_i, y_i, t_i) and (x_j, y_j, t_j) is the collection of 3-tuples consists of space-time coordinates, (*x, y, t*), that satisfies the following three conditions (Kuijpers and Othman, 2009).

$$\begin{cases} (x,y) \in \boldsymbol{RN} \\ d_{RN}\big((x_i,y_i),(x,y)\big) + d_{RN}\Big(\big(x_j,y_j\big),(x,y)\Big) \le \big(t_i - t_j\big) \\ t_i + d_{RN}\big((x_i,y_i),(x,y)\big) \le t \le t_i - d_{RN}\Big(\big(x_j,y_j\big),(x,y)\Big) \end{cases} \quad (1)$$

Where ***RN*** is the set of locations on a road network; $d_{RN}(\underline{\boldsymbol{p,u}})$ is the shortest path travel time between two locations within ***RN***.

2.2 Path and Prism Similarity Measures

Path similarity measures provide a quantitative assessment of resemblance among paths based on geometric and sequential properties. Two major classes of path similarity measures are shape-based and time-based (Yuan and Raubal 2014). Shape-based measures consider only the geometric characteristics of a path. This includes classical Euclidean distance and Hausdorff distance. Time-based methods conceptualize a space-time path as multidimensional time series data. It includes synchronized Euclidean distance, Fréchet distance, dynamic time warping (DTW) and longest common sub-sequences (LCSS) (See Long and Nelson (2013) for the review of these measures). The applications of path similarity in transportation studies include clustering methods for movement trajectories, comparison of individual mobility patterns, and querying for paths in a database that are similar to a reference path.

STPs have been widely used in human and ecological studies as measure of individual accessibility and exposure to environments (Espeter and Raubal 2009; Long and Nelson 2012). They can also measure space-time path uncertainty based on sampled locations (Pfoser and Jensen 1999). As with space-time paths, we also may wish to measure similarity between STPs for assessing differences in accessibility, exposure and mobile object location uncertainty. We also may wish to cluster, aggregate and search for similar prims based on geometric and/or semantic properties. However, until recently, researchers have paid little attention to similarity measures for STP.

Miller, Raubal and Jaegal (2016) develop a methodology for measuring STP similarity based on dimensionality reduction. They temporally swept STPs to generate time series data of geometric and semantic properties of STPs. We can assess STP similarity by applying existing path similarity measures such as DTW to the resulting temporal profile curves. The next section of this paper describes the temporal sweep method and its extension to measure NTP similarity.

3. Methodology

3.1 Temporal Sweeping Method

Figure 1 provides an illustration of the general strategy. The method records a selected property of the NTP at discrete moments in time. The left side of Figure 1 presents an example NTP constructed by network analysis tool available in ArcGIS 10.3. The right side of Figure 1 shows that we can reduce the dimensionally of NTP into a temporal profile curve. Since this sequence is a time series, existing path similarity measures discussed in the previous section can assess the resemblance between different temporal curves.

We can measure and record a wide range of relevant NTP properties to generate the temporal profile curves. These include geometric properties such as the size and shape of the spatial region covered by the NTP, as well as semantic properties such as the spatial distribution of activities, resources and people within that region. In this paper, we apply graph theoretic measure to assess the topology and connectivity of the NTP subnetwork at each moment in time. We will now discuss these measures.

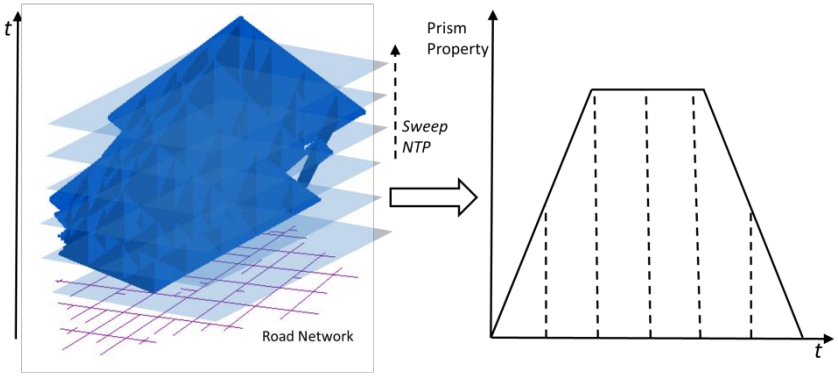


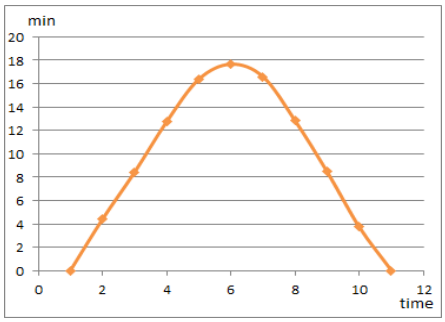
Figure 1. Temporal sweeping for NTP

3.2 Graph Theory

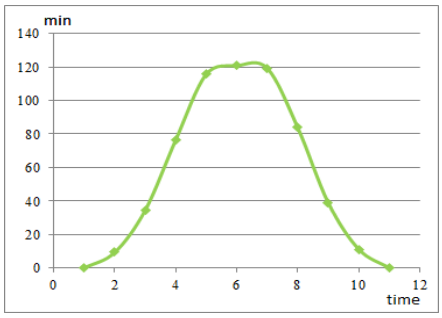
Since the temporal sweeping method for NTPs generates a series of subnetworks at each moment in time, we can use analytical tools available from graph theory to assess properties of NTP for similarity assessment. A graph is a symbolic representation of network using vertices, *v*, and edges, *e*. (Rodrigue, Comtois and Slack, 2013). Graph theory is a mathematical study of the encoding and measurement of the graph.

Two key aspects of a network affecting mobility are size and connectivity. The simple graph theoretic measures for network size include the number of vertices, e_n, and edges, v_n, but they provide no information about travel time between vertices. Other size measures considering transportation times on the network include cost, diameter and Pi index. *Cost* is the sum of network travel time of all edges on a network. *Diameter* is the shortest path between the two farthest apart vertices in the graph. The *Pi* index is the ratio of cost to diameter. The connectivity measures from graph theory include Beta index and Gamma index. The *Beta* index is the ratio of the number of vertices to edges. It ranges from zero for a network with no edges and increases in value with greater connectivity. It exceeds unity for more complex networks with more than one circuit or path that begins and ends at the same vertex (Bhaduri, 1992). The *Gamma* index uses the number of all possible edges instead of the number of vertices as the numerator.

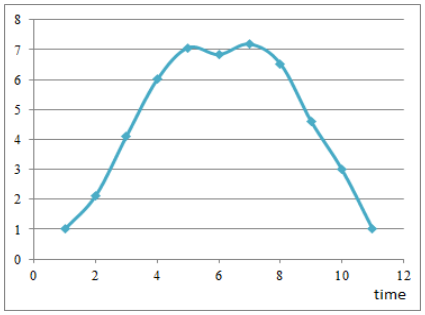
We apply these measures to the example NTP in Figure 1 to demonstrate the concept. Figure 2 presents the temporal curves of selected graph theoretic measures. Figure 2a and 2b illustrates the diameter and cost capturing effectively the change in the size of NTP with respect to time. The Pi index curve in Figure 2c also captures the changes in network size, and it would be useful when comparing NTP from different networks since it is standardized by the diameter of the network. Lastly, Figure 2d shows that Beta index capturing the change of connectivity over time in the NTP.



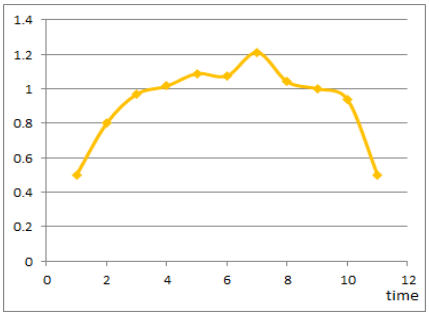
(a) Diameter



(b) Cost



(c) Pi Index



(d) Beta Index

Figure 2. Temporal profile curves for the example NTP

4. Concluding remarks

This paper develops a methodology for measuring the similarity between NTPs based on a temporal sweeping method for reducing the dimensionality of NTP and demonstrates the approach using selected graph theoretical measures for assessing the size and connectivity of the NTP over time. We can compare other NTP properties in a similar manner, such as the geometric size and shape of the space-time regions, and the spatial distribution of activities, resources contained within the NTP at different moments in time.

The next steps for this research include assessing the performance of a wider range of graph theoretic measures for different types of networks and NTPs. After assessing the performance of different measures for carefully crafted experimental networks, the next step is to apply the methodology to empirical transportation networks and NTPs. Finally, we will assess the performance of the similarity measures in NTP clustering and aggregation.

References

Bhaduri S, 1992, *Transport and Regional Development: A Case Study of Road Transport of West Bengal*. Concept Publishing Company

Espeter M and Raubal M, 2009, Location-based decision support for user groups. *Journal of Location Based Services*, 3(3): 165-187.

Hägerstrand T, 1970, What about people in regional science? *Regional Science Association. Papers and Proceedings*, 24: 7-21

Kuijpers B and Othman W, 2009, Modeling uncertainty of moving objects on road networks via space–time prisms. *International Journal of Geographical Information Science*, 23(9): 1095-1117

Long, JA and Nelson TA, 2012, Time geography and wildlife home range delineation. *The Journal of Wildlife Management*, 76(2): 407-413

Long JA and Nelson TA, 2013, A review of quantitative methods for movement data. *International Journal of Geographical Information Science*, 27(2): 292-318

Miller HJ, Raubal M and Jaegal Y, 2016, Measuring space-time prism similarity through temporal profile curves. In *Geospatial Data in a Changing World*, Springer International Publishing, 51-66.

Ranacher, P. and Tzavella, K, 2014, How to compare movement? A review of physical movement similarity measures in geographic information science and beyond. *Cartography and geographic information science*, 41(3), 286-307

Rodrigue JP, Comtois C and Slack B, 2013, *The Geography of Transport Systems*. Routledge, Abingdon, Oxon, UK

Yuan Y and Raubal M, 2014, Measuring similarity of mobile phone user trajectories: A spatio-temporal edit distance method. *International Journal of Geographical Information Science*, 28: 496-520

Applied CyberGIS in the Age of Complex Spatial Health Data

M. M. Jankowska[1], J. Schipperijn[2], J. Kerr[1], I. Altintas[1]

[1]University of California San Diego, 9500 Gilman Drive La Jolla CA 92093-0811 USA
Email: {majankowska; jkerr}@ucsd.edu; altintas@sdsc.edu

[2]University of Southern Denmark, Campusvej 55 Odense 5230 Denmark
Email: jschipperijn@health.sdu.dk

Abstract

Advances in data acquisition in a number of fields throughout the health spectrum are resulting in large, complex, and diverse data sets. With proliferation of sensor and spatial data acquisition, the analysis and processing of complex spatial health data analytics will become a pressing problem. CyberGIS can offer solutions in this realm, however no systems have been developed that cater to the specific challenges associated with complex spatial health data such as privacy, real-time analytics, data standardization, data integration, workflow provenance, and a front end interface that is accessible to individuals in the public health realm. In this paper we present SPACES, an in-development CyberGIS to address some of these challenges. We discuss the architecture, and present two test case examples of utilizing SPACES for understanding environmental influences on physical activity.

1. Introduction

Public health, health care, and medical advances are increasingly looking to the collection of voluminous data sets. Movements such as Quantified Self, where individuals engage in self-tracking of biological, physical, behavioural, and environmental information (Fawcett 2015), or the Precision Medicine Initiative Cohort announced by President Obama in 2015, which will include extensive data collection and tracking of over one million U.S. participants (Ashley 2015), are driving the need to develop cyberinfrastructures and methodologies for big and complex spatial health data integration, processing, and analysis. Complex spatial health data may include temporally-linked objective measures of behavior and health, molecular data such as genomics, proteomics, or metabolomics, life course data including movement trajectories, self-reported survey and demographic data, environmental data assessing exposures, and finely resolved spatial data.

CyberGIS may assist in addressing the challenges of complex spatial health dat. It represents a new-generation GIS based on the synthesis of advanced cyberinfrastructure, geographic information science, and spatial analysis and modeling (Wang 2010). The field of CyberGIS is rapidly developing and the US National Science Foundation funded a 5-year major initiative titled 'CyberGIS Software Integration for Sustained Geospatial Innovation' (http://cybergis.org). However, advances in the application of CyberGIS to health specific problems are lacking. Goldberg *et al.* (2013) presented a comprehensive vision of a Spatial-Health CyberGIS Marketplace, which tackled many of the opportunities as well as challenges of a health focused CyberGIS including confidentiality and privacy protections, real-time analytic methods, data standardization, and a comprehensive end-to-end ecosystem architecture. We would add to this list the need for shareable workflows to promote inter-field collaboration, diverse data type integration, and replicability of analytic processes.

Recently, the NSF funded a platform called DELPHI (Data E-platform for personalized population health) (Katsis *et al.* 2013), one of the first cyberinfrastructures in development specifically catered toward meeting the goals of integrating complex and diverse health data

with an accessible API for individual use, clinician intervention and patient management, and population-level research. However, DELPHI has not been developed to operate in a HIPAA-compliant environment (HIPAA Privacy Rule establishes national standards for protection individual medical records and personal health information), it does not cater to the complexities of performing analyses on spatial data, and finally it does not offer workflows to promote collaboration and replication of processing and analytic tasks. For this reason, we have integrated DELPHI into a larger CyberGIS called SPACES (Secure Physical ACtivity and Environment Software).

2. SPACES

SPACES is an in-development CyberGIS specifically geared at solving and promoting data integration, analysis, and collaboration with complex spatial health data. SPACES aims to address four core problems in existing CyberGIS applications for health: 1) conceptual and applied challenges of complex spatial and biomedical data integration, data integrity, and common data standards, 2) lack of easy-to-use data integration computation infrastructure, 3) lack of documented and shared workflows that will promote scalability, provenance (understanding of origin of results and repeatability of scientific process), and knowledge base development, and 4) lack of systems designed to function in HIPAA-compliant environments.

Here, we focus on one specific application of complex spatial health data. The current obesity and inactivity epidemics have instigated a surge of research into the spatial factors that influence physical inactivity and obesity (Berrigan *et al.* 2015). Technological and methodological developments have led to the ability to examine dynamic, high-resolution measures of temporally-matched location and physical activity behavior data through GPS, accelerometry, and GIS (Jankowska *et al.* 2015), a nascent field called 'spatial energetics' (Jankowska *et al.* 2016). This field sees two promising paths by improving population health through environmental modifications and improving individual health through targeted mobile-based health interventions (*Sallis et al.* 2006; Adams *et al.* 2013). Spatial energetics offers an ideal case study for the development of CyberGIS for complex health data by offering different types of data collection methods (real-time, data logging, individual self-report, survey), diverse levels of data collection (biological, minute level, individual demographic, community environment), large data sets (data sampled at the sub-minute level on large populations quickly grows to terabytes of data), and often focuses on health outcomes beyond obesity such as cancer or diabetes that requires HIPAA privacy protections.

SPACES is designed to support and train researchers on the spatial and temporal analysis of large volumes of physical activity, geographic, and contextual data. It is housed in a HIPAA and FISMA-compliant, private computing cloud housed at the San Diego Supercomputer Center called Sherlock; however the structure can be replicated in other HIPAA compliant clouds. The key organizing principle of SPACES is workflows, and a unified data communication layer through which the development of common data elements can be achieved. The workflows are provided by Kepler Scientific Workflow System (Altintas *et al.* 2006), which provides a graphical user interface for designing workflows composed of a linked set of extensible and configurable components. Kepler can also call on distributed programming that supports interfacing to different components of the cyberinfrastructure and computing platforms. In SPACES, Kepler workflows are a programming abstraction that describes which computational elements (serial or parallel) and data elements need to fit together, and what order they must be processed in. In Figure 1 below, an example of such a workflow is given with processing and analysis blocks working through the integration of minute level accelerometer, GPS, and heart rate data.

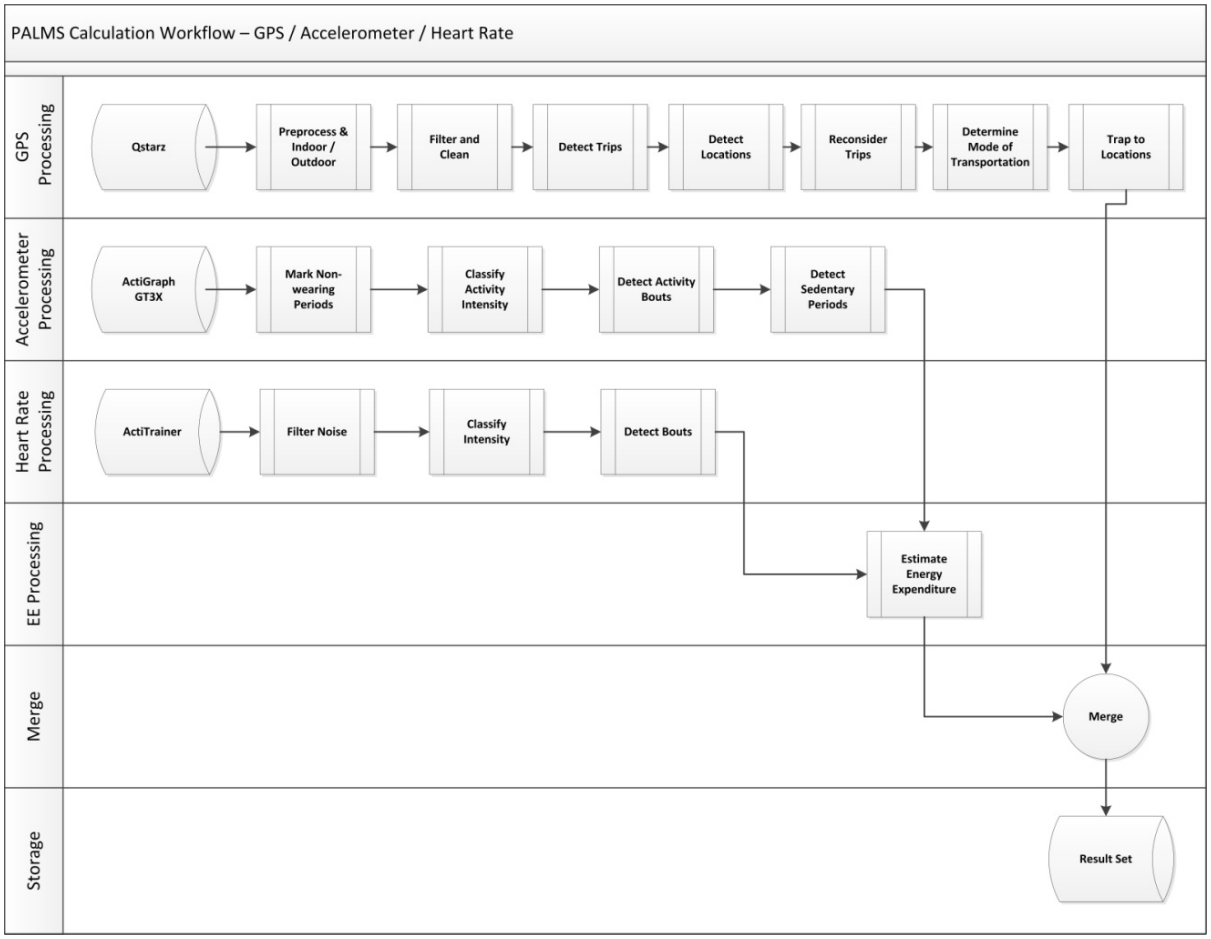


Figure 1. An example workflow of GPS, accelerometer, and heart rate data.

Sitting below the workflow(s) are the systems and software deployments required for reliable execution, as well as the data communication layer for efficient data organization with is built upon the existing DELPHI infrastructure and adds in a PostgreSQL geodatabase for spatial data. A key element of SPACES is the creation of common data elements through existing physical activity and spatial energetics processing algorithms such as the Personal Activity Location Measurement System (Demchak *et al.* 2012), machine learning algorithms (Ellis *et al.* 2014), and spatial element extraction and analysis (Thierry *et al.* 2013). Common data elements and outputs can directly enhance and influence the training efforts for the health research community and promote common data organization standards. The use of the Sherlock environment as a platform (or another HIPAA compliant platform) where all data integration and analysis takes place ensures security of sensitive data while allowing access to users to analyze derived datasets without compromising the raw data. Finally, a friendly user front end with training modules and ability to share workflows will be developed to make SPACES accessible to various researchers and end users.

3. Applications

In this section we will briefly detail two current applications for the SPACES system. A group of researchers at UCSD recently completed a walking intervention for over 300 older adults living in retirement communities throughout San Diego County with intervention activities occurring at the individual, interpersonal, and community levels. The intervention included measurement of walking behaviour with accelerometer and GPS devices for one week periods at each of the five time points of the study. The team was interested in assessing the amount of change in walking behaviour that occurred over the intervention period both as

an assessment of per person walking time increase, but also spatially to see if community-level interventions such as cleaning up walking paths promoted walking both on the retirement campuses and in the surrounding community. PALMS accelerometer and GPS match algorithms and geodatabase processing algorithms were employed in the SPACES environment to assess change over time in walking behaviours of the older adults. Figure 2 illustrates results for one retirement site, demonstrating an increase in walking behaviours on the retirement campus as well as along a recommended walking path by the beach. Because SPACES employs a PostgreSQL geodatabase structure for spatial data, results can be directly exported to ArcGIS or QGIS sitting in the Sherlock environment for visualization purposes.

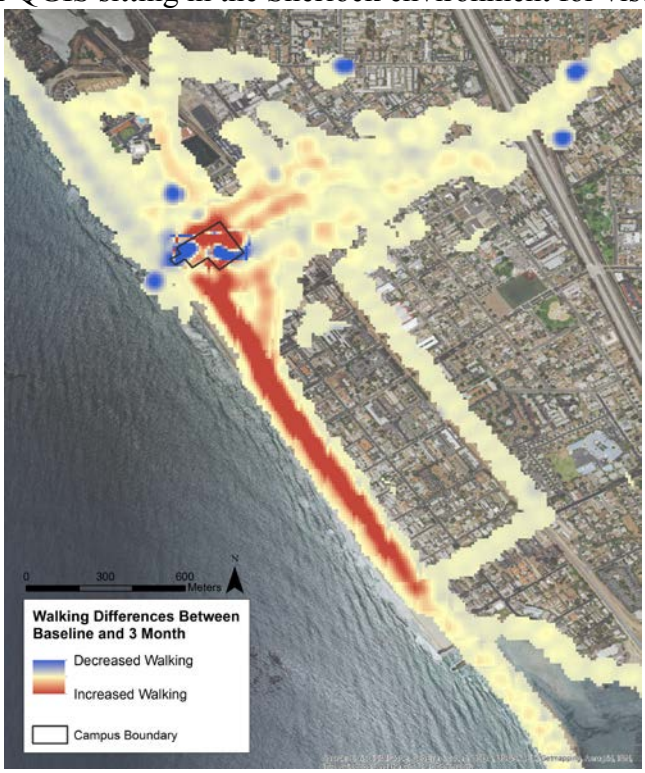


Figure 2. One intervention site and surrounding area; difference in walking as measured by accelerometers and GPS between baseline and 3 months into intervention.

In Denmark, the University of Southern Denmark used the SPACES environment among others to evaluate the effect of renovating 7 schoolyards on students' physical activity behaviour during school recess. The team was interesting in assessing the amount of change in moderate to vigorous physical activity as well as time spent sedentary as a result of the intervention. PALMS accelerometer and GPS match algorithms and geodatabase processing algorithms were employed in the SPACES environment to assess change over time in physical activity and sedentary behaviour. Physical activity behaviour patterns were analysed both on the individual level, but also for each schoolyard to determine which schoolyard elements were most conducive to physical activity. In Figure 3 a vegetated area of the schoolyard proves to be one of the most popular locations for physical activity.

Figure 3. Physical activity hotspots of children from one schoolyard with an unexpected activity hot spot on school grounds – a vegetated trail.

Acknowledgements

This research was supported by the National Institutes of Health National Cancer Institute research grant R01 CA179977-01, Principle Investigator Jacqueline Kerr, titled "PQA4: GPS exposure to environments and relations with biomarkers of cancer risk". Additional funding was provided by the National Institutes of Health Transdisciplinary Research on Energetics and Cancer Grant U54 CA155435-02, pilot funding awarded to PI Marta Jankowska, titled "Assessing the impact of spatial uncertainty on relationships between the built environment and physical activity and sedentary behavior."

References

Adams M, Sallis J, Norman G, Hovell M, Hekler E, and Perata E, 2013, An adaptive physical activity intervention for overweight adults: a randomized controlled trial. *PloS One*, 8(12).

Altintas I, Barney O, and Jaeger-Frank E, 2006c Provenance Collection Support in the Kepler Scientific Workflow System. *Work*, 4145: 118–132.

Ashley E, 2015, The precision medicine initiative: a new national effort. *JAMA*, 21:E1-2.

Berrigan D, Hipp A, Hurvitz P, James P, Jankowska M, Kerr J, Laden F, Leonard T, McKinnon R, Powell-Wiley T, and Tarlov E, 2015, Geospatial and contextual approaches to energy balance and health. *Annals of GIS*, 21(2):157–168.

Demchak B, Kerr J, Raab F, Patrick K, and Kruger I, 2012, PALMS: A modern coevolution of community and computing using policy driven development. *45th Hawaii International Conference on System Sciences*: 2735–2744.

Ellis K, Kerr J, Godbole S, Wing D, and Marshall S, 2014. A random forest classifier for the prediction of energy expenditure and type of physical activity from wrist and hip accelerometers. *Physiological Measurement*, 35(11):2191.

Fawcett T, 2015, Mining the quantified self: Personal knowledge discovery as a challenge for data science. *Big Data*, 3(4):249–266.

Goldberg D, Cockburn M, Hammond T, Jacquez G, Janies D, Knoblock C, Kuhn W, Pultar E, and Raubal M, 2013, Envisioning a future for a spatial-health CyberGIS marketplace. *Proceedings of the Second ACM SIGSPATIAL International Workshop on the Use of GIS in Public Health*:27–30.

Jankowska M and James P, Marx C, Hart J, Berrigan D, Kerr J, Hurvitz P, Hipp J, and Laden F, 2016, Unanswered questions in "spatial energetics" research. *American Journal of Preventive Medicine*. Accepted

Jankowska M, Schipperijn J, and Kerr J, 2015, A framework for using GPS data in physical activity and sedentary behavior studies. *Exercise and Sports Science Reviews*, 43(1):48–56.

Katsis Y, Baru C, Chan T, Dasgupta S, Farcas C, Grisowld W, Huang J, Ohno-Machado L, Papakonstantinou Y, Raab F, and Patrick K, 2013, DELPHI: Data e-platform for personalized population health. In *IEEE Workshop on Service Science and Health (SSH)*, Lisbon.

Sallis J, Cervero R, Ascher W, Henderson K, Kraft K, and Kerr J, 2006, An ecological approach to creating active living communities. *Annual review of public health*, 27:297–322.

Thierry B, Chaix B, and Kestens Y, 2013, Detecting activity locations from raw GPS data: a novel kernel-based algorithm. *International journal of health geographics*, 12:14.

Wang S, 2010, A CyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers*, 100(3):535–557.

Scalability in Participatory Planning: A comparison of online PPGIS methods with face-to-face meetings

Piotr Jankowski[1,2], Michał Czepkiewicz[2], Marek Młodkowski[2] Michał Wójcicki[2] Zbigniew Zwoliński[2]

[1]San Diego State University, Department of Geography San Diego, CA 92182-4493
Email: pjankows@mail.sdsu.edu

[2]Institute of Geoecology and Geoinformation
Adam Mickiewicz University in Poznań, Poznań, Poland

1. Introduction

A traditional approach to participatory planning involves face-to-face engagement of interested public in the tradition of town-hall meetings. A limitation of this approach has been its inability to scale public participation out to involve more people and up to involve participants from a wider geographical area (Nyerges and Aguirre 2011, Innes and Booher 2004). Although online Public Participation GIS (PPGIS) methods are not a panacea for scaling public participation, they offer a potential for collecting views and preferences of some residents who typically do not participate in open planning processes. The questions of who participates and whose views are represented have been at the heart of theorizing about PPGIS (Schlossberg and Shufford 2005, Sieber 2006), and were highlighted as core research questions in a recent review (Brown and Kyttä, 2014).

This paper contributes to the literature by comparing online PPGIS and face-to-face methods in two aspects of scalability: number of participants (scaling out) and spatial extent (scaling up). Additionally, the demographic representation across the methods is also assessed. The evaluation is based on an empirical study involving four participatory planning processes, which took place between May 2014 and July 2015 in the City of Poznań (pop. 554 thousand), Poland. The processes were focused on local land use plan for a centrally located, multi-functional area in the City of Poznań, including a park, recreational facilities, allotment gardens, single- and multi-family housing, and a public school. Two of the four processes were traditional town-hall meetings (May 2014 and June 2015) whereas the other two (October 2014 and July 2015) employed online PPGIS applications. The dichotomy between the two modes of participatory planning (same-place/same-time and distributed) affords a unique opportunity to compare demographic and spatial scalability of face-to-face meetings with online PPGIS methods.

2. Methods

Both of traditional, town-hall style meetings were organized by the municipal planning office in Poznań. The purpose of the first meeting (May 2014) was to familiarize participants with the planning procedure and to facilitate a discussion, during which participants had an opportunity to voice their concerns, opinions, and expectations regarding the planning process. During the

second meeting (June 2015) the participants were presented with a draft project of land use plan, and commented on land use designations and functions. Information about the participants including their age and address were collected by researchers through a questionnaire given to the participants during both meetings.

The two online PPGIS Web applications involved a geo-questionnaire (October 2014) and a geo-discussion (July 2015). Geo-questionnaire is an online questionnaire coupled with an interactive map facilitating data collection of two types: object descriptions linked to geographical features, and descriptions without an explicit spatial reference (Jankowski et al. 2016). In geo-questionnaire the geographical features are sketched by participants or selected from an interactive map. Depending on the geo-questionnaire design, sketching or selecting a geographical feature may trigger one or more questions pertinent to feature's location.

Geo-discussion is a Web application tool developed by the authors, consisting of an online structured discussion forum coupled with an interactive map with rich geographical context. The tool allows participants to annotate map objects with discussion contributions. The discussion forum has several standard functions, such as adding new threads on a map, commenting on threads added by other participants, subscribing to posts and threads, adding "like" and "dislike" reactions, adding attachments, and sorting and searching posts. The map module allows participants to measure distances and surfaces, search by address, toggle between map layers, filter objects from a selected area, and retrieve attribute information about land use plan.

The geo-questionnaire was used to collect the preferences of Poznań residents in regard to the plan area and its future land use organization. The details describing the geo-questionnaire recruitment process, response rate, content, and data concerning public preferences have been reported in Jankowski et al. (2016). The geo-discussion was carried out to collect comments and ideas from Poznań residents about the project of local land use plan for the park including land use designations and functions.

3. Results

Table 1 provides the comparison of face-to-face and online modes of participation by participant number, age, gender, and college education attainment. The differences are clear in the number of participants and their mean age. The online mode attracted on average between five (geo-discussion) and 40 times (geo-questionnaire) more participants than the face-to-face meeting mode. The online participants were on average 7 to 10 years younger than the meeting participants. There is no or very little difference in gender breakdown and percentage of college educated participants.

Table 1. Comparison of traditional with online participation mode.

Participation method		No. of participants	Mean age of participants	Percent women	Percent with higher education
Public meetings	1st meeting	32	45.4	37.5	90.6
	2nd meeting	23	42.5	47.8	76.2
Geo-questionnaire		1087	35,8	47.7	74.5
Geo-discussion		128	35,3	46.9	79.5

Spatial extent and the distribution of participation rates are presented in Figure 1. Three main differences in the distribution and the overall spatial extent of participation are noteworthy: 1) the rates of participation in the online modes were on average four times higher than for the face-to-face mode, 2) residents from 37 out of 42 city neighborhoods participated in the online mode in comparison to residents from 10 neighborhoods in the face-to-face mode, and 3) the online participation mode tracks better the city population by neighborhood than the face-to-face mode. In all methods, there is a disproportionally higher participation from the neighborhood where the plan area is located and from the adjacent neighborhoods. A preliminary spatial analysis of participant distribution based on Average Nearest Neighbor Index shows statistically significant clustered patterns with higher clustering of participants observed in the online mode (Table 2). This obviates two things; first, that participation in a local planning problem is also local - regardless of the mode of participatory process. Second, surprisingly the online mode did not produce a geographically dispersed pattern of participation despite a geographically wider (than the face-to-face mode) participation extent. This suggests that participants self-selected for the survey based on their interest in the area covered by the plan.

Table 2. Average Nearest Neighbor Index of participant distribution for the Poznań area.

Method	Observed mean distance [m]	Expected mean distance [m]	Nearest Neighbor Index	N	Z-score	p-value
First meeting	627.85	1907.25	0.329	18	-5.445	0.000
Second meeting	565.05	1962.55	0.288	17	-5.617	0.000
Geo-discussion	395.55	872.56	0.453	86	-9.699	0.000
Geo-questionnaire	238.14	425.30	0.560	362	-16.017	0.000

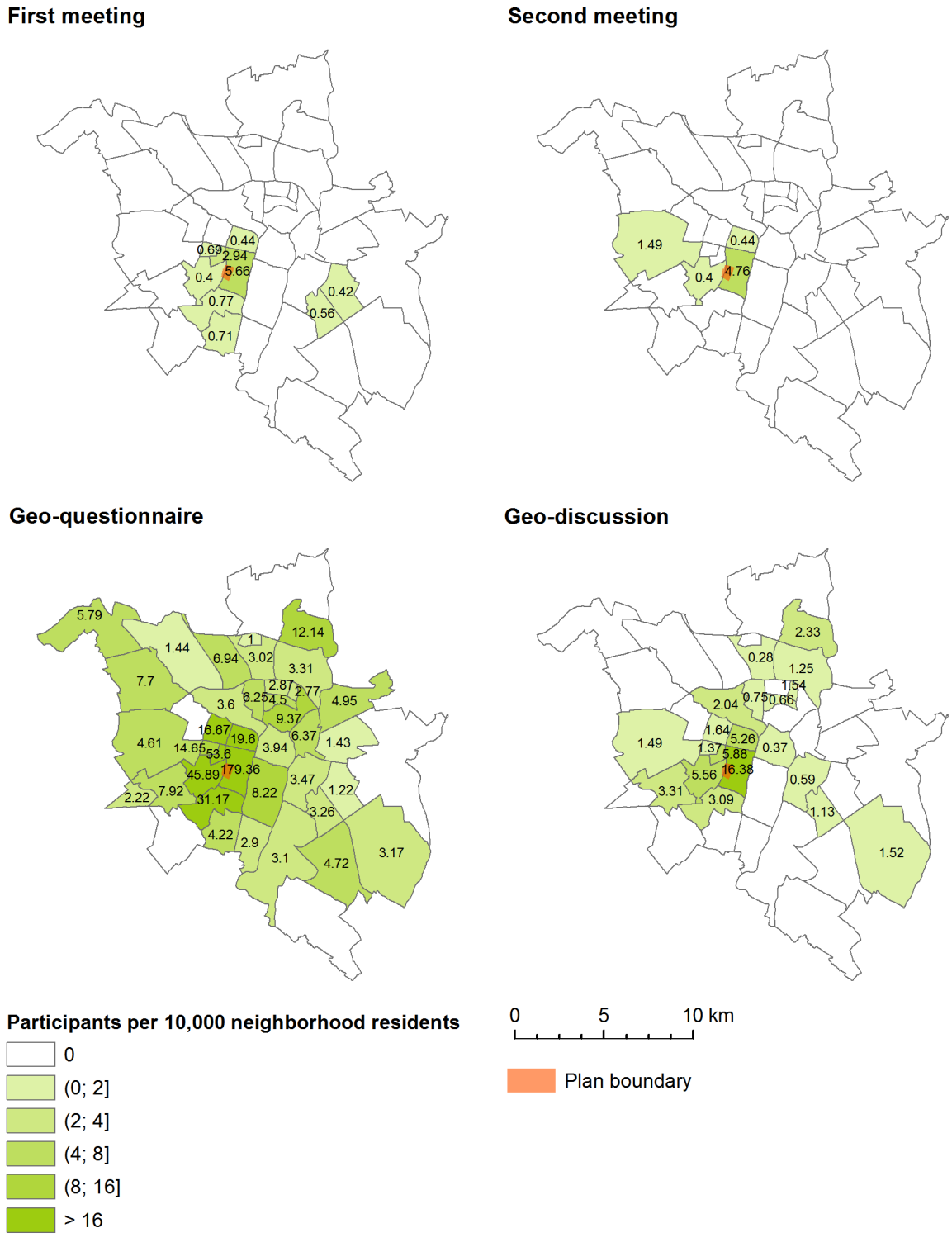


Figure 1. Spatial distribution of participation rates by city neighborhoods

The comparison of demographic distribution shows the overrepresentation of online participants in younger age groups (15-19 and 20-24), and underrepresentation in middle-and older-age groups (50 and older) (Figure 2). The overrepresentation among the younger participants is largely due to the Internet use patterns in Poland, and prevalence of learning about online participatory processes from social media, which was by far the most effective recruitment tool for both the geo-questionnaire and the geo-discussion (Jankowski et al., 2016).

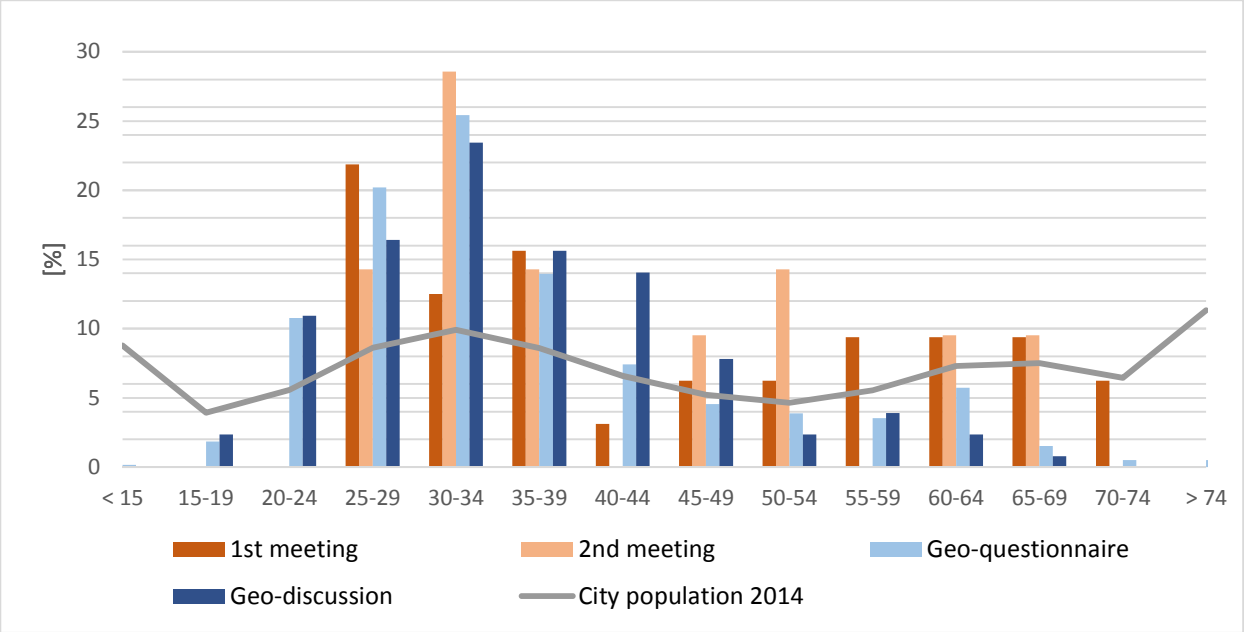


Figure 2. Demographic distribution by mode of participation compared to the city age structure

4. Conclusion

The results show that online PPGIS methods, in particular geo-questionnaire and geo-discussion do scale out and up in comparison with face-to-face meetings employed as a traditional mode of public participation in land use planning processes. Scalability of public participation methods, however, is confounded by the demographic representativeness of those who participate. The reported results provide an argument for combining online mode with face-to-face mode in order to improve the representation across the age groups. They also demonstrate no difference between the modes in respect to social-educational status of those who participate. Hence, the issue of scaling public participation across social groups remains a challenge. Involving underrepresented social groups may require using sampling approaches directed specifically at them in both online and offline settings.

References

Brown, G., and Kyttä, M. (2014) Key issues and research priorities for public participation GIS (PPGIS): A synthesis based on empirical research. *Applied Geography*, 46: 122-136.

Innes J.E., and Booher D.I. (2004) Reframing Public Participation: Strategies for the 21st Century. *Planning Theory & Practice* 5 (4): 419–36.

Jankowski, P., Czepkiewicz, M., Mlodkowski, M. Zwolinski, Z. 2016. Geo-questionnaire: A Method and Tool for Public Preference Elicitation in Land Use Planning. *Transactions in GIS*, in press DOI: 10.1111/tgis.12191

Nyerges, T., and Aguirre, R.W. (2011) Public participation in analytic-deliberative decision making: Evaluation a large-group online field experiment. *Annals of the Association of American Geographers*, 101(3): 561-586.

Schlossberg, M. A., and Shuford, E. (2005) Delineating "public" and "participation" in PPGIS. *Urban and Regional Information Systems Association (URISA) Journal*, 16(2): 15–26.

Sieber, R. (2006) Public participation geographic information systems: A literature review and framework. *Annals of the American Association of Geography*, 96(3): 491–507.

Multi-resolution, pattern-based segmentation of very large raster datasets

J. Jasiewicz[1,2], J. Niesterowicz[1], T. F. Stepinski[1]

[1]Space Infromatics Lab, University of Cincinnati, 401 Braunstein Hall, Cincinnati, OH 45221-0131, US
Email: {niestejk; stepintz}@mail.uc.edu

[2]Instititute of Geoecology and Geoinformation, Adam Mickiewicz University, Dziegielowa 27, Poznan, Poland
Email: jarekj@amu.edu.pl

Abstract

We present an algorithm which efficiently segments very large categorical rasters based on patterns of their categories. It operates on a grid of motifels – square blocks of raster cells representing a local pattern. Our algorithm is based on the seeded region growing principle but it uses a novel grid topology and seeds stack with individual thresholds. It has a single free parameter – the spatial scale of a pattern. Algorithm was proven to be robust on land cover data, topographic landforms data, and high resolution color-quantized RGB images. We present a multi-scaled segmentation of NLCD2011 as an example. Potential applications of the new algorithm include ecology, geomorphology, pedology, forestry, agriculture, and urban studies.

1. Introduction

Segmentation is the process of partitioning a raster dataset into multiple homogeneous segments. The goal of segmentation is to spatially generalize a raster so it provides more insight and is easier to analyze. The bulk of the existing work (Zhang *et al*. 2008) has focused on segmentation of images of relatively small scenes. However, segmentation of datasets that originated from remote sensing and cover large, continental or even global-scale areas, are also important, but existing segmentation algorithms are ineffective for such large datasets and the custom algorithms are lacking. Examples of such datasets are the National Land Cover Dataset (NLCD), which covers the conterminous US (CONUS) with the resolution of 30 m, or the SRTM-based DEM, which has a world-wide extent at 90 m resolution. Segmentation of NLCD could yield landscape types – precursors to ecoregions, and segmentation of world-wide DEM could delineate physiographic regions.

In this paper we describe a segmentation algorithm especially designed for very large rasters. Specificities of such datastes are as follows. (1) They are the mosaics of multiple datasets, thus it is better to segment a secondary product of uniform quality (for example, a land cover) rather than a montage of primary data of variable quality (for example, a montage of Landsat scenes). (2) They are large; for example, the NLCD consists of ~8 billion cells and has the size of ~16 GB. (3) The goal of the segmentation is to identify regions characterized by patterns which are homogeneous on the scale that is large in comparison to the resolution of the raster, since the need for pattern-based segmentation.

To deal with a large size of the input the proposed algorithm is based on the concept of Complex Object-Based Image Analysis (COBIA) (Vatsavai 2013, Stepinski *et al*. 2015). In COBIA the

raster is divided arbitrarily into a regular grid of local blocks of cells at minimal computational cost. These blocks of cells are the basic units of analysis; we will refer to them as *motifels* – the smallest processing elements containing local motifs (patterns) of raster variable. The size of the motifel sets the scale of patterns to be segmented. Using COBIA reduces the size of the problem by orders of magnitudes as elementary grid elements change from cells to motifels. To simplify a description of motifels we only consider categorical rasters. This is less restrictive than it may appear as one major application, land cover, is already a categorical raster, and the second, DEM, can be easily converted into categorical raster using geomorphons algorithm (Jasiewicz and Stepinski, 2013). With categorical input we represent patterns within motifels by a category co-occurrence histogram and we use the Jensen-Shannon divergence (JSD) (Lin, 1991) to measure a degree of dissimilarity between motifels.

The proposed segmentation algorithm works within the GeoPAT toolbox (Jasiewicz *et al*. 2015) – a collection of GRASS-GIS modules for pattern-based geoprocessing. It is based on the seeded region growing concept, but, in addition of being applied to motifels instead of pixels, it has two original innovations: (a) a novel *brick*-topology grid, and (b) a novel method for ordering seeds into a stack in order of increasing local inhomogeneity.

2. Pattern-based segmentation algorithm

The algorithm consists of five steps: 1) building brick-grid; 2) analyzing local homogeneity; 3) ordering seeds; 4) small areas removal (optional); 5) homogeneity enhancement (optional).

The result of segmentation depends on the topology of the grid. Four possible topologies are shown in Fig.1. The square, 4-connected grid is a standard in segmentation because the square, 8-connected grid may lead to segments that permeate each other. However, directions of the growth of segments are overly limited by the 4-connectivity. The hexagon grid has no directional bias but calls for hexagonal motifels formed out of square cells – a significant complication. We partition a raster into motifels using 6-connected square grid with brick topology which reduces directional bias.

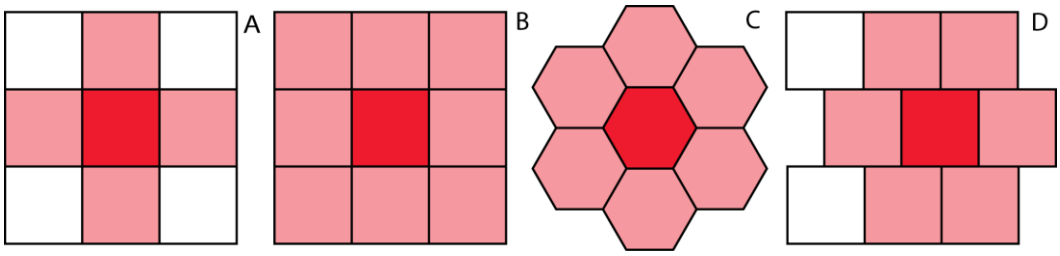


Figure 1: Different grid topologies: (A) square, 4-connected grid; (B) square, 8-connected grid; (C) hexagon, 6-connected grid, (D) brick-like square, 6-connected grid.

This grid is segmented using histograms as motifels' signatures and JSD as distance measure. The results of the segmentation depend on the selection of the seeds and the order they are grown into segments. Our idea is that priority should be given to growing segments from motifels located in the most homogeneous local neighborhoods. For a focus motifel we calculate a JSDs between it and motifels 18-motifels neighborhood. Using Fisher's Linear Discriminant we identify a subset of the neighborhood consisting of motifels most similar to the focus motifel.

The proxy for local inhomogeneity is an average JSDs between motifels in this subset and the focus motifel. All motifels in the grid are ordered into a stack of increasing values of their local inhomogeneities.

Segments are grown from the top of the stack. The size of emerging segments is automatically determined but the preference for larger or smaller segments could be set. Starting from a single motifel, all motifels forming a perimeter of a growing segment are checked for accruement. A motifel with the smallest value of an average linkage (AV - mean dissimilarity to all motifels in the segment) is added to the segment and the process is repeated until the value of the minimum AV is larger than the linkage threshold for this seed. The threshold is equal to the seed's inhomogeneity value plus the standard deviation of JSDs used to calculate the seed's inhomogeneity value. All motifels in the newly formed segment are removed from the stack and the next segment is grown from the remaining top ranked seed. After segment growing ends the segmentation may be optionally enhanced by removal of small segments and possible swapping of motifels at the segments boundaries to increase an overall homogeneity of segments.

3. Results

The algorithm was tested on the entire NLCD, the 30m CONUS DEM classified to topographic landforms, and a high resolution color-quantized RGB image (Niesterowicz *et al*., this conference). Here, we present the results of segmenting the NLCD2011. The algorithm has only a single parameter – the size of the motifel. We segmented the NLCD at the scales from 128 cells (~4km) to 4096 cells (~123km) each time increasing the scale by the factor of two. Fig.2 shows segmentations at four selected levels. Most segments have values of inhomogeneity below 0.1 (the range is between 0 and 1). Limited number of small segments has inhomogeneity values of ~0.3. The segmentation on the scale of 2048 cell resembles most closely the level-IV ecoregions division, but there are significant differences between the two partitioning because much more information than just the land cover pattern is used to delineate ecoregions. The computational time depends on the size of motifel. Segmentation on the scale of 128 cells took 187s on 8-core I7 computer, while segmentation on the scale of 1024 cells took less than 2s. Segmentations (shapefiles) of NLCD at scales of 128, 256, 512, 1024, 2048, and 4096 cells can be downloaded from http://sil.uc.edu/ (in the Data section).

4. Conclusions and future work

The algorithm is fast, stable, and convenient to use (the only parameter is the scale of pattern). Using different datasets other than those we have tested may require changing the motifel representation and dissimilarity function. Segmentation module is available as a part of GeoPAT software (http://sil.uc.edu). Applications include ecology, geomorphology, pedology, forestry and agriculture, and urban studies. Next step is to extend the algorithm to perform segmentation using multiple inputs thus becoming directly relevant to delineation of ecoregions.

Acknowlegements

This work was supported by the University of Cincinnati Space Exploration Institute, by Grant NNX15AJ47G from NASA, and by the National Science Center (NCN) grant DEC-2012/07/B/ST6/01206.

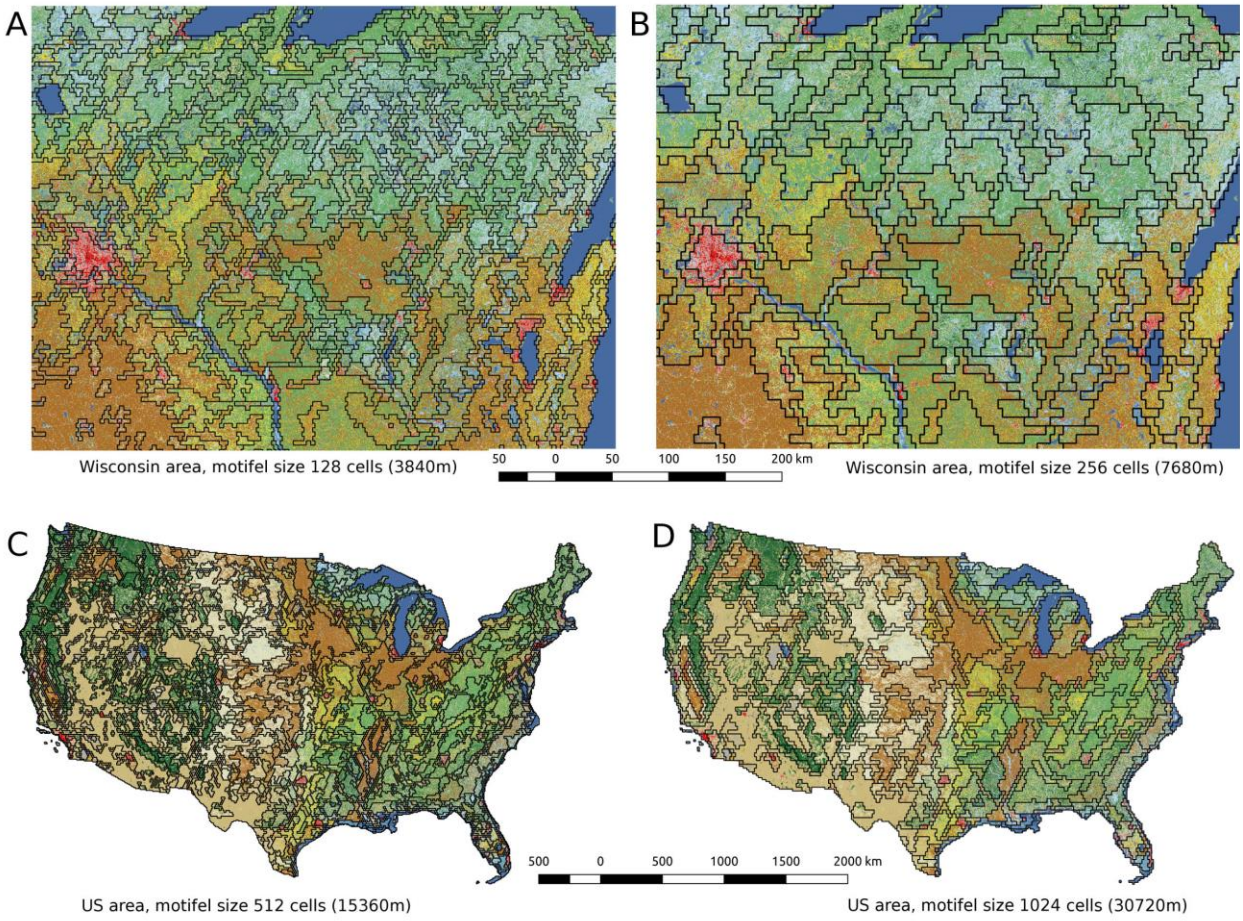


Figure 2: Selected results of segmentation of the NLCD2011. NLCD classes are shown using the standard colors

References

Jasiewicz, J., Netzel, P. & Stepinski, T. 2015, GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases. *Computers & Geosciences*, 80:62–73.

Jasiewicz, J. and Stepinski, T.F. 2013, Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182:147-156.

Lin, J. 1991, Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Stepinski, T.F., Niesterowicz, J. & Jasiewicz, J. 2015., Pattern-based Regionalization of Large Geospatial Datasets Using Complex Object-based Image Analysis. *Procedia Computer Science*, 51:2168–2177.

Vatsavai, R.R. 2013, Object based image classification: state of the art and computational challenges. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. pp. 73–80.

Zhang, H., Fritts, J.E. & Goldman, S.A. 2008, Image segmentation evaluation: A survey of unsupervised methods. *Computer Vision and Image Understanding*, 110(2):260–280.

Geospatial Internet of Things: Framework for fugitive Methane Gas Leaks Monitoring

L. J. Klein, R. Muralidhar, F. J. Marianno, J.B. Chang, S. Lu, H.F. Hamann
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

Abstract

We present a framework for wireless sensor network monitoring and detection of methane leaks from natural gas well pads. The wireless sensor network can measure methane concentrations across a well pad and combined with advanced analytics it can locate and determine the leak rate. Simulations of the inverse and forward modeling problems indicates that methane leaks can be localized within a 1 m distance from their original locations. The wireless sensor network and real time analytics can be extended to monitor multiple methane leaks and methane background levels.

1. Introduction

Methane has a much larger global warming potential compared to carbon dioxide. Methane gas is emitted by agricultural and waste management sources, however more than 30% of emitted methane gas is coming from energy exploration sites (natural gas and petroleum system and coal mining). With more than half a million natural gas well pad sites developed in USA, understanding the impact of methane gas on human health and long term climate impact became important.

In the past, standalone high precision sensors were used to measure methane gas leaks. These measurements can offer a very precise methane concentration assessment but the spatial coverage is limited. Current methane measurement and modeling techniques are lacking the capabilities to localize leaks on a well pad (Zavala-Araiza 2015; Lyon 2015; Foster-Wittig 2015). An alternative method to detect methane leaks is using satellite observations. While the satellite methods offer a large scale geospatial observation, the spatial resolution of the detection method is too coarse for single leak detection (Turner 2015; Veefkind 2012). There is certainly a need to combine the high accuracy local wireless sensor measurements with large scale satellite observations for (1) accounting all methane leaks over a regional area and (2) attribute methane leaks to emission sources.

Here we present a novel methane monitoring solution based on wireless sensor network. Methane sensitive sensors are distributed on a 10 m grid and are measuring in real time the methane concentration. In addition, the wind direction and speed is measured as well. We note that each well pad have construction on their perimeters and an associated infrastructure (storage tanks, well heads, etc). These structures will cause turbulence to wind flow pattern. The well pad layout can be extracted from high resolution satellite or drone imagery. The layout is used for three dimensional reconstruction of a gas well pad and generate a Computer Aided Design (CAD) models. The CAD model is a necessary input into CFD for dispersion modeling. The advantage of our proposed method is that each methane leak can be identified and localization on a well pad.

The concept for large area methane measurements, across multiple well pads, in presented in Figure 1. Each of the 4 well pads have its own Wireless Sensor Network that measure methane and wind for that specific locations. Inter well pads communication is enabled using a Wide Area Network (WAN) to transfer the sensor data and computational load between well pads. Each well pad may have one or more Raspberry Pi computers that act as a computational platform and data gathering device (edge devices). If no leak is present on a well pad, the sensor values can be sampled every hour, however in case of large methane leaks the sampling rate may have to be increased to a measurement each second. Since significant amount of data may be generated on each well pad, data processing needs to be carried out on the edge device and only aggregated and processes sensor values are sent to the cloud platform.

Locations of the sensors on the well pad as well as the site layout are geospatially located. Measurement on a single well pad may be affected by a leak on a different well pad. This scenarios can

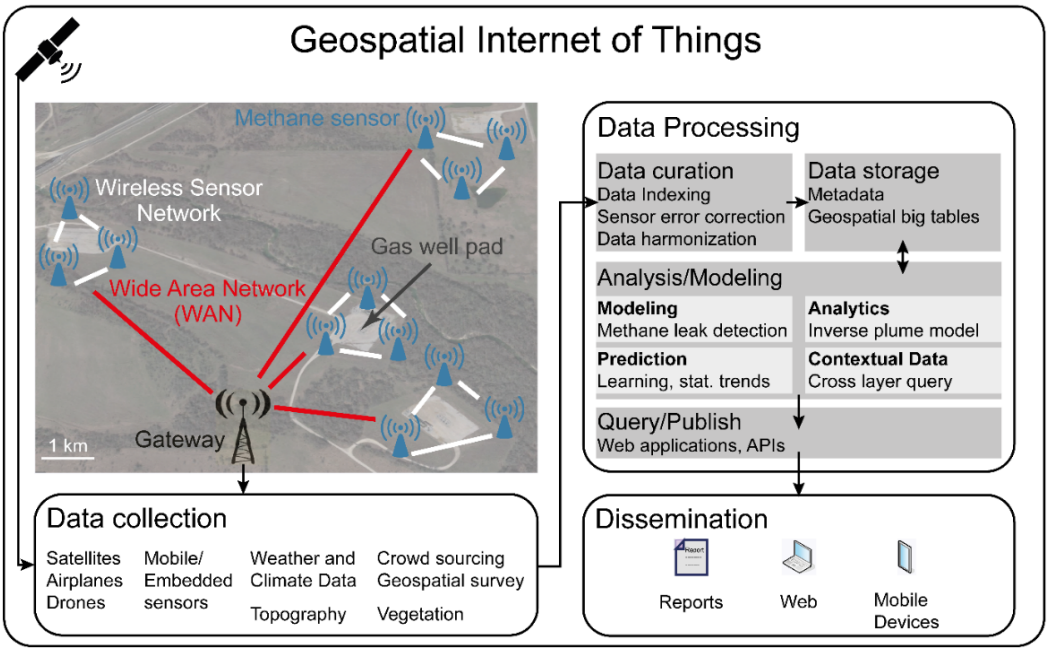


Figure 1 Geo-spatial Internet of Things architecture for methane leak sensing, modeling, and visualization.

be only addressed if the analytics models will utilize geospatial data; e.g. proximity of well pads are included into modeling along with topography and/or local vegetation. In our approach, wireless sensor network on a well pad is explored and the sensor point measurements are spatially linked with geospatial data (Klein, 2015). In the future, the IoT sensor measurement with GIS based analytics (Veefkind 2012) framework could be extended to multi pad methane leak monitoring.

2. Approach

A wireless mesh sensor network based on volatile organic compound sensors (VOC) is currently tested for 20 ppm methane plume detection sensitivity. Sensor analytics is developed to self-calibrate each sensors and compensate sensor reading for ambient temperature, humidity, and wind flow variations. The diffusion and transport of gas in the atmosphere strongly depends on local wind conditions. This fact plays a critical role in the detection and localization of gas sources. The modeling approaches described below assume that wind speed and direction are uniform within 10 m x 10 m. In order to get a better understanding of the statistical behavior of the wind, the distribution of wind speed and direction was calculated from two simultaneously acquired wind sensor readings (Figure 2 and 3). The angle mismatch between two sensors can be determined by calculating the maximum of the cross-correlation of two wind direction readings. The related difference in the angles is used to compensate for the mismatch.

The turbulence is homogeneous across the spatial length that separates the two wind sensors, although there are short term fluctuations. In case the sensors are separated by a building or infrastructure, these measurements for similarity in detected values would change significantly due to the local turbulence. Furthermore, the auto-correlation time of the wind speed and direction are between 2-3 s, meaning that the wind speed and direction stay constant within this time frame and will determine the required sampling rates for methane sensors. Hence, this quantity is critical when locating gas sources, since it implies that methane sensor sampling interval needs to be close to the autocorrelation time.

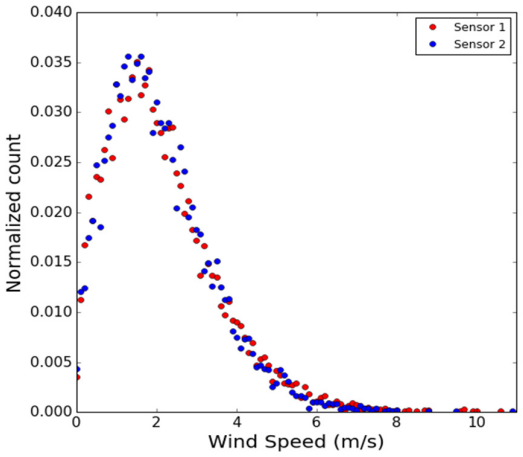


Figure 2. Distribution of wind speed for two sensors .

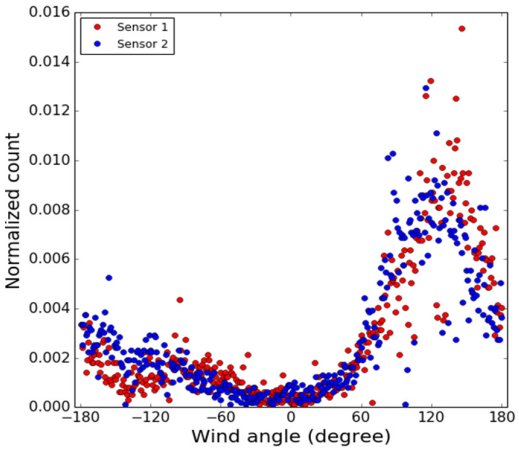


Figure 3. Distribution of wind directions compensated for 5 degree mismatch.

The analytics of source attribution from various types and length scales of geospatial data is dependent to some extent on the length scale of observation. The simplest and most studied model corresponds to the well-known Gaussian plume that establishes itself from a source when steady wind and atmospheric conditions remain stationary over a sufficient length of time and are spatially homogeneous. In the case of a planar terrain in a x-y plane, the plume dispersion characteristics for a source of strength q at $(0,0,h)$ are described by

$$c(x,y,z) = \frac{q}{2\pi u \sigma_y \sigma_z} \exp\left(-\frac{y^2}{2\sigma_y^2}\right)\left\{\exp\left(-\frac{(z-h)^2}{2\sigma_z^2}\right) + \exp\left(-\frac{(z+h)^2}{2\sigma_z^2}\right)\right\} \quad (1)$$

where, c is the pollutant concentration, q is the mean wind speed, and $\sigma_x$and σ_y are the dispersion parameters in the lateral and vertical directions and depend on atmospheric turbulence characteristics and distance downwind from the leak source, x. The x direction coincides with wind direction. For a spatially distributed network of sensors, and a non-local leak, the recorded concentrations are given by

$$s_i = A_{ij} q_j, \qquad i = 1,2 \ldots m \quad (2)$$

where, A is the discretized Green's function and $\{q_j\}$ with $j = 1, \ldots, n$ are source strengths at spatial grid point j. The discretized Green's function A_{ij} represents the concentration at sensor i arising from a unit source at spatial grid point j and depends on wind speed as well as wind direction. This is the so called *forward problem*. The *inverse problem* consists of determining the source distribution $\{q_j\}$ with $j = 1, \ldots, n$ from a knowledge of the sensor readings measured in a gas leak scenario. Since the number of sensors, m is typically less than the number of spatial grid points, n multiple wind directions are needed to make the inverse problem well –posed. To illustrate the source attribution, a network of 25 sensors were placed in a rectangular 10 m x 10 m grid. The forward problem was solved to predict the concentrations at the sensors for 6 different wind conditions. To mimic experimental Gaussian noise a signal-to-noise (SNR) of 10 was added to the sensor readings. The resulting sensor data were used to locate the leak by solving the inverse problem using least-squares with regularization.
Thus the source vector, $\boldsymbol{q}$, is determined by a minimization of

$$||\boldsymbol{Aq} - \boldsymbol{s}||^2 + \lambda||\boldsymbol{q}||^2 \quad (3)$$

where, λ, is a small regularization parameter. Figure 4 shows the result of inversion for a scenario where source is positioned 0.5 m above the ground. The distribution of normalized counts (sum of all counts is equal to 1) for the inversion is successful, most of the time, with a mean source location error of about 1.04 m (Figure 5).

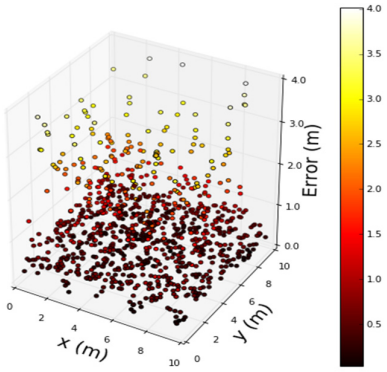


Figure 4. Randomly generated 1000 leak coordinates (x, y) and the error in source location from the inverse problem.

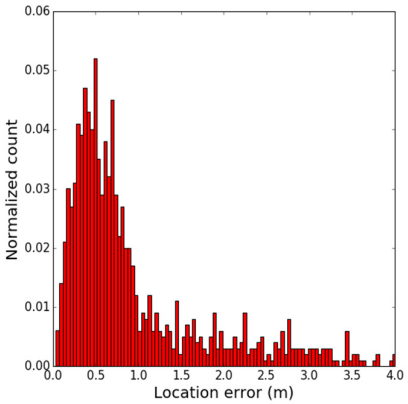


Figure 5. Histogram of source location errors for 1000 simulations.

While a simple scenario has been used for illustration, the method can be generalized to realistic situations such as gas well pads by using a Green's function appropriate for the situation. Such Green's functions can be obtained for instance by computational fluid dynamics (Crank 1975) and integrated into edge devices analytics that carry out methane leak calculation on well pad sites.

3. Conclusion

A scalable solutions to monitor methane leaks is proposed based on wireless sensor network and advanced analytics. We demonstrated that inverse and forward modeling can locate methane leaks with an accuracy of 1 m and quantify emission rates on individual well pad sites. Distributed computing on edge devices and cloud platforms require integration of GIS with IoT sensor data to pinpoint methane leaks and to distinguish small leaks from background methane fluctuations. The solution can be easily adaptable to other industries like agriculture, livestock or waste management where methane emission has a significant impact.

Acknowledgment

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represented that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

References

Turner A J *et al.*, 2015, Estimating global and North American methane emissions with high spatial resolution using GOSAT satellite data. *Atmospheric Chemistry and Physics*, 15:7049-7069.

Veefkind, J P *et al.*, 2012, TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment* 120:70-83.

Zavala-Araiza D *et al.*, 2015, Reconciling divergent estimates of oil and gas methane emissions. *Proceedings of the National Academy of Sciences*, 112.51:15597-15602.

Lyon D R *et al.*, 2015, Constructing a spatially resolved methane emission inventory for the Barnett Shale region. *Environmental science & technology*, 49.13:8147-8157.

Foster-Wittig T A *et al.*, 2015, Estimation of point source fugitive emission rates from a single sensor time series: A conditionally-sampled Gaussian plume reconstruction. *Atmospheric Environment*, 115:101-109.

Klein L J *et al.*, 2015, PAIRS: A scalable geo-spatial data analytics platform, *IEEE International Conference on Big Data*, Santa Clara, USA, 1290-1298.

Crank J., The Mathematics of Diffusion, Oxford University Press, Second Edition, 1975.

Understanding spatial patterns of biodiversity: How sensitive is phylogenetic endemism to the randomisation model?

S.W. Laffan [1], A.H. Thornhill[2], J.T. Miller[3], N. Knerr[4], C.E. Gonzales-Orozco[5], B.D. Mishler[2]

[1]Centre for Ecosystem Science, School of Biological, Earth and Environmental Science, UNSW, Sydney, Australia, 2052
Email: shawn.laffan@unsw.edu.au
[2]University and Jepson Herbaria, and Dept. of Integrative Biology, University of California, Berkeley, CA 94720-2465, USA.
Email: {andrew.thornhill, bmishler}@berkeley.edu
[3]Division of Environmental Biology, National Science Foundation, Arlington, Virginia, USA
Email: joe@acaciamulga.net
[4]National Research Collections Australia, CSIRO National Facilities and Collections, GPO Box 1600, Canberra ACT 2601, Australia
Email: Nunzio.Knerr@csiro.au
[5]Corporación Colombiana de Investigación Agropecuaria, Corpoica, km 17 Vía Puerto López, Meta, Colombia
Email: cegonzalez@corpoica.org.co

Abstract

Mapping spatial patterns of phylogenetic diversity helps identify regions of unique evolutionary history warranting conservation. Randomisations form an integral component of this process. Here we test the sensitivity of a method used to identify unusual concentrations of old and new evolutionary history to the underlying randomisation. The results indicate low sensitivity to models of complete spatial randomness and spatial structure (proximal allocation and random walks).

1. Introduction

Knowledge of the spatial distribution of biodiversity is essential for the allocation of scarce conservation resources and for understanding the evolutionary histories of a region's biota. Biodiversity is many-faceted, and can be measured using components such as species, phylogenetic, and trait diversity (Laffan 2014).

Biodiversity indices are typically aggregate measures of the taxon assemblage found in a location, with geographic surfaces of indices commonly generated. The most commonly used index is species richness, calculated as the number of unique species in a sample. However, closely related species represent less unique diversity than do distantly related species (see Fig 1), e.g., a sample comprising a human, a gorilla, and an orangutan has less unique diversity than one comprising a snake, a cow, and a squid. If one is interested in the conservation and analysis of biodiversity at an evolutionary level then one needs to use phylodiversity indices (Laity et al. 2015).

Phylogenetic Diversity (PD; Faith 1992) is the simplest phylodiversity measure and is calculated as the sum of a tree's branch lengths in a sample (Figure 1). Phylogenetic Endemism (PE; Rosauer et al. 2009) is calculated in the same way as PD, but the branches are weighted by the fraction of their geographic ranges found in a location, such that wide-ranged branches contribute less than narrow-ranged branches of the same length. PE is used to identify regions containing lineages that are found in few other places.

A more recent development of PE is the CANAPE method (Categorical Analysis of Neo- and Palaeo-Endemism; Mishler et al. 2014). CANAPE uses PE with a randomisation test to identify regions of geographically restricted long or short branches. Regions of palaeo-endemism can be considered as museums of evolutionary history currently found in few other places, while regions of neo-endemism can be considered as cradles of new diversity. CANAPE classifies the remaining cells into three other classes, two that contain some

mixture of palaeo- and neo-endemism at differing levels (mixed and super), and those that are not significant given the randomisation test.

The randomisation test is integral to CANAPE as it is used to allocate cells to the different classes. The randomisation test used in studies to date follows a model of complete spatial randomness (CSR). In each random realisation, each species is allocated to cells randomly across the data set, with the dual constraints that each species is found in exactly the same number of cells as in the observed data set, and that each cell has exactly the same number of species as it does in the observed data set (thus range size and richness are held constant). A swapping process is used to ensure all species occurrences are allocated to satisfy the range constraint.

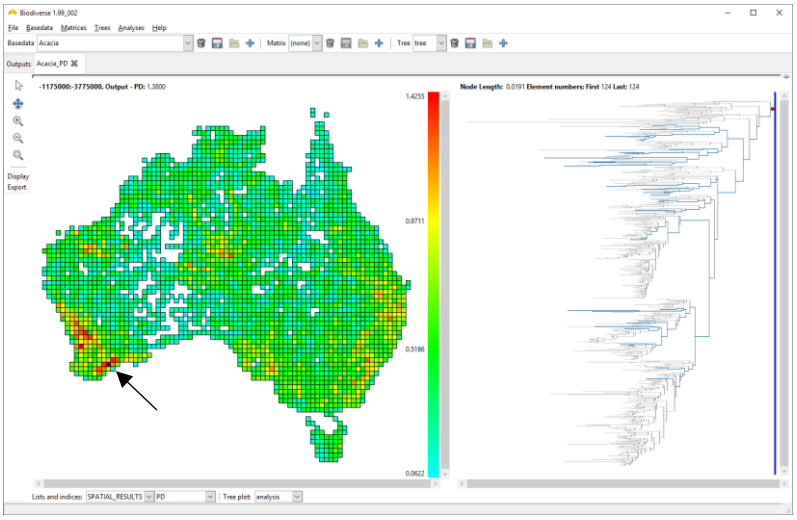


Figure 1. Phylogenetic diversity (PD) is measured as the sum of branch lengths on a phylogenetic tree that are found in a location. The example shows PD using a phylogenetic tree containing 506 *Acacia* species across Australia. Branches highlighted in blue occur in a cell in south-west Western Australia (arrow) and the sum of their lengths is the PD score for that cell.

The limitations of the CSR model are well documented (see for example O'Sullivan and Unwin 2010), with spatially structured randomisations potentially offering a more rigorous test of the results (O'Sullivan and Perry 2013). Such effects have previously been explored for species level endemism indices (Laffan and Crisp 2003), but not for more complex phylogenetic indices such as CANAPE. An understanding of the effect of more spatially structured randomisations on the CANAPE analysis is therefore needed.

2. Spatial randomisations

Two additional models have been implemented in the Biodiverse software (Laffan et al. 2010), proximal allocation (PA) and a random walk (RW) (Figure 2). Both models use the swapping approach from the CSR model to ensure all occurrences are allocated.

In the PR model a species and seed cell are selected, and the species occurrence is allocated to that cell. Subsequent occurrences of that species are then allocated in order of increasing distance from the seed cell, with random selection in the event of ties. Cells are skipped if they have already reached their richness target. A spatial window centred on the seed cell controls how far species will be allocated from the seed cell. Once all allocatable cells in the window have been used, a new seed location is chosen and the process continues until all occurrences are allocated or there remain no cells to assign to. This approach is similar to the circular model used in Laffan and Crisp (2003), but with greater flexibility as

Biodiverse supports arbitrarily complex spatial windows. (The allocation order can also be random instead of proximal, but that is not used here).

The RW model is a long established approach (O'Sullivan and Perry 2013) and is simply a variation on the PA approach that uses a different allocation method. From the seed cell, the method selects a neighbouring cell to which it allocates the next species occurrence. It then allocates to one of that cell's neighbours, and the process repeats until all occurrences have been allocated or no more cells can be assigned to. If a cell has no assignable neighbours then the system backtracks to the most recently allocated cell with such neighbours, or it restarts at a new seed location. The RW model has the advantage that the random distributions will remain within a region if there are gaps that the spatial window does not span, for example islands.

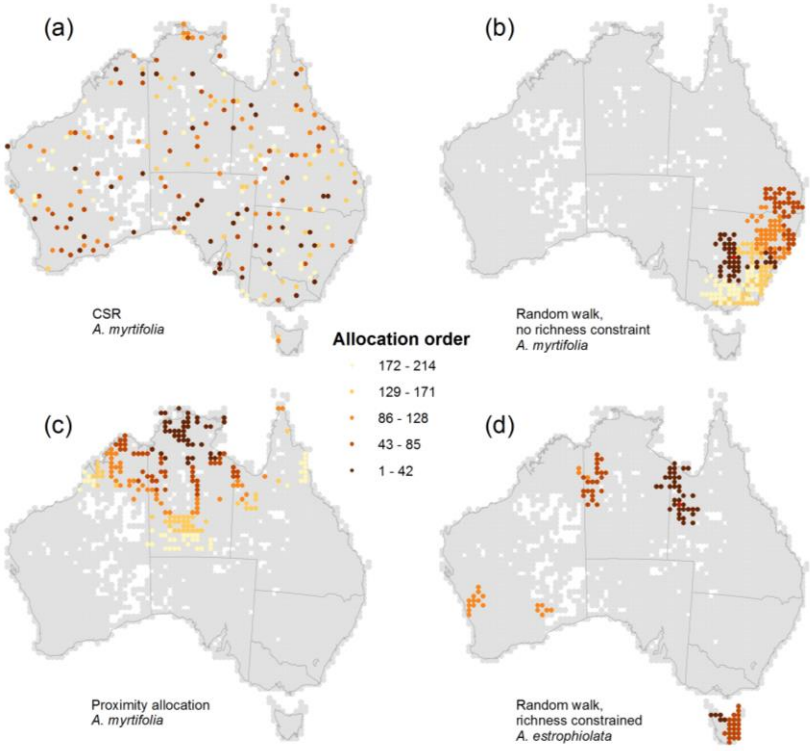


Figure 2. Example randomised distributions. (a) CSR, (b) RW without richness constraints, (c) PA, and (d) RW with richness constraint. Colours show allocation order.

3. Analyses and discussion

The CANAPE index was calculated for a data set of 506 *Acacia* species aggregated to 3037 cells at a 50 km resolution (Mishler et al. 2014). Four randomisations were used: CSR, RW with 100 km radius windows, PA with 100 km radius windows, and PA with no window constraint.

The results indicate little difference in the overall patterns (Figure 3). There is a small increase in the neo-endemic locations using the spatial randomisations, and the unconstrained PA has more non-significant cells, but the general patterns remain the same.

It is likely that the richness constraints have a large influence on initial allocations, with the swapping process further disrupting the spatial structure of the initial distributions. This is the likely reason the internal branch ranges remain larger than the observed data, and are not substantially different among the models (data not shown). This is also broadly consistent with the results of Laffan and Crisp (2003) for a single cell analysis.

Further testing will assess the effect of spatial scale on the results, considering the degree to which branches in a cell are restricted to regions surrounding them. More complex RW

models could include random back-tracking instead of sequential to generate shorter paths, and constraining the overall size and shape of a distribution to create more compact walks.

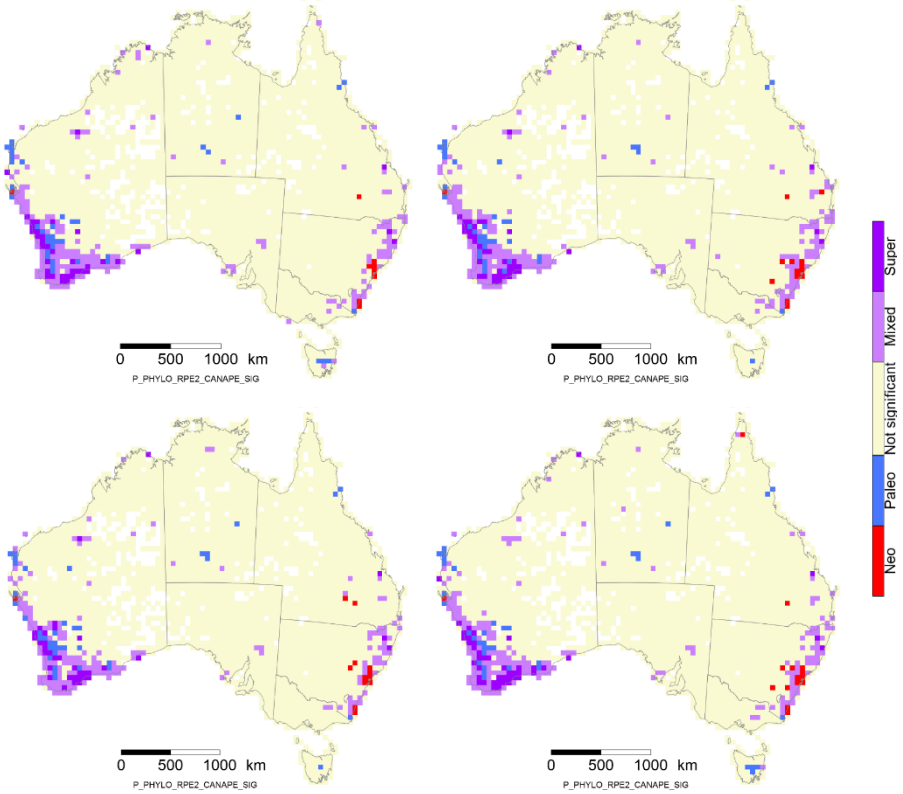


Figure 3. CSR model (upper left), random walk model (upper right), proximal allocation (100km radius; lower left), proximal allocation to any cells (lower right).

Acknowledgements

This manuscript includes work done by JTM while serving at the National Science Foundation. The views expressed in this article do not necessarily reflect those of the National Science Foundation or the United States Government. NSF grant DEB- 1354552 provided partial support for the UC Berkeley portion of this project.

References

Faith, D.P. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation,* 61, 1-10.

Laffan, S.W., 2014. GeoComputation: Applications in Biology. *In:* Abrahart, R. J. and See, L. eds. *GeoComputation, Second Edition.* 2 ed. London: CRC Press, 125-142.

Laffan, S.W. and Crisp, M.D. 2003. Assessing endemism at multiple spatial scales, with an example from the Australian vascular flora. *Journal of Biogeography,* 30, 511-520.

Laffan, S.W., Lubarsky, E. and Rosauer, D.F. 2010. Biodiverse, a tool for the spatial analysis of biological and related diversity. *Ecography,* 33, 643-647.

Laity, T.*, et al.* 2015. Phylodiversity to inform conservation policy: An Australian example. *Science of the Total Environment,* 534, 131-143.

Mishler, B.D.*, et al.* 2014. Phylogenetic measures of biodiversity and neo- and paleo-endemism in Australian Acacia. *Nature Communications,* 5, 4473.

O'Sullivan, D. and Perry, G.L.W., 2013. *Spatial Simulation: Exploring Pattern and Process.* Chichester, UK: John Wiley & Sons, Ltd.

O'Sullivan, D. and Unwin, D.J., 2010. Geographic Information Analysis, 2nd edition. *Geographic Information Analysis.* 2nd ed. Hoboken, USA: John Wiley & Sons, Inc., 406.

Rosauer, D.F.*, et al.* 2009. Phylogenetic endemism: a new approach to identifying geographical concentrations of evolutionary history. *Molecular Ecology,* 18, 4061-4072.

Retrieving Indigenous Knowledge to a Digital Map: the Case of the Traditional Farming System in a *Hñahñu* (Otomí) Community, Mexico

José María León Villalobos [a], Enrique Ojeda Trejo [a], Michael K. McCall [b]
Verónica Vázquez García [c].

[a] Posgrado de Edafología, Colegio de Postgraduados, Campus Montecillo, Carretera México – Texcoco km. 36.5, CP 56207, Texcoco Estado de Mexico, Mexico, Email: {jomalevi@yahoo.com.mx; enriqueot@colpos.mx; verovazgar@hotmail.com }

[b] CIGA, Centro de Investigaciones en Geografía Ambiental, UNAM, Universidad Autónoma de México, Antigua Carretera a Pátzcuaro No. 8701. Col. Ex –Hacienda de San José de la Huerta C.P. 58190 Morelia, Mexico Email: mccall@ciga.unam.mx

[c] Posgrado en Desarrollo Rural, Colegio de Postgraduados, Campus Montecillo, Carretera México – Texcoco km. 36.5, CP 56207, Texcoco Estado de Mexico, Mexico, Email: verovazgar@hotmail.com.

Abstract

Indigenous classification systems represent cognitive experiences of human groups in the geographical space. Formalization efforts of indigenous knowledge impose their own concepts, and therefore, it is often decontextualized. This research aims to formalize the farm land management system of an *Hñahñu* (*otomí*) community into a map using their own geographical concepts. A semantic analysis with Participatory Geographical Information System and Google Earth visualization is proposed as a method. Results show that farm land management system developed by *Hñahñu* include a set of geographical categories and subcategories. It was found that the *Hñahñu* classify them using the plot location in the landscape and the technique for providing water to grow crops as attributes. Although this recognition allowed the drawing of boundaries, the *Hñahñu* conceptualization of space challenged the conventional map, this led into a Google earth map. Google Earth showed the potential for improving indigenous knowledge representations within the community.

1. Introduction

Language is a good starting point to understand the way indigenous people perceive, conceptualize and understand their geographical space (Giannakopoulou *et al*. 2013). Although numerous researches on indigenous knowledge formalization have been conducted most of them impose their own scientific and technical concepts. Therefore, indigenous knowledge is often decontextualized and incompletely represented (Chapin *et al*. 2005). Using the indigenous geographic concepts and terms to formalize knowledge is a reasonable way to approach the indigenous view of geographical space. This research took place in the *Hñahñu* community of Huitexcalco in the dry Mezquital Valley, Mexico. The low precipitation and shallow soils led to the *Hñahñus* to trap water and soil in terraces which are typically built in gullies or up hillsides for cropping. This research aims to formalize the *Hñahñu* farm land management system into a spatial representation using their own concepts and terms. The formalization process encompassed both the elicitation of the *Hñahñu* farm terms and concepts in workshops and the production of a conventional map into a GIS and then into Google Earth.

2. Methods

The research proposal was presented to the community, and then they selected the people who would participate in the whole process. This selection included five elders ranging in ages from 70 – 78 years and eight farm experts between 18 to 50 years of age. Six of them were men and the others were females. The *Hñahñu* farmland management terms were elicited in a workshop using the words provided by Granados et al. (2004). Participants wrote the terms in colorful papers and discussed their meanings verbally. Using the listed words we undertook three field trips to different farm plots in the community aiming to clarify the meanings of some terms. Participants related with more precision each term with their most distinguished attributes such as the plot location on the landscape and the technique for providing water to grow crops. Once there was mutual agreement on the meanings, the participants offered a generic translation of terms in the Spanish language in the fashion of the semantic analysis performed by Wellen and Sieber (2013).
The formalization process encompassed both the visual spatial allocation of the categories and the drawing of boundaries on a photomap with satellite image, scale 1:25 000, as a PGIS tool (McCall and Dunn 2012). Participants were involved in a lengthy process of comparing and contrasting the most distinguished characteristics of number of farm plots in order to define the category to which they belong and its spatial extent. Finally, boundaries were digitized into a GIS to produce a conventional map. However, it received negative feedback and hence the map was transformed into Google Earth visualization.

3. Results

The research revealed four farm land categories: *Ngat'i, Ngats'i, Ndants'i* and *Mothe*. All of them, except *Mothe*, are positional terms. For instance, *Ngat'i* refers to a farmland plot that is 'in the down slope and near a deep gully'. Also, for the *Mothe* category, four specific subcategories were distinguished: M*øinñe, Ngadñe, Ñøt'athee* and *Mothee*. Categories and subcategories are complementary terms both make sense to farmland management plots. For instance, in *Mothe – møinñe, Mothe* is a functional term that means 'where water is retained in gullies or caught by rain', whereas *Møinñe* may be translated as a farmland plot 'where water is retained in the gully' this is again a positional term. Table 1 summarizes the semantic analysis by each category and subcategory.

Table 1. Semantic meanings of the *hñahñu* farmland plot categories and subcategories.

Category	Sub-category	Words	Translation meaning
Mothe		*Mo*= retains; *the*= water	Where water is retained in gullies.
	Møinñe	*Møin*= belly; *ñe* = gully.	Where water is retained in the gully (inside of the gully's belly)
	Ngadñe	*Ngad* = nex to; *ñe*= gully.	Where water is retained next to the gully.
	Ñøt'athee	*Ñøt'a* = flows; *thee*= water.	Where water flows from the gully.
	Mothee	*Mo*= retains; *the*=water	Where water is retained in gullies.
Ngat'i		*Ngat'i* = next to the gully.	In the down slope and near a deep gully
Ngats'i		*Ngats'i* = on the slope.	In the middle of the slope
Ndants'i		*Ndants'i* = on the top of the hill.	On the top of a hill

The semantic analysis provided useful insights to understand the way *Hñahñu* conceptualize their farmland plots. The farm plot location in the landscape, the water supplies strategies and the slope remains as the major attributes of the *hñahñu* farmland system. The central role given by *hñahñu* to those attributes can be explained by the rough environmental conditions in the community. Similarly, Barrera-Bassols *et al*. (2006) has pointed out that other indigenous groups have employed such attributes for soils and land uses classification purposes in central Mexico because they reflect their immediate potentials and constraints.

The conventional representation on a map of the farmland management categories and subcategories was at odds with the *hñahñu* view of geographical space. Participants experienced confusion in placing themselves on a map oriented north to south, and also faced difficulties to identify their own farm plots in a bi-dimensional projection. The *Hñahñu*, as other indigenous groups, have a direct relationship with the land. They use hilltops and gullies as the central spatial references (Oliveira 2005). In order to overcome this, the map was transformed into a Google Earth visualization provided with a digital terrain model. Such representation enabled participants to explore their farmland plots by categories and subcategories in a fairly realistic and detailed three dimensional view shown in Figure 1.

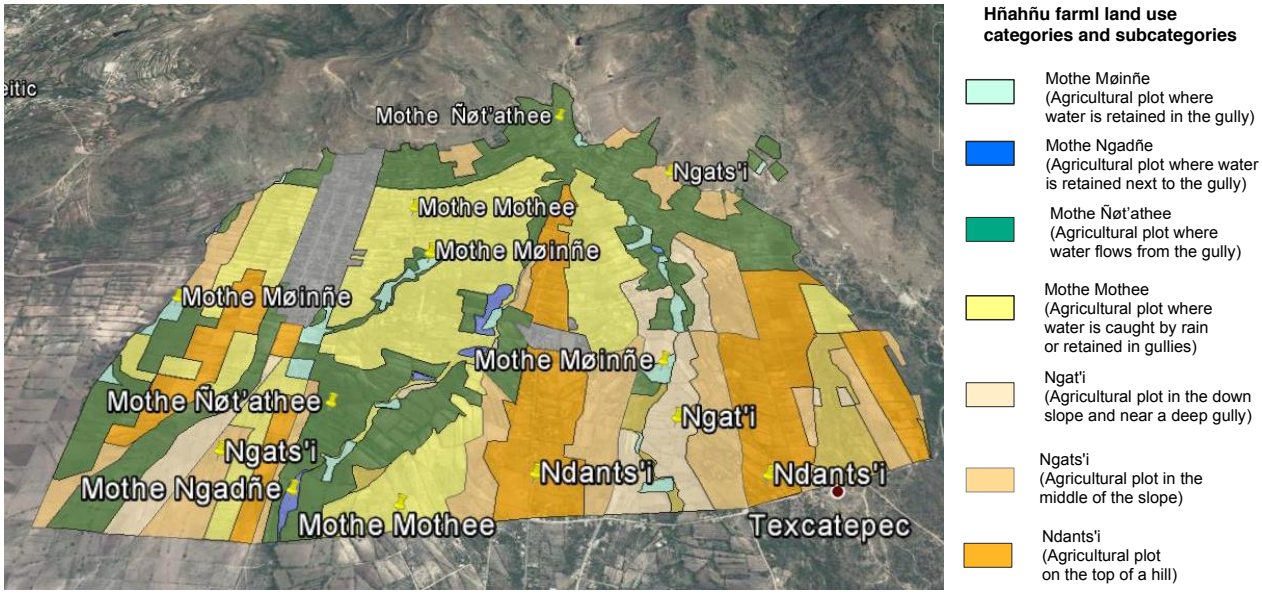


Figure 1. *Hñahñu* farm plot categories and subcategories represented in Google Earth.

4. Conclusions

The research revealed that *Hñahñu* classify their farms lands in categories and subcategories by combining two central attributes the farm plot location in the landscape and water supply strategies. It is show that *Hñahñu* indigenous knowledge representation in two dimensional maps was inappropriate and Google Earth proved to be more effective in communicating the *Hñahñu* cultural cognition around their farm lands. This methodology can be employed in similar researches, aiming to document and visualize traditional spatial indigenous knowledge in central Mexico.

5. References

Barrera-Bassols N, Zinck J. A. and Van-Ranst E, 2006, Symbolism, knowledge and management of soil and land resources in indigenous communities: Ethnopedology at global, regional and local scales. *Catena,* 65(1)**:**118-137.

Chapin M, Lamb Z and Threlkeld B, 2005, Mapping indigenous lands. *Annual Review of Anthropology*, 34(1):619-638.

Giannakopoulou L, Kavouras M, Kokla M and Mark D, 2013, From compasses and maps to mountains and territories: experimental results on geographic cognitive categorization. In: Cartwright W, Gartner G, Meng L and Peterson M P (eds), *Cognitive and Linguistic Aspects of Geographic Space*, Heidelberg, Springer, New York, USA, 63-81.

Granados SD, López RG, Hernández HJ, 2000, Agricultura nhanñhu-otomí del Valle del Mezquital, Hidalgo. *TERRA Latinoamericana* 22 (1): 117-126.

McCall M K and Dunn C, 2012, Geo-information tools for participatory spatial planning: fulfilling the criteria for a good governance?. *Geoforum*, 43(1):81-94.

Oliveira K R K, 2005, Two worldviews war: the struggles of a Hawaiian connection to the land. In: G. Cant, Goodall A. and Inns J (eds), *Discourses and silences: indigenous peoples, risks and resistance,* Department of Geography, University of Canterbury, Christchurch, NZ, 115-121.

Wellen C C and Sieber R, 2013, Toward an inclusive semantic interoperability: the case of Cree hydrographic features. *International Journal of Geographical Information Science,* 27(1):168-191.

A Field-Based Time Geography for Wildlife Movement Analysis

J. A. Long[1]
[1]School of Geography and Geosciences, University of St Andrews,
Irvine Building, North Street, St Andrews, Fife, UK KY169AL
Email: jed.long@st-andrews.ac.uk

Abstract

Field-based time geography is proposed as a new, specialized model for estimating wildlife utilization distributions (UDs) and home ranges. Field-based time geography represents the combining of classical time geography with least-cost path analysis. Here the derivation of field-based time geography is emphasized, paying particular attention to how it can be implemented in wildlife analysis. An example showing caribou movement in northern British Columbia is used to demonstrate field-based time geography and compare it with the Brownian bridge model, a popular method for delineating wildlife UDs. The results show how field-time geography is able to better represent the structure and barriers of the landscape, and provide alternative insights into wildlife space use. The entire process is implemented in R, making it attractive to movement ecologists.

1. Introduction

One of the fundamental pieces of spatial information used in the study of wildlife movement is the home range. A home range is broadly defined as the area an animal uses in its normal day-to-day activities (Burt, 1943). From a GIScience perspective, a home range is a polygon, and thus represents this area discretely. Recognizing that animals typically do not use their home ranges evenly, alternatively a utilization distribution (UD) can be estimated, which represents the home range as a two dimensional probability density surface, where the values indicate the probability of observing the animal at different locations within (and outwith) the home range (Worton, 1989). Typically, a UD is represented as a raster grid after selecting an appropriate spatial resolution. After creating a UD it is common to estimate the home range simply as a % volume contour of the UD (most commonly the 95% volume contour).

Many methods have been proposed for estimating UDs and home ranges and there is little consensus on which is best. Historically, kernel density estimation (Worton, 1989) has been the preferred approach, however in recent years Brownian bridges (Horne et al., 2007) have become increasingly popular. Time geographic methods have also been proposed as powerful alternatives for delineating home ranges (Long and Nelson, 2012) and UDs (Downs et al., 2011). However, one limitation that is present in all currently available home range and UD methods is the assumption that wildlife move within a homogeneous arena. That is, the underlying environment has no influence on the model used to calculate the home range and UD. By failing to consider how the underlying landscape characteristics (e.g., topography, structure, land cover) contribute to the formation of home ranges and UDs, current approaches fail to adequately capture how a heterogeneous environment contributes to observed movement patterns. In many cases, the natural environment constrains the area potentially visited by an animal, and thus home ranges and UDs are incorrectly estimated.

To address this limitation, this paper proposes field-based time geography as a new tool for estimating wildlife UDs. Field-based time geography was originally proposed by Miller and Bridwell (2009) in the study of transportation, and thus focused on applications on a spatial network. Here, the conceptual framework outlined by Miller and Bridwell (2009) is extended to the study of wildlife movement in a two-dimensional field. In this context, field based time

geography represents the marriage of time geographic theory (Hägerstrand, 1970) with popular GIS methods for studying least-cost paths in spatial fields. This paper outlines the formulation of field-based time geography for generating wildlife UDs. Using an example where caribou (*Rangifer tarandus*) were tracked via GPS in northern British Columbia, Canada, field-based time geography is compared to the Brownian bridge estimator, demonstrating how the approach can be used in real applications.

2. Methods

2.1 Field-based time geography

Field-based time geography extends the classical definition of time geography (Hägerstrand, 1970) in an effort to estimate the unequal movement probabilities within the bounding structure of space-time prisms (Miller and Bridwell, 2009). Movement possibilities are limited not only by the upper bounds on animal movement (e.g., maximum travelling speed) but also the characteristics of the landscape through which the animal moves. Thus, field-based time geography requires a cost (resistance) surface which defines the speed at which an animal can navigate the landscape, which is practically implemented as a raster grid.

For a given pixel i with 8 cardinal neighbours (other definitions are possible) let c_{ij} be travel cost (in units of time) to go from pixel i to neighbour j. In practice c_{ij} will be related to the movement ability of the animal and, importantly, to the characteristics of the landscape, such as topography, land cover, and the presence of barriers. Then let $\boldsymbol{N}$ represent the network graph which consists of all transitions c_{ij} for all pixels in the study area. Using a network optimization algorithm (e.g., Dijkstras algorithm), based on $\boldsymbol{N}$, we can compute the minimum cost of movement between any two locations on the graph $\boldsymbol{N}$ in units of time.

The construction of field-based time geography follows classic time geography by considering the intersection of space-time cones. For an intermediate time point t between two fixes a and b, where $t_a < t < t_b$, first calculate C_{ai}, which is the cost (in units of time) from the location of fix a to pixel i based on the network $\boldsymbol{N}$ (similarly compute C_{ib}). If location (pixel) i is accessible at time t (i.e., $C_{ai} \leq t\text{-}t_a$; and $C_{ib} \leq t_b\text{-}t$) then location i is within the potential path space at time t (PPS_t). Next we compute the minimum time budget required to reach each location within PPS_t:

$$TB_i = C_{ai} + C_{ib} \quad \forall\, i \in PPS_t \tag{1}$$

Here we estimate the probability an animal visited a given pixel at time t as:

$$\hat{P}_{it} \propto \frac{1}{TB_i} \tag{2}$$

$$P_{it} = \frac{\hat{P}_{it}}{\sum_{\forall i} \hat{P}_{it}} \tag{3}$$

Equation (2) states that the probability of an animal visiting a pixel i at time t is proportional to the inverse of the time budget required to get to location i. The formulation of $\hat{P}_{it}$ in equation (2) has the desired effect of associating visit probabilities with travel cost, and simultaneously considers this cost relative to the minimum of all possible routes in the potential path space, but other definitions of $\hat{P}_{it}$ are possible, for example inverse aquared. With equation (3) the $\hat{P}_{it}$ are standardized so that $\sum P_{it} = 1$ for any time t. Standardizing so that $\sum P_{it} = 1$ is necessary to account for variations in the size of the PPS_t.

In order to compute animal UDs, we are interested in the cumulative visit probability P_i for any location i over the entire time interval between t_a and t_b, which can be defined straightforwardly as:

$$P_i = \int_{t_a}^{t_b} P_{it}\, dt \tag{4}$$

In practice, this integral is not easily defined, but we can approximate it using a set of equally spaced time spaces between t_a and t_b and performing numerical integration using the trapezoid

rule. The surface given by $P_{\{i\}}$ will represent the UD for the animal between t_a and t_b. We can compute the surface $P_{\{i\}}$ recursively for each of the n-1 pairs of consecutive fixes in the telemetry dataset to estimate an overall UD.

2.2 Example: Caribou in northern British Columbia

For this example, we look at an individual caribou tracked with GPS during the year 2000 with a location fix recorded every 4 h. A cost surface to represent the time-cost of caribou navigation was derived using two datasets: a DEM derived slope dataset, and a land cover dataset, both with a spatial resolution of 25 m. Slope was translated into movement speed using a modified version of Tobler's (1993) hiking function with a maximum travelling speed of 3.6 km/h (Fancy and White, 1987). Realistically, caribou do not maintain such a speed over long periods, and three movement behaviour states (slow, medium, fast) were identified with maximum travel speed reduced appropriately in the slow (0.3 km/h) and medium (0.9 km/h) classes. Land cover was used to further restrict maximum travel speeds (based on Johnson et al., 2002). Specifically the maximum speed caribou could cross the following land cover classes was adjusted by: forest – 0.8, wetland – 0.5, snow/ice/rock – 0.5, and water – 0.1. To demonstrate the field-based time geography approach and compare with the Brownian bridge model a segment of particularly active movement (between 11h – 15h on 24-01-2000) was chosen, as well as the entire trajectory during the months Jan-Mar 2000 (n = 482 fixes). For this study, all analysis was conducted in R and specifically operations for computing travel costs were implemented using the package '*gdistance*' (Van Etten, 2015).

3. Results and Discussion

The field based time geography model offers unique insight into predicting animal movement probabilities when compared to Brownian bridges (Figure 1). In this segment, there is clearly a non-straight lowest cost path between the two fixes. In the area surrounding the fix on the right hand side we see the presence of topographical barriers that restrict movement, and thus shift probable movements away from the direct path between the two fixes.

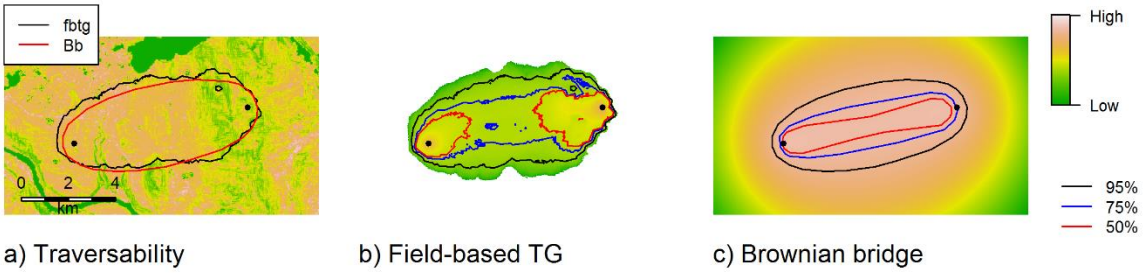


Figure 1: Comparison of a) 95% volume contours for field-based time geography and Brownian bridges overlain on the traversability map (i.e., the inverse of travel cost), b) field-based time geography and c) Brownian bridge models for estimating movement probabilities (and subsequently utilization distributions) between two fixes of a single caribou between 11h and 15 h on 24-01-2000. The 95%, 75%, and 50% volume contours are shown for comparison.

When we compute the overall UD and home range for the months of Jan-Mar 2000 we see a similar result, whereby the field-based time geography UD places less emphasis in certain areas that represent barriers to movement. It is the ability to model such landscape heterogeneity that is lacking from current approaches. Field-based time geography % volume contours are similar in size to the Brownian bridge model which will make it attractive for users wishing to perform similar analysis. However, with field based time geography the output structure of the UD and size of the home range is dependent on the parameters used to generate the cost surface.

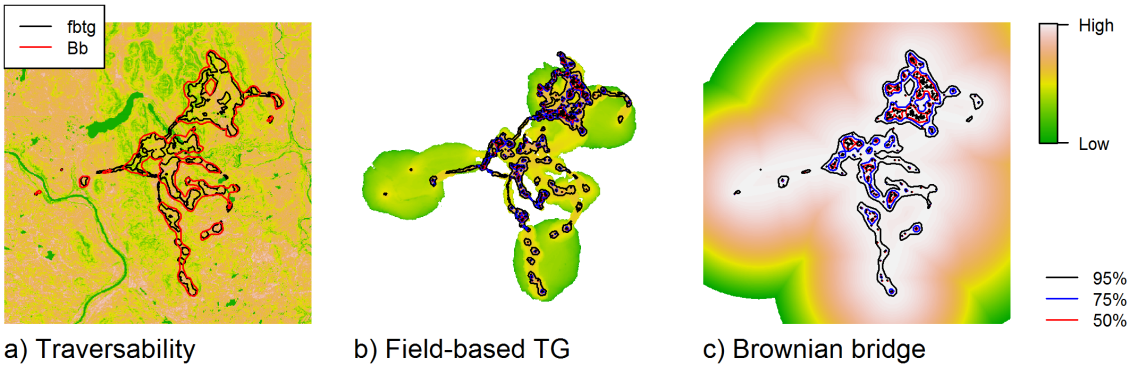


Figure 2: Comparison of a) 95% volume contours for field-based time geography and Brownian bridges overlain on the traversability map (i.e., the inverse of travel cost), b) field-based time geography and c) Brownian bridge models for estimating animal UDs for an individual caribou tracked via GPS during Jan-Mar 2000.

Field-based time geography represents a significant modification to current models used to estimate animal UDs and home ranges. To date no approach has taken advantage of modern GIS techniques for estimating travel costs between two locations, and subsequently, these approaches fail to adequately capture the heterogeneous nature of the landscape. Here field-based time geography was demonstrated using a complex natural landscape in northern British Columbia, but these same concepts could be applied in other environments, for example in marine and avian movement where currents and wind patterns will greatly influence movement potential. The approach has been implemented entirely in the statistical software R (see http://jedalong.github.io/wildlifeTG) which will make it attractive to movement ecologists.

Acknowledgements

The caribou data used here were made available by the British Columbia Ministry of Environment. Thank-you to U. Demšar who read and commented on an earlier version.

References

Burt, W.H., 1943. Territoriality and home range concepts as applied to mammals. J. Mammal. 24, 346–352.

Downs, J.A., Horner, M.W., Tucker, A.D., 2011. Time-geographic density estimation for home range analysis. Ann. GIS 17, 163–171.

Fancy, S.G., White, R.G., 1987. Energy expenditures for locomotion by barren-ground caribou. Can. J. Zool. 65, 122–128.

Hägerstrand, T., 1970. What about people in regional science? Pap. Reg. Sci. Assoc. 24, 7–21.

Horne, J.S., Garton, E.O., Krone, S.M., Lewis, J.S., 2007. Analyzing animal movements using Brownian bridges. Ecology 88, 2354–2363.

Johnson, C.J., Parker, K.L., Heard, D.C., Gillingham, M.P., 2002. Movement parameters of ungulates and scale-specific responses to the environment. J. Anim. Ecol. 71, 225–235.

Long, J.A., Nelson, T.A., 2012. Time geography and wildlife home range delineation. J. Wildl. Manage. 76, 407–413.

Miller, H.J., Bridwell, S.A., 2009. A field-based theory for time geography. Ann. Assoc. Am. Geogr. 99, 49–75.

Tobler, W., 1993. Three presentations on geographical analysis and modeling: Non-isotropic geographic modelling; Speculations on the geometry of geography; and global spatial analysis. University of California, Santa Barbara, CA.

Van Etten, J., 2015. gdistance: Distances and Routes on Geographical Grids (v 1.1-9). R Foundation for Statistical Computing.

Worton, B., 1989. Kernel methods for estimating the utilization distribution in home-range studies. Ecology 70, 164–168.

Multi-Scale Extraction of Regular Activity Patterns in Spatio-Temporal Events Databases: A Study Using Geolocated Tweets from Central Mexico

Pablo Lopez-Ramirez[1] and Oscar Sanchez-Siordia[1]

[1]Centro de Investigación en Geografía y Geomática Ing. Jorge L. Tamayo, CentroGeo

Abstract

This paper proposes a new technique for the extraction of regular activity patterns at different scales, mined from the micro-blogging platform Twitter. The approach is based on the recursive application of the DBSCAN clustering algorithm to the geolocated Twitter feed. This technique includes a novel way to obtain averaged regular activity zones based on the rasterization and aggregation of the Concave Hull of the clusters identified at each resolution level. Since the proposed technique uses only the spatio-temporal characteristics of the geolocated Twitter feed and it does not depend on the messages, it can be extended to work with different spatio-temporal event sources such as mobile telephone records. An experiment was carried out to demonstrate the effectiveness of the technique in the extraction of known activity patterns in the Mexico Central Region.

1 Introduction

The digital breadcrumbs left by social media users have proven to be a valuable source of geographic insights. In the GIS field, they have been used for the detection of events (Atefeh and Khreich) or the characterization of zones through social media activity (Frias-Martinez and Frias-Martinez; Lee et al.), among other things.

In most cases, extraction and characterization of regular activity patterns is of great importance. In this paper we propose a technique for the extraction of such patterns that relies only on the spatio-temporal properties of the Twitter feed. The purpose of this is, on the one hand, to improve on the current available techniques (Lee et al.; Frias-Martinez and Frias-Martinez) and, on the other hand, to be as independent from the nature of the Twitter feed as possible.

The proposed technique is based on the observation that human activity patterns exhibit a wide range of scales (Arcaute et al.), and that the current methods for determining this activity from Twitter messages do not consider this. The proposed approach is based on the recursive application of a clustering algorithm. This approach demands the development of a novel way for averaging the spatial patterns extracted from the data.

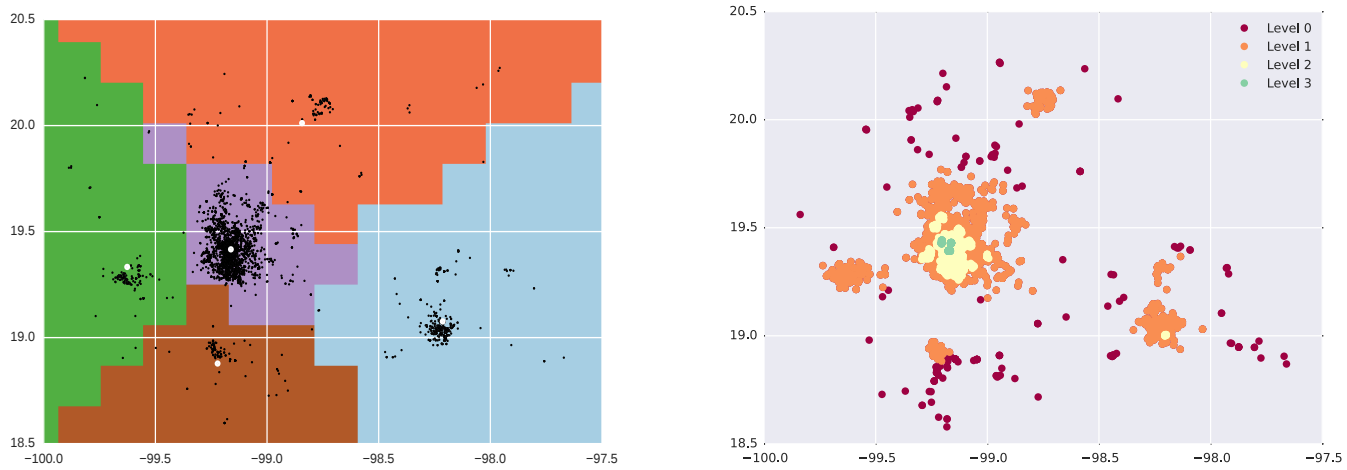
2 Methodology

The idea behind the technique proposed in this paper, is that activity patterns may exhibit a range of scales, and that this scales cannot be represented by a flat tessellation. To overcome this limitation, we propose the use of a recursive clustering strategy that is able to extract the structures present at different scales in the database.

2.1 Recursive Clustering

Several techniques produce a hierarchical representation of point samples, the focus being often on extracting the most significant clusters across all scales. In the case of our study, we need to find clusters that are representative at each resolution level. For this, we will recursively apply DBSCAN (Ester et al.) to the data and thus obtain a hierarchy of clusters across resolution levels.

Initially eps_0 and $MinPoints$ are selected using the *k-dist* plot (Ester et al.) for the whole sample in the corresponding time slice. Then for each iteration of the algorithm, the value for eps_0 is halved while $MinPoints$



(a) K Means. **(b) Recursive DBSCAN.**

Figure 1: Comparison of two clustering strategies over the same points sample. Figure (a) shows the K means clusters and Voronoi polygons around the centroids. Figure (b) shows the clusters obtained in each iteration of the recursive application of DBSCAN.

is held constant. This process is repeated until the number of points in the largest cluster obtained falls below a predefined threshold. In Figure 1 we show a comparison between the iterative application of DBSCAN and the flat tessellation obtained with K-Means.

2.2 Spatio-Temporal Averaging

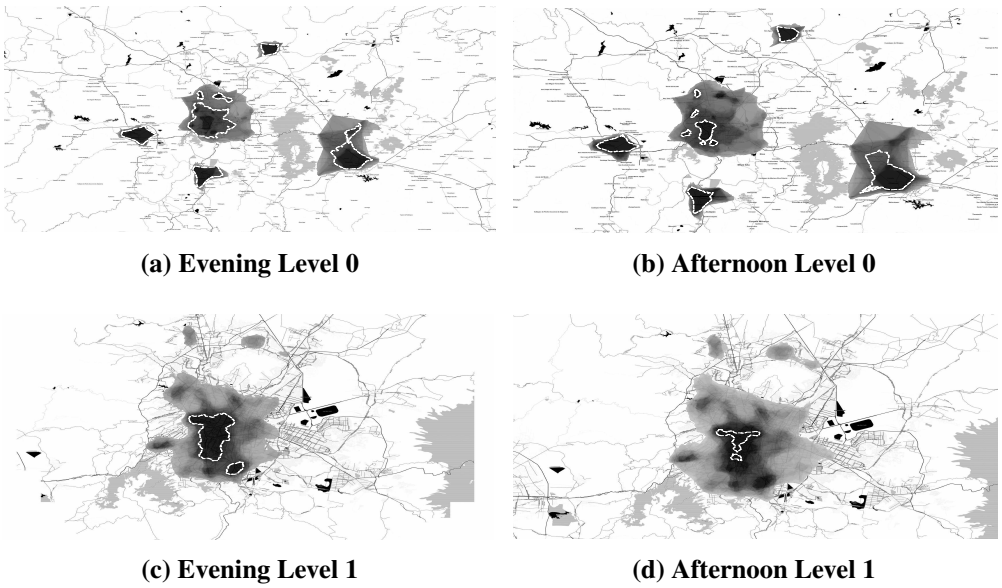
For every day in the input geolocated Twitter feed, a time segmentation similar to those used in Lee et al. and Frias-Martinez and Frias-Martinez will be performed. Then, for each resulting time slice, the recursive clustering strategy described in Section 2.1 will be used. This process is carried out for the whole study period, thus ending with a hierarchical representation for each day.

To average this representations and obtain a single hierarchy representing the whole period for each time slice, the following procedure is carried out:

- A polygon for each cluster is extracted using the *Optimal Alpha Shape* (Edelsbrunner) of the cluster points.
- The polygons obtained are then rasterized to obtain images that have a value of 1, if the pixel belongs to a cluster, or 0 otherwise.
- The resulting images are aggregated to obtain a single image whose values represent the number of days a given pixel has belonged to a cluster.
- The aggregated images are polygonized by cutting them with a threshold value- the number of days a pixel must belong to a cluster in order for it to be considered part of the regular activity. This allows for further characterization of activity zones, such as assigning user counts or other activity measures.

3 Experiment

An experiment was carried out to demonstrate the application of the proposed methodology to the extraction of regular activity patterns around Central Mexico. The database consists of geolocated tweets in the area from October 10 2014 to April 4 2015, (5,415,827 tweets). Prior to the extraction of regular activity zones, we need



(a) Evening Level 0 **(b) Afternoon Level 0**

(c) Evening Level 1 **(d) Afternoon Level 1**

Figure 2: Aggregated activity rasters with threshold cut polygons in dashed lines.

to deal with the *pollution* commonly encountered in the tweeter feed, this means that we must perform some preprocessing to clean up the database. In this case, we filtered out tweets by users that have more than one update within 100 meters of their original location in the same time period. The rationale behind this filtering is that this kind of behavior might be representative of bots, or that it might artificially alter the shape of the clusters without representing the regular activity of the population. The resulting activity patterns for levels 0 and 1 for the Afternoon (14:00 to 18:00 hours) and Evening (18:00 to 22:00 hours) periods are shown in Figure 2.

4 Results Discussion

At the smaller scale (Level 0 in Figure 2), our technique is able to detect the greater metropolitan areas of Mexico City, Puebla, Toluca, Pachuca and Cuernavaca, in the Central Mexico region. As we increase the resolution, the point density in the smaller cities (Puebla, Toluca, Pachuca and Cuernavaca), does not allow for the recursive algorithm to detect larger scale activity. The opposite is true for Mexico city, where we are able to detect up to three scales within the city (not all are shown in the figure).

By comparing the patterns found for the Afternoon and Evening intervals, we see that in the latter, the activity is more dispersed and Level 0 (Figure 2a) shows activity peaks in the northern low-income housing suburbs. On the other hand, the activity for the Afternoon segment (Figure 2b) is more concentrated around the Central Business District (CBD) of the city. This results are consistent with known activity patterns for Mexico City. For example, Suarez and Delgado (Suarez and Delgado) performed a study in the Job-Housing ratio and found the same T-shaped pattern for the CBD. From the same study, we can see that the job to housing ratio of the northern low-income suburbs is very low, which means it is mostly a residential area. This is in line with our results that show activity peaks for those areas only at the Evening intervals.

5 Conclusions

The main improvements of the proposed technique over the available methods are:

1. The ability to detect patterns at different scale levels. This allows us to detect both major urban areas and activity zones within those areas which have a high enough activity density.

2. Using the Alpha Shapes to polygonize the clusters allows us to account for the shape of the activity zones. This represents an improvement compared to the use of Voronoi tessellations.

Qualitative analysis of the regular activity zones obtained, show great accordance with the known activity patterns for the study area, mainly with the spatial distribution of the Job-Housing ratio.

6 Further Work

The next step is quantitative validation of the results obtained. For this, activity data for the study area is essential. One approach would be coupling Job-Housing ratio maps with mobility patterns extracted from Origin-Destination surveys to disaggregate the latter to the scales needed for our analysis.

Also, a more tractable way of setting parameter values (such as eps_0 and $MinPoints$ at each iteration) is needed. Ground truthing against measured activity distributions would provide basis for a calibration-validation approach. In the same line of thought, it would be important to test different clustering algorithms, such as HDBSCAN, which does not introduce additional parameters or STDBSCAN, which is purely spatio-temporal.

Finally, the multi-scale regular activity zones could be used to detect unusual crowd activity at various scales. The rationale behind this is that unusual events also exhibit scale differences. For example, it is known that important large scale events, such as the Super Bowl or the Arab Spring, produce a general increase of messages in the social networks, while localized small-scale occurrences, such as festivals, demonstrations or accidents, produce small clusters of messages around the locations affected.

References

Elsa Arcaute, Carlos Molinero, Erez Hatna, Roberto Murcio, Camilo Vargas-Ruiz, Paolo Masucci, Jiaqiu Wang, and Michael Batty. Hierarchical organisation of Britain through percolation theory. URL http://arxiv.org/abs/1504.08318.

Farzindar Atefeh and Wael Khreich. A Survey of Techniques for Event Detection in Twitter. 31(1):132–164. ISSN 1467-8640. doi: 10.1111/coin.12017. URL http://onlinelibrary.wiley.com/doi/10.1111/coin.12017/abstract.

Herbert Edelsbrunner. Smooth surfaces for multi-scale shape representation. In P. S. Thiagarajan, editor, *Foundations of Software Technology and Theoretical Computer Science*, number 1026 in Lecture Notes in Computer Science, pages 391–412. Springer Berlin Heidelberg. ISBN 978-3-540-60692-5 978-3-540-49263-4. doi: 10.1007/3-540-60692-0_63. URL http://link.springer.com/chapter/10.1007/3-540-60692-0_63.

Martin Ester, Hans-peter Kriegel, Jörg S, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press.

Vanessa Frias-Martinez and Enrique Frias-Martinez. Spectral clustering for sensing urban land use using Twitter activity. 35:237–245. ISSN 0952-1976. doi: 10.1016/j.engappai.2014.06.019. URL http://www.sciencedirect.com/science/article/pii/S0952197614001419.

Ryong Lee, Shoko Wakamiya, and Kazutoshi Sumiya. Discovery of unusual regional social activities using geo-tagged microblogs. 14(4):321–349. ISSN 1386-145X, 1573-1413. doi: 10.1007/s11280-011-0120-x. URL http://link.springer.com/article/10.1007/s11280-011-0120-x.

M. Suarez and J. Delgado. Is Mexico City Polycentric? A Trip Attraction Capacity Approach. 46(10):2187–2211. ISSN 0042-0980. doi: 10.1177/0042098009339429. URL http://usj.sagepub.com/cgi/doi/10.1177/0042098009339429.

Deserts in the Deluge: TerraPopulus and Big Human-Environment Data

S. M. Manson[1], T. A. Kugler[2], D. Haynes II[2]

[1] Department of Geography, Environment, and Society. University of Minnesota. 414 Social Sciences, 267 19th Avenue South, Minneapolis, MN 55455, USA
Email: manson@umn.edu

[2] Minnesota Population Center, 50 Willey Hall, 225 – 19th Avenue South, Minneapolis, MN 55455
Email: takugler; dhaynes@umn.edu

Abstract

Terra Populus, or TerraPop, is a cyberinfrastructure project that integrates, preserves, and disseminates massive data collections describing characteristics of the human population and environment over the last six decades. TerraPop has made a number of GIScience advances in the handling of big spatial data to make information interoperable between formats and across scientific communities. In this paper, we describe challenges of these data, or 'deserts in the deluge' of data, that are common to spatial big data more broadly, and explore computational solutions specific to microdata, raster, and vector data models.

1. Introduction

Over the past six decades, the world's population more than doubled. Sharp interregional differences in growth rates—together with unprecedented urbanization and international migration—led to dramatic spatial redistribution of population. Economic changes were equally remarkable, as world per-capita gross domestic product roughly doubled (Rosa *et al.* 2010; Bloom 2011). This extraordinary global demographic and economic growth has ushered in alarming environmental degradation, resource depletion, and climate change (Ehrlich, Kareiva, and Daily 2012).

Scientific and policy bodies have called for more richly-detailed data to support the research and informed decisions necessary to meet the challenges of rapid social and environmental change (Millett and Estrin 2012). There is particular interest in the 'data deluge' or 'big data', or research based on datasets that are vastly larger than those traditionally used in most fields, and which in turn entail new forms of processing and analysis.

However, there are deserts in the deluge of data. At the level of data as such, scholars untangling human-environment interactions face a dearth of spatially-detailed multidecadal data. While some relevant data are available, such as climate observations, there is surprisingly little detailed information about many social and natural features for most of the globe before the year 2000 (Nelson *et al.* 2010). At the level of methods, there are similar shortfalls in our ability to store, manipulate, and analyze spatial big data (Wang and Liu 2009). And at the level of theory, we face many unresolved challenges in representing social and biophysical entities and relationships that operate at multiple levels of organization, over space, and through time (O'Sullivan and Manson 2015). TerraPop address these deserts in deluge of big spatial data.

2. TerraPop

TerraPop addresses challenges in data, methods, and theory in big spatial data by using location-based data integration to make heterogeneous data interoperable and thereby break down barriers to interdisciplinary research. Researchers can combine data across three major data classes – microdata, raster, and area-level. For example, TerraPop can summarize raster data derived from satellite images to determine the percentage of each municipio in Brazil

covered by trees, and then attach that contextual information to each record of census microdata (or data that represents an individual person or set of household). TerraPop has population data for over 170 countries, global long-term climate data, a variety of global land cover and land use datasets, and the geographic boundaries necessary to support integration across the collection. Many of these data sets are unique (especially those on demographics and socioeconomic characteristics) and help address one of the primary deserts in the deluge, the dearth of data on human populations for much of the globe prior to 2000. TerraPop makes these global datasets interoperable across time and space, disseminates them to the public and to multiple research communities, and preserves these resources for future generations.

3. Spatial high-performance computing

We addressed a number of fundamental GIScience challenges in order to integrate and disseminate this vast data collection. We developed workflows and supporting software tools for processing data and metadata. We developed a suite of Python-ArcGIS tools that enable efficient boundary data processing of current and historic population datasets, automate temporal harmonization, and manage regionalization to protect respondent confidentiality (Kugler *et al.* 2015). We also developed a metadata management application that tracks data provenance from the original sources through all TerraPop processing steps and produces complete descriptions of the final data.

At the core of the TerraPop infrastructure is a set of spatial high-performance computing solutions that transform microdata, vector data, and raster data. We have microdata describing 250 billion microdata characteristics, 300 billion vector data points, and over a trillion pixels of raster data. These data are available via a web interface or application program interface (Figure 1). These large datasets create research opportunities and challenges. Parallel computation, the usual solution for such problems, is fundamentally difficult for big spatiotemporal data (Eldawy and Mokbel 2015). Many computational problems are "embarrassingly parallel" because they can be solved by partitioning and distributing data among nodes in a computing cluster, solving the problem for a subset of data on each node, and then collating the results.

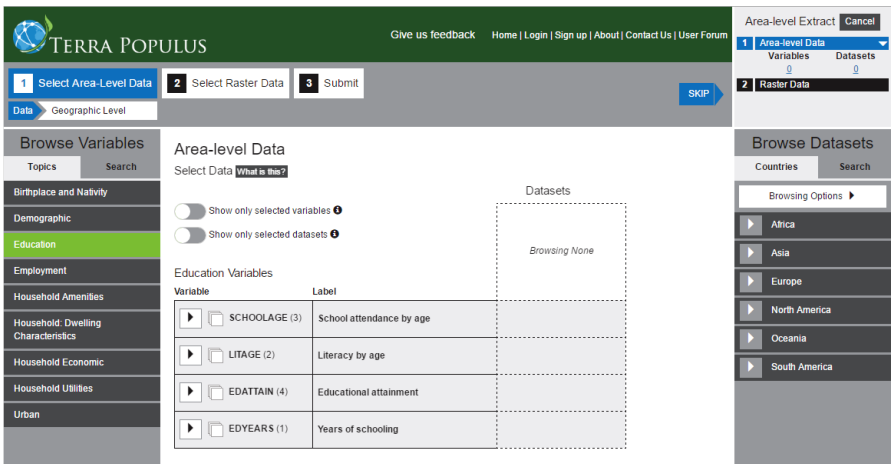


Figure 1. TerraPopulus web interface for combining and abstracting data.

Standard high-performance computing approaches are often inefficient or unworkable for spatiotemporal data, due to the difficulty of preserving spatial and temporal relationships across nodes (Ding and Densham 1996). Microdata require a distribution algorithm that preserves relationships between individuals and their households. Raster data and vector data embody

complex spatial and topological relationships that are essential for answering most spatial problems, relationships that must be preserved when partitioning across nodes. Parallel computing for spatial big data is an area of active research, but most existing computing platforms cannot handle multiple spatial data models or perform spatial data handling and analytics commonly found in even the most basic geographic information systems (Ray *et al*. 2015). In particular, most approaches to parallelization are limited in scope or provide extensions to existing frameworks such as MapReduce and column store databases. While this work is very promising, no existing systems are robust or wide-ranging enough for GIScience production environments like TerraPop (Haynes *et al.* 2015).

Microdata. Microdata are most often stored as hierarchical fixed-width text or binary files, where each line represents an individual person or set of household characteristics. Challenges to high-speed processing of individual-level data derive from the size and complexity of the data and the need to conduct complicated queries across multiple samples with thousands of attributes and multiple embedded relationships. We implemented Apache Spark's Parquet columnar storage database and found significant performance gains across these queries over standard Java-based approaches. Query execution speed has increased by a factor of 10 to 300 for a variety of common operations and using Parquet promises further gains because Parquet offers record shredding and assembly (Armbrust *et al*. 2015).

Vector data. Large vector datasets are difficult to parallelize because spatial relationships such as adjacency and connectivity must be preserved across nodes (Ray *et al*. 2013; Puri and Prasad 2013). We use the leading open-source spatial computing framework, PostgreSQL/PostGIS, because it offers deep data handling and analytical capabilities. However, PostgreSQL does not natively support parallel queries, though multiple projects are trying to scale PostgreSQL onto machine clusters (e.g., GridSQL, Stado, Postgres-XC, CitusDB and Postgres-XL). We extensively tested these projects and determined that they do not support parallel spatial processing well (although several projects are working on the problem) so we have been developing a prototype vector analytic engine that partitions a PostgreSQL database across computing nodes. We chose this approach based on evidence that parallel relational databases like PostgreSQL can perform significantly better than MapReduce systems (Pavlo *et al.* 2009). Our work to date has significantly improved performance in analyzing vector datasets, offering near linear speedup when adding nodes by sharding spatial queries across a cluster of machines where a PostgreSQL database instance is run on each node for simple topographical operations such as determining whether a polygon intersects a line or other polygon (Haynes *et al.* 2015; Ray *et al.* 2014). This work thereby addresses fundamental research needs in spatial high performance computing (Vo, Aji, and Wang 2014).

Raster data. Large raster datasets are difficult to parallelize because of the sheer volume of data involved, the need to preserve spatial relationships among grid cells, and the large number of varying raster operations that are needed to manipulate data. While the combination of PostgreSQL/PostGIS offers a comprehensive set of raster analytics, that approach does not handle large rasters well because row limit sizes are often exceeded by raster datasets (Stonebraker *et al.* 2011). By experimenting with array data structures, we have doubled or tripled performance for most operations while ensuring that larger raster layers do not fail outright. We are also experimenting with web applications to offer easy and fast access to these data via textual and web mapping interfaces (Manson *et al.* 2012).

4. Conclusion

TerraPop incorporates the largest and most comprehensive available collections of data on human activities and behavior, along with important global environmental datasets. The population and environmental data are multiscale over time and space, have multiple levels of

hierarchy, and cover a remarkable range of topics. To manage the scale, complexity, and heterogeneity of the data, we will engage the leading edge of data science and develop new technologies and processes. Innovative solutions are needed through the entire data life cycle, including collection, preservation, analysis, dissemination, and long-term access and management. TerraPop will provide open-source software, metadata, and workflows that can overcome these challenges and that can readily be adapted to spatiotemporal data in multiple scientific domains. In particular, our work on spatial high-performance computing will address critical bottlenecks in the integration and dissemination of massive spatiotemporal datasets.

Acknowledgements

This work is supported in part by the National Science Foundation OCI: Terra Populus: A Global Population/Environment Data Network (0940818), the National Institutes of Health supported Minnesota Population Center (R24 HD041023), and the Resident Fellowship program of the Institute on the Environment.

References

Armbrust, Michael, Reynold S Xin, Cheng Lian, Yin Huai, Davies Liu, Joseph K Bradley, Xiangrui Meng, Tomer Kaftan, Michael J Franklin, and Ali Ghodsi. 2015. "Spark Sql: Relational Data Processing in Spark." In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, 1383–94. Melbourne, Australia: ACM.

Bloom, David E. 2011. "7 Billion and Counting." *Science* 333 (6042). American Association for the Advancement of Science: 562–69.

Ding, Yuemin, and Paul J Densham. 1996. "Spatial Strategies for Parallel Spatial Modelling." *International Journal of Geographical Information Systems* 10 (6). Taylor & Francis: 669–98.

Ehrlich, Paul R, Peter M Kareiva, and Gretchen C Daily. 2012. "Securing Natural Capital and Expanding Equity to Rescale Civilization." *Nature* 486 (7401): 68–73.

Eldawy, Ahmed, and Mohamed F Mokbel. 2015. "The Era of Big Spatial Data: A Survey." *Information and Media Technologies* 10 (2). Information and Media Technologies Editorial Board: 305–16.

Haynes, David, Suprio Ray, Steven M Manson, and Ankit Soni. 2015. "High Performance Analysis of Big Spatial Data." In *Big Data 2015: IEEE International Conference on Big Data*, 1953–57. Santa Clara, California: Institute of Electrical and Electronics Engineers.

Kugler, Tracy A, David C Van Riper, Steven M Manson, David A Haynes II, Joshua Donato, and Katie Stinebaugh. 2015. "Terra Populus: Workflows for Integrating and Harmonizing Geospatial Population and Environmental Data." *Journal of Map & Geography Libraries* 11 (2). Taylor & Francis: 180–206.

Manson, S. M., L. Kne, K. Dyke, J. Shannon, and S. Eria (2012). Using eye tracking and mouse metrics to test usability of web mapping navigation. *Cartography and Geographic Information Science* 39 (1): 48-60

Millett, Lynette I, and Deborah L Estrin. 2012. *Computing Research for Sustainability*. National Academies Press.

Nelson, E., H. Sander, P. Hawthorne, M. Conte, S M Manson, and S. Polasky. 2010. "Projecting Global Land Use Change and Its Affect on Ecosystem Service Provision and Biodiversity with Simple Techniques." *PLoS ONE* 5 (12): e14327.

O'Sullivan, D. and S. M. Manson (2015). Do Physicists Have 'Geography Envy'? And What Can Geographers Learn From It? *Annals of the Association of American Geographers* 105 (4): 704-722.

Pavlo, Andrew, Erik Paulson, Alexander Rasin, Daniel J Abadi, David J DeWitt, Samuel Madden, and Michael Stonebraker. 2009. "A Comparison of Approaches to Large-Scale Data Analysis." In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, 165–78. Providence, Rhode Island: ACM.

Puri, Shruti, and Sushil K Prasad. 2013. "Efficient Parallel and Distributed Algorithms for GIS Polygonal Overlay Processing." In *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 27th International*, 2238–41. Boston, Massachusetts: IEEE.

Ray, Suprio, Angela Demke Brown, Nick Koudas, Rolando Blanco, and Anil K Goel. 2015. "Parallel in-Memory Trajectory-Based Spatiotemporal Topological Join." In *Big Data (Big Data), 2015 IEEE International Conference on*, 361–70. Santa Clara, California: Institute of Electrical and Electronics Engineers.

Ray, Suprio, Bogdan Simion, Angela Demke Brown, and Ryan Johnson. 2013. "A Parallel Spatial Data Analysis Infrastructure for the Cloud." In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 284–93. ACM.

———. 2014. "Skew-Resistant Parallel in-Memory Spatial Join." In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*, 6–12. Aalborg, Denmark: Association for Computing Machinery.

Rosa, Eugene A, Andreas Diekmann, Thomas Dietz, and Carlo Jaeger. 2010. *Human Footprints on the Global Environment: Threats to Sustainability*. Cambridge, MA: MIT Press.

Stonebraker, Michael, Paul Brown, Alex Poliakov, and Suchi Raman. 2011. "The Architecture of SciDB." In *Scientific and Statistical Database Management*, 1–16. Berlin: Springer.

Vo, Hoang, Ablimit Aji, and Fusheng Wang. 2014. "SATO: A Spatial Data Partitioning Framework for Scalable Query Processing." In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 545–48. Dallas, Texas: Association for Computing Machinery.

Wang, Shaowen, and Yan Liu. 2009. "TeraGrid GIScience Gateway: Bridging Cyberinfrastructure and GIScience." *International Journal of Geographical Information Science* 23 (5). Taylor & Francis: 631–56.

Uber vs. Taxis: Event detection and differentiation in New York City

Grant McKenzie[1], Carlos Baéz[2]

[1] Department of Geographical Sciences, University of Maryland, College Park, USA
[2] Department of Geography, University of California, Santa Barbara, USA
Email: gmck@umd.edu; carlos.baez@geog.ucsb.edu

Abstract

The recent rise of ride-sourcing services such as Uber have significantly changed the transportation landscape. This work takes a first step in differentiating Uber and taxi transportation methods through events attended by their passengers. Using a sample of Uber and taxi pick-up times and locations in New York City, we show that events can be detected within each platform. Through identification of a select few of these events, this work takes a preliminary step in showing that there is a difference in the types of events that are attended by Uber users and taxi passengers.

1. Introduction

Historically, taxicab companies have controlled the largest share of the *for-hire vehicle* (FHV) market in the United States. Over the past few years, however, alternative transportation options have arisen such as *Uber* and *Lyft* that rely on the use of online-enabled platforms to connect passengers with drivers. Together with others, these types of ride-sourcing companies, often called *Transportation Network Companies*(TNC), have significantly disrupted the traditional transport model, namely taxi service. By some accounts (Certify, 2015), TNCs now account for 46% of some U.S.-based FHV markets. This dramatic shift in the means of transportation has spurred a lot of research and discussion on its impact and significance (NRC-TRB, 2015; Hall et al., 2015). From a spatiotemporal research perspective, this shift has also lead to some important questions related to the differences between these services as well as the people that use them.

This short paper presents a *first step* in exploring the differences between traditional taxi services and TNCs as described through events[1] in New York City. **Specifically, this work addresses the following research questions**.

- Is it possible to detect events based on passenger pick-up times and locations in publicly available Uber and taxi data?
- Do events detected in the Uber dataset differ from those detected in the taxi dataset?
- Do these findings support existing research showing that there are differences between TNC users and taxi riders? We approach this question through identifying a select sample of detected events.

As stated in this last question, existing work in this area indicates that there are differences in the demographics of taxi and TNC passengers. Specifically, TNC user surveys suggest that, relative to taxi users, TNC passengers are younger and posses a higher average level of education (Rayle et al., 2016). Our work continues on this

[1]See work by Worboys (2005) in discussing and defining events.

thread by exploring differences in the types of events that are attended by these two groups of passengers. Furthermore, this research builds off of work by Zhang et al. (2015) on detecting events in Chinese taxi data, although we take it several steps further in comparing taxi-based events with those discovered in TNC data.

2. Event Detection

Data for this work was accessed via the New York City Taxi & Limousine Commission. In total, 80,295,320 Yellow taxi pick-up locations[2] and 4,534,327 Uber pick-up locations[3] were used for this research (drop-off locations are not available for the Uber data). The data is spatially bounded by the extent of New York City and temporally bounded between April 1 and September 30, 2014. The attributes used in this work include geospatial coordinates and timestamps for passenger pick-ups.

For each of the two datasets, the following analysis took place. The timestamps for each pick-up were rounded to the nearest hour and aggregated to counts by intersecting with the New York city census tract spatial data from 2014. The mean and standard deviation for the number of pick-ups per census tract, aggregated to the hour of a typical week (24 values) were calculated. This produced mean pick-ups and standard deviation values for 168 hours in a week in 2,162 census tracts. Setting a minimum threshold mean of 10 pick-ups per hour, per census tract significantly reduced this value to 114 census tracts and 152 hours of the week.

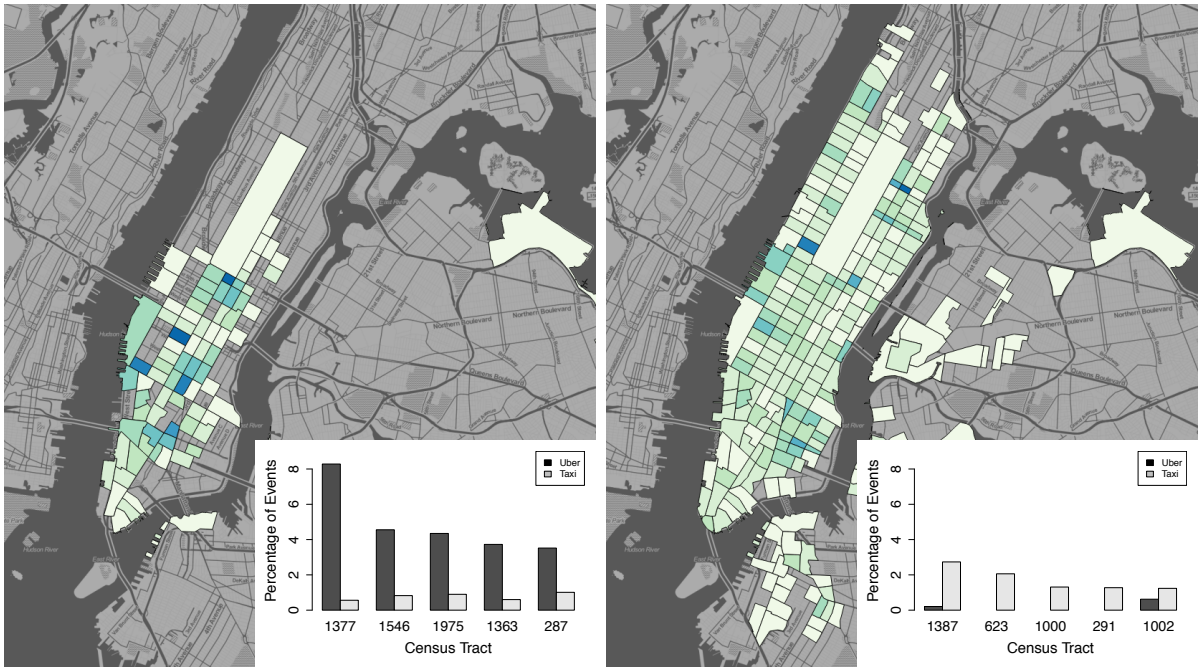
Events were detected in each dataset by comparing the number of pick-ups on any given day, time and census tract with the amount of pick-ups typical for that hour of the week (mean count). An *event* was recorded if the number of pick-ups was above three standard deviations from the mean. Using this approach, 485 events were discovered in the Uber dataset and 2,671 in the taxi data.

3. Event Differentiation

Given the events detected in both datasets, the next step was to identify events that were detected in both datasets and those that were specific to the taxi or Uber data. Setting a temporal buffer of two hours before and after an Uber-identified event, we searched for events in the taxi data that occurred in the same census tract within this specified temporal window. This resulted in 17 events identified in both the taxi and Uber datasets, meaning the majority of events were identified exclusively in the taxi or Uber data and not in both. Figure 1 depicts detected events through two chloropleth maps (quantile breaks) ranging from a high (blue) to low (light green) number of events for both Uber and taxi data. The inset bar plots show the top five census tracts by percentage of overall events detected for each mode of transportation. Notably, each dataset's highest event counts occurred in different census tracts. This suggests that certain types of events are aligned with Uber users while others lend themselves to taxi passengers. The difference in census tracts also suggests that there may be regional influences, although further investigation is beyond the scope of this short paper.

Given these common and mode-specific events, we manually investigated a number of the events based on their location and temporal parameters. Using application pro-

[2] http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
[3] https://github.com/fivethirtyeight/uber-tlc-foil-response



(a) Uber (b) Taxi

Figure 1: Choropleth map showing number of events by census tract normalized by area. Inset bar plots show top 5 census tracts by percentage of overall events detected for that type (Uber or taxi). Base map by Stamen maps.

gramming interfaces (APIs) such as *Eventful*[4] along with venue specific websites (e.g., Madison Square Gardens), we were able to identify a number of events which are shown as examples in the following sections.

3.1 Taxi-specific Events

A large number of events were identified in the taxi dataset that did not appear in the Uber data. Examples of these types of events are *Summer by the Sea* on June 9 at Rockefeller Center and *David Gray* on June 23 at Madison Square Gardens. *Summer by the Sea* is an annual celebrity chefs' tribute to city meals-on-wheels while *David Gray* is an adult-contemporary musician who performed a live concert. Arguably, both of these events appeal to an older age demographic rather than young adults or teenagers. While these are only two examples, it is interesting to note that these events were only detected in the taxi data.

3.2 Uber-specific Events

Similar to the taxi-specific events, a number of Uber identified events could not be matched to the taxi data. For example, *NBC Upfront* on May 12 at Jacob Javits Convention Center and *Eric Prydz* on September 27 at Madison Square Gardens were identified in the Uber data, but not the taxi data. *NBC Upfront* is a large television media event organized by the *NBC* network to showcase new television shows for advertisers and *Eric Prydz* is a popular electronic music recording artist who gave a live performance. One possible explanation as to why these events were only detected in the Uber data is that on

[4] http://www.eventful.com

average, attendees of these events are more technologically savvy than those that attend events via taxi. Electronic music is quite popular among young adults and large television media advertising events are attended by those that actively use the latest technology. Given TNC users tend to be younger and more technologically aware than non-users (Rayle et al., 2016; Murphy, 2016), it is notable that these events are only detected in the Uber data.

3.3 Common Events

Finally, a number of events were identified that exist in both the taxi and Uber datasets. Two examples of these include *Dinner in White* on August 25 at Battery Park City and *Bruno Mars & Pharrell* on July 14 at Madison Square Gardens. *Dinner in White* was described as "one of the biggest summer events in New York City" and *Bruno Mars & Pharrell* was a music concert event featuring two of the highest selling recording artists at the time. Both large events appeal to a wide demographic, both young and old from diverse technological backgrounds.

4. Conclusions & Next Steps

This work presents a first step in differentiating Uber and taxi transportation through events attended by their passengers. Using a sample of Uber and taxi pick-up times and locations in New York City, we showed that events can be detected. Moreover, while some events were identified in both the taxi and Uber data, a larger number were detected within only one of the datasets. Through identification of a few of these events, we have taken a preliminary step in showing that there is a difference in the types of events that are attended by TNC users and traditional taxi passengers.

Next steps in this work will be to explore event detection at a variety of spatial and temporal resolutions. Census block groups were suitable as a preliminary boundary, but additional events may be discovered through increasing or decreasing the spatial resolution. Additionally, the time-window during which events occur ranges significantly depending on the type of event. This will be explored in greater detail along with any demographic information associated with the individuals that attend these events.

Lastly, the long term goal of this work is to understand and differentiate motivating factors for human urban mobility. This will have a significant impact on transportation planning and policy, emergency management, and urban infrastructure.

References

Certify (2015). Sharing the road: Business travelers increasingly choose uber. https://www.certify.com/infograph-sharing-the-road.aspx. Accessed: 2015-04-04.

Hall, J., Kendrick, C., and Nosko, C. (2015). The Effects of Uber's Surge Pricing: A Case Study.

Murphy, C. (2016). Shared mobility and the transformation of public transit. Technical report, American Public Transportation Association.

NRC-TRB (2015). Between public and private mobility: Examining the rise of technology-enabled transportation services. Technical report, National Academy of Sciences, Washington, D.C.

Rayle, L., Dai, D., Chan, N., Cervero, R., and Shaheen, S. (2016). Just a better taxi? a survey-based comparison of taxis, transit, and ridesourcing services in san francisco. *Transport Policy*, 45:168–178.

Worboys, M. (2005). Event-oriented approaches to geographic phenomena. *International Journal of Geographical Information Science*, 19(1):1–28.

Zhang, W., Qi, G., Pan, G., Lu, H., Li, S., and Wu, Z. (2015). City-scale social event detection and evaluation with taxi traces. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3):40.

Crowd-sorting: reducing bias in decision making through consensus generated crowdsourced spatial information

H. D. McNair[1], L. M. Arnold[2]

[1]University of Canterbury, Private Bag 4800, Christchurch 8020, New Zealand
Email: hamish.mcnair@pg.canterbury.ac.nz

[2]Curtin University, Kent Street, Bentley, 6102, Western Australia
Email: lesley.arnold@curtin.edu.au

[1, 2] Cooperative Research Centre for Spatial Information, Australia

1. Introduction

Crowdsourcing complements existing planning processes. It enables people to contribute information to decision-making processes in a way that promotes outcomes better suited to their needs. People have a unique insight into the world they know, interact with, and live in, and therefore have an affinity with local issues. However, by the time traditional public consultation (e.g. town hall meeting) is undertaken, unfavourable planning decisions may have inadvertently been made. Potential oversights can be mitigated by including insights from the crowd earlier in the planning process.

Spatial data and information is used to model a truth that represents the existing landscape and identifies issues that the planning process needs to address. Spatial data and information, however, are fundamentally distinct. Data is an abstraction of the real world based on measured properties, whereas information concerns the interpretation of a scenario or event – whether by way of a first-hand account or analysis of data (Ackoff 1989). Experts use interpreted information to decide what issues should be addressed by the planning process and how. On the other hand, crowdsourced spatial information can be obtained directly from the crowd, not via expert processing. This *crowd truth* offers a balance to the *single truth* that may result from mathematical data processing (Aroya and Welty 2015).

While crowdsourced spatial information can help define the needs of a community it must be compatible with existing processes to influence change. This paper introduces Sensibel as a mechanism for crowdsourcing data and information about the current performance of cycling infrastructure. We then present an approach to *crowd-sorting* this information to increase both its usability and trustworthiness. Crowd-sorting uses the spatial characteristics of the information alongside its provenance. It considers how the crowd interacts with individual observations and the areas they describe. The aim is to use these interactions to crowd-sort contributions and produce more constructive inputs to the planning process.

1.1 Crowdsourcing in Planning

Inclusion of crowdsourcing in planning processes often overlooks local insights by focussing on rudimentary data elements, not prevailing local perspectives. Such a process is depicted in Figure 1. For example, assessing cycle routes using crowdsourced data from Strava (http://strava.com) simply shows which routes people use the most. The issue being that high volume does not necessarily equate to a well performing route. Incorporating crowdsourced information early in the planning process may better address the needs of the community. Brown (2015) indicates that land use planning can be improved by spatially arranging comments from the public. This spatial reference provides a link between data centric processes and crowdsourced information. A similar approach was applied during the formation of Helsinki's master plan (Kahlia-Tani *et al*. 2016). However, distilling public

contributions into discernible courses of action is challenging given the potentially broad range of comments.

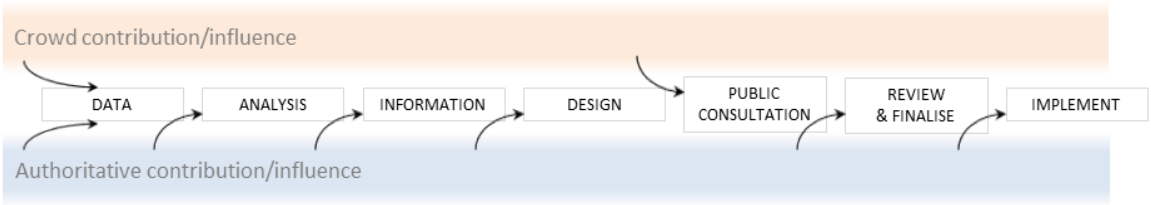


Figure 1. Contributions and influences in the planning process.

2. Generating Consensus Using Sensibel and PROV

Sensibel (https://fabriko.squarespace.com/sensibel) allows the crowd to define locations where cycle infrastructure is performing above or below expectation. The bicycle bell sized unit attaches to a cyclist's handlebars and has positive and negative buttons which, when pressed, indicate the cyclist's opinion of that location. The hardware synchs via Bluetooth with the associated mobile app. The smartphone's GPS captures the route (line feature) and the location of individual positive/negative points. Data is submitted to a central repository where spatial clustering of all cyclists' point data produces locations that the crowd can comment on via a web app. They can also agree (up vote) or disagree (down vote) with these opinions.

A possible use case is presented in Figure 2 where users **a** and **b** have negative experiences and submit points representing these ($\mathbf{P_a}$, $\mathbf{P_b}$). A cluster ($\mathbf{Cl_1}$) forms at this location. Users **e** and **f** also pass through the area but do not contribute points. Users **c** and **d** have not passed through this area but may contribute opinions via web app. All contributors (**a** through **f**) are able to submit comments about the area the cluster represents and up or down vote these comments according to their opinion of them.

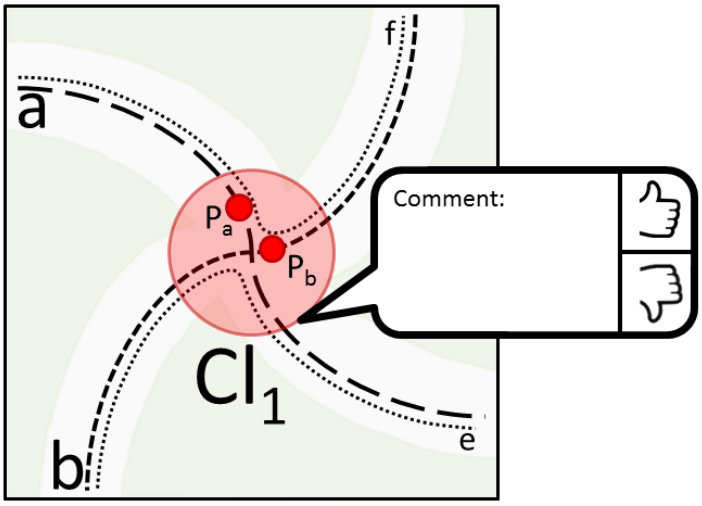


Figure 2. Sensibel in action.

The elements contributing to the formation of crowd consensus can be assessed based on how information is created and used. Provenance details an entity's creation and subsequent use or modification. The W3C developed PROV Data Model (Moreau and Missier 2013) enables detailed provenance information to be recorded and its semantic capabilities permit analysis of the relationships between descriptive elements. PROV defines three general classes – entity, activity and agent:

- An **entity** (real, digital, or otherwise). For Sensibel, entities include the data/information representing **positive/negative points**, **cyclists' routes**, **comments**, and **up/down votes**.
- An **activity** uses or acts upon entities. Users **submit** positive/negative points, **pass through** cluster areas, **contribute** comments, or **up/down vote** comments.
- An **agent** has authority over activities performed or entities used. These are Sensibel **users**.

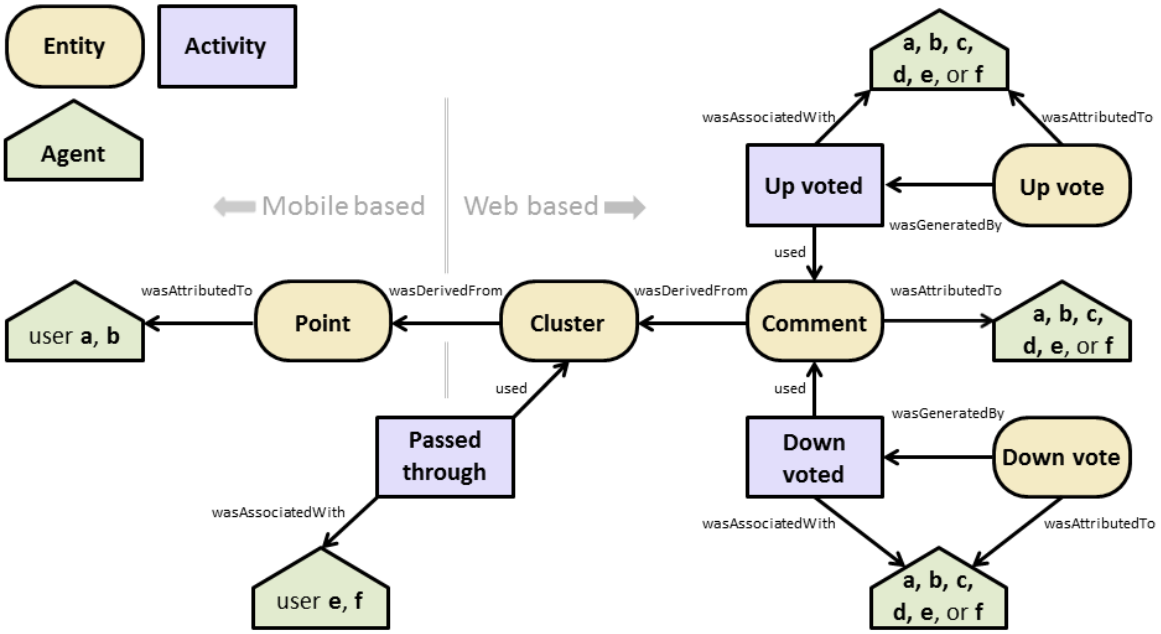


Figure 3. A simplified provenance graph detailing the Sensibel use case from Figure 2.

By defining types of user we can investigate how different kinds of interactions influence the resulting information. We propose 3 tiers of participation and their likely contribution to crowd-sorted information:

- Active-participants contribute points that form clusters (e.g. users **a** and **b**). These cyclists identify something in the area they consider important. We foresee their comments and votes as providing valuable insights into positive or negative issues associated with a site.
- Passive-participants pass through a cluster-area without identifying issues (e.g. users **e** and **f**). We predict they will have enough knowledge of the area to provide supporting evidence via web app voting.
- Remote-participants aren't registered as having visited the area (e.g. users **c** and **d**). Their comments may prove useful, but we are reliant on those familiar with the area verifying their opinions via web app voting.

PROV permits filtering by these 3 tiers – the three different types of **agent** in Figure 3 – so we can observe the impact each has on crowd consensus. This enables the contributions of each type of user to be assessed so emphasis can be placed on the most relevant contributions at each stage in the Sensibel system. It also allows for a broader examination of the importance associated with being in the place about which you are providing information. This concept was employed by Celino (2015) when using the crowd and an individual's location (with PROV) to curate a crowdsourced spatial data set. In a planning context this allows separation of the perspectives of active-participants who have recently interacted with

the site and remote-participants providing comment based on their general – and potentially dated – opinion of the area.

Crowd-sorting seeks to produce inputs to the planning process that represent the views of the crowd in a format that complements existing planning processes, as illustrated by Figure 4. The top comments for each cluster are assigned its location and can be used in conjunction with other spatial data sets. For example, a spatial join with asset data sets may highlight reoccurring issues (such as an aversion to a type of railway crossing) that induces improvements in infrastructure design.

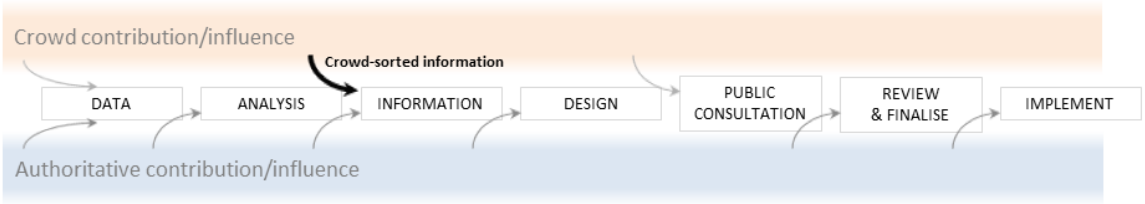


Figure 4. Crowd-sorted information in the planning process.

3. Conclusion and future work

This paper discussed the value of including crowdsourced spatial information in the planning process and introduced Sensibel as a system for collecting crowdsourced geographic data and information relating to cycle infrastructure performance. Crowd-sorting is a mechanism by which these contributions can be used to produce more representative and constructive inputs to the planning process.

The next stage is to collect data and information from a test group of users. Planners will be engaged to determine characteristics of the information that should be emphasised in order for it to have the greatest impact in the planning process. Once the elements most conducive to producing crowd consensus have been identified the semantic capabilities of PROV can be implemented to achieve effective and practical information sorting.

Acknowledgements

This work has been supported by the Cooperative Research Centre for Spatial Information (CRCSI), whose activities are funded by the Australian Government.

References

Ackoff RL, 1989. From data to wisdom. *Journal of applied systems analysis*, 16(1):3-9.

Aroyo L and Welty C, 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15-24.

Brown G, 2015. Engaging the wisdom of crowds and public judgement for land use planning using public participation geographic information systems. *Australian Planner*, 52(3):199-209.

Celino I, 2015. Geospatial dataset curation through a location-based game. *Semantic Web*, 6(2):121-130.

Kahila-Tani M, Broberg A, Kyttä M and Tyger T, 2015. Let the citizens map—public participation GIS as a planning support system in the Helsinki master plan process. *Planning Practice & Research*, 31(2):1-20.

Moreau L and Missier P (eds), 2013. PROV-DM: The PROV Data Model. World Wide Web Consortium.

Modeling Place as a Relationship between a Person and a Location

J. Mennis[1], M.J. Mason[2]

[1]Department of Geography and Urban Studies, Temple University, 1115 Polett Walk, 309 Gladfelter Hall, Philadelphia, PA 19122, USA
Email: jmennis@temple.edu

[2]Department of Psychiatry, Virginia Commonwealth University, 515 North 10th Street, P.O.Box 980489, Richmond, VA 23298-0489 USA
Email: michael.mason@vcuhealth.org

Abstract

Place can be understood, and represented, not only as an attribute of a location but also as the emotional attachments that characterize a relationship between an individual and a location. We demonstrate the development and application of this principle in place-based geographic information systems (GIS) through the use of georeferenced ecological momentary assessment (EMA), an approach for gathering real-time, in-situ data on individuals' mood states, behaviors, and social interactions via brief surveys delivered on a mobile phone. As a case study, we focus on the representation and analysis of the effect of activity space exposure to vacant housing on perceived safety among a sample of 139 urban adolescents enrolled in a longitudinal, georeferenced EMA study.

1. Introduction

Researchers have distinguished place-based geographic information systems (GIS) as emphasizing subjective interpretations of locations (Elwood *et al*. 2013). It has been suggested that geographic information scientists should look to research in cognitive science and humanistic geography for inspiration to developing representational and analytical approaches to place-based GIS (Winter and Freksa 2012). Such a perspective has informed research on extracting place information from linguistic descriptions, formalizing computational notions of place, georeferencing place data, and reasoning with place data (Gao *et al*. 2013; Scheider and Janowicz 2014; Vasardani *et al*. 2013). Health behavior is a domain that is particularly appropriate for place-based GIS (Goodchild 2015), as environmental influences on health behaviors are often moderated by the characteristics of the individual, including not only group characteristics such as race, age, and gender, but also the unique prior life experiences each individual carries with them (Mennis and Mason 2011). Thus, environmental influences on health behaviour can be interpreted to occur through subjective constructions of place.

Drawing from research in humanistic geography that posits place-making as construed through personal experience (Tuan 1977), we argue here that place-based GIS can represent place not only as a property of a location (as with conventional GIS), but also as a relationship between an individual and a location – a relationship that is molded by the unique characteristics and experiences of the individual (Mennis *et al*. 2013). We demonstrate the application of these principles using georeferenced ecological momentary assessment (EMA), an approach for gathering real-time, in-situ data on individuals' mood states, behaviors, and social interactions via brief surveys delivered on a mobile phone. EMA combined with global positioning systems (GPS) for georeferencing enables the collection of integrated longitudinal survey and activity space data (Epstein *et al*. 2014; Mason *et al*. 2015). As a case study, we focus on the representation and analysis of place-based perceptions of safety among a sample of 139 adolescents enrolled in a longitudinal, georeferenced EMA study of substance use in Richmond, Virginia.

2. Case Study Sample and Research Question

The sample is composed of 13 and 14 year old adolescents recruited between 2012 and 2014, primarily from an adolescent medicine outpatient clinic, and followed for one year. Subjects received georeferenced EMA surveys 4-6 times per day over a four day period every other month, yielding 1,629 georeferenced EMA responses with no missing data. These EMA responses include only those taken outside the home, e.g. at a friend's house, at a park, etc. Here, we investigate whether an environmental variable, the percentage of vacant housing within a U.S. Census tract, influences the perception of safety at that activity space location, measured according to the EMA item "How safe are you right now?" with responses given on a 1 ("Not at all safe") to 9 ("Very safe") scale.

Drawing on theories of neighborhood disorder we hypothesize that higher vacant housing is associated with lower perceived safety. We also hypothesize that this association will differ according to characteristics of the individual, both in terms of gender and the percent vacant housing that occurs at the subject's home neighborhood, as we consider that an individual's personal experience with vacant housing nearby their residence will affect their emotional response to vacant housing in their activity space away from home. To this end, we condition the EMA activity space location percent vacant housing variable on the percent vacant housing value that occurs at each subject's home tract by subtracting the latter from the former, a variable we call the 'relative percent vacant housing.' Positive values of relative percent vacant housing indicate a subject traveling to an activity space location with a higher level of vacant housing as compared to their home tract.

3. Conceptual Representation

The basic conceptual model for data representation is shown in Figure 1. There are three entities: subjects, locations, and places, the last of which is defined as time-varying relationships between subjects and locations. Subjects have attributes which do not change, collected during a full battery assessment at baseline, which include attributes such as gender, age at enrolment, and race. Location in this case study is represented by a tessellation of U.S. census tracts with associated attributes, including percent vacant housing, which is assumed to be constant over the one year span of the study. Place is represented as the relationship between a subject and location as captured by an EMA observation at particular time, when a subject is at a particular location, for example the perception of safety an individual has at a particular activity space location when completing an EMA survey.

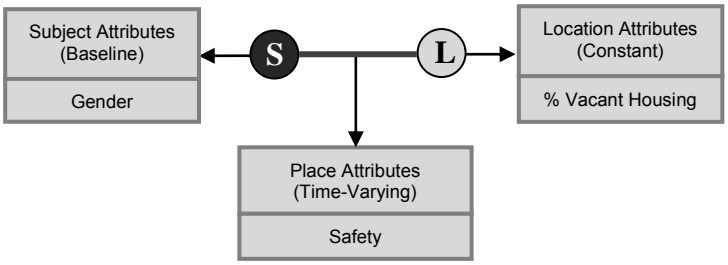


Figure 1. Conceptual model of subject (S), location (L), and place data.

4. Case Study Analysis and Results

To test the hypothesis that higher relative percent vacant housing is associated with lower perception of safety we employ Generalized Estimating Equations (GEE), which allows us to control for effects at the tract, subject, and survey wave levels, while also controlling for age, race, and gender. We then test whether the effect of relative percent vacant housing on safety

differs between boys and girls by entering an interaction term. Results are presented in Table 1, where Model 1 shows that higher relative percent vacant housing is associated with lower perceived safety (β=-0.015, p<0.05). Model 2 shows that this association is moderated by gender (β=-0.051, p<0.005), as is illustrated in Figure 1, which shows the relationship between relative percent vacant housing and perceived safety for boys versus girls. Clearly, the effect of exposure to vacant housing on perceived safety is stronger for boys as compared to girls, for whom the effect is near zero. We speculate that this is due to the greater propensity for violence among boys and the greater risk of violence, or of other illicit activities, in communities with high levels of disorder.

Table 1. Results of GEE Models of Perceived Safety (controlling for age and race; coefficients reported; *p<0.05, ***p<0.005).

Year	Model 1	Model 2
Female	0.174	0.163
Rel. % Vac	-0.015*	-0.049***
Fem. * Rel. % Vac		0.051***
Intercept	7.365***	7.376***

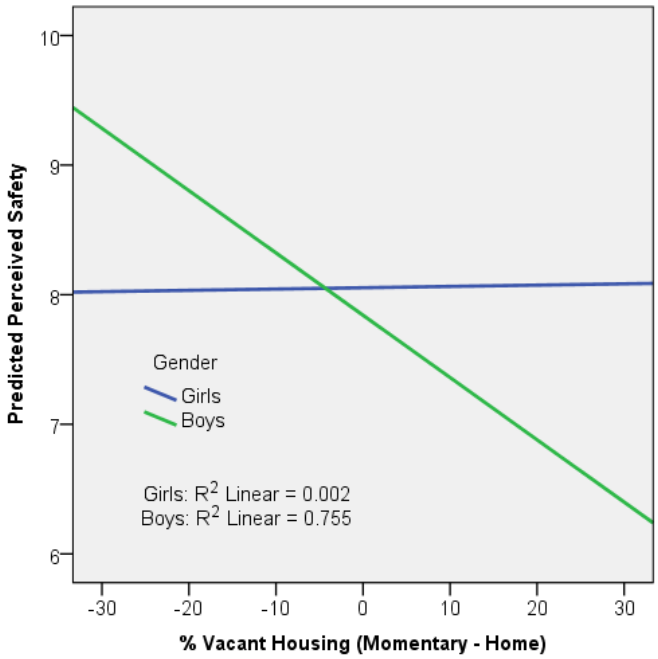


Figure 2. Gender moderates the effect of relative percent vacant housing on perceived safety.

5. Conclusion

This research demonstrates an approach to place-based GIS that differs from previous approaches in terms of how place is conceptualized and encoded, and in how place may be operationalized analytically. Here, we emphasize that place characteristics can be viewed as subjective and ambiguously defined interpretations of one's environment, and can thus be represented as a relationship between an individual and a location. We explicitly distinguish

between objectively observable attributes of a location, e.g. the percent vacant housing in a tract, and the emotional interpretation of that location at a given moment in time, e.g. the sense of safety an adolescent has at that location. This approach thus agrees with research that posits place-attachment as a relationship between an individual's perceptions and the qualities of a location (Brown *et al*. 2015).

Further, by incorporating gender as a moderator, and by conditioning the momentary exposure to vacant housing on the exposure experienced in an adolescent's home tract, we demonstrate that modeling place characteristics on individual outcomes can benefit by incorporating characteristics of the individual, including prior experiences. We believe this approach has the potential to better incorporate a humanistic geography perspective into place-based GIS for health applications, where place-based models of environmental effects on health behaviors can more effectively capture how people subjectively relate to places, as compared to more common analytical approaches where an objectively observable environmental condition is hypothesized to exert some effect on an entire population regardless of individual differences and personal experiences. We note that although we focus here on a relatively simple example of modeling the effect of exposure to vacant housing on the perception of safety, the same approach can easily be extended to address a variety of more complex environmental conditions and affective and behavioural outcomes.

Acknowledgements

This research was supported by grant No. 1R01 DA031724-01A1 from the National Institutes of Health. The findings and conclusions are those of the authors and do not necessarily represent the views of the National Institutes of Health.

References

Brown G, Raymond CM, Corcoran J, 2015. Mapping and measuring place attachment. *Applied Geography*, 57: 42-53.

Elwood S, Goodchild, M, and Sui, D, 2013, Prospects for VGI research and the emerging fourth paradigm. In: Sarah Elwood, Michael F Goodchild and Daniel Sui (eds), *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) and Theory and Practice*, Springer, London.

Epstein DH, Tyburski M, Craig IM, Phillips KA, Jobes ML, Vahabzadeh M, Mezghanni M, Lin J-L, Furr-Holden CDM, Preston KL, 2014, Real-time tracking of neighborhood surroundings and mood in urban drug misusers: Application of a new method to study behavior in its geographical context. *Drug and Alcohol Dependence*, 134: 22-29.

Gao S, Janowicz K, McKenzie G, and Li L, 2013, Toward palatial joins and buffers in place-based GIS. In: *Proceedings of SIGSPATIAL '13: 21st SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Orlando, FL, USA, 42-49.

Goodchild MF, 2015, Space, place and health. *Annals of GIS*, 21: 97-100.

Mason M, Mennis J, Way, T, Light, J, Rusby, J, Westling, E, Crewe, S, Flay, B, Campbell, L, Zaharakis, NM, and McHenry, C, 2015, Young urban adolescents' activity space, peers, and substance use. *Health and Place*, 34: 143-149.

Mennis J and Mason MJ, 2011. People, places, and adolescent substance use: Integrating activity space and social network data for analyzing health behavior. *Annals of the Association of American Geographers*, 101: 272-291.

Mennis J, Mason MJ, and Cao Y, 2013, Qualitative GIS and the visualization of narrative activity space data. *International Journal of Geographical Information Science*, 27(2): 267-291.

Scheider S and Janowicz K, 2014. Place reference systems: A constructive activity model of reference to places. *Journal of Applied Ontology*, 9: 97-127.

Tuan Y-F, 1977, *Space and Place: The Perspective of Experience*, University of Minnesota Press, Minneapolis, MN.

Vasardani M, Winter S, and Richter K-F, 2013. Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27: 2509-2532.

Winter S and Freksa C, 2012, Approaching the notion of place by contrast. *Journal of Spatial Information Science*, 5: 31-50.

Shape and resolution: quantifying feature morphology change due to coarsening spatial resolution using UAV-based images from vertical transects

S. W. Mitchell[1], T. K. Remmel[2]

[1]Department of Geography & Environmental Studies, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6
Email: Scott.Mitchell@carleton.ca

[2]Department of Geography, York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3
Email: remmelt@yorku.ca

Abstract

Aside from tone or colour, shape is one of the most readily utilized characteristics to interpret remotely sensed imagery. As spatial resolution is coarsened, the level of observable spatial detail diminishes, thus the shapes of features become less distinct. This study controlled the observation platform, a Tetracam ADC snap multispectral camera mounted on a Phantom 2 quad-copter, to collect images along vertical transects. This design simulates continuously coarsening spatial resolution as the aircraft ascends, while maintaining consistent imaging parameters. Vegetation visible on all scenes was analysed using common shape characterizations. Scaling relationships were assessed between original shape complexity and spatial resolution (a proxy for scale). Initial results show strong scaling relationships at ultra-fine scales but an interesting area of stability which, if repeatable, could have important implications for the use of UAVs in environmental research.

1. Introduction

The characterization and comparison of spatial pattern characteristics or landscape morphological elements are sectors of GIScience that intersect with mathematics and landscape ecology, and have been developing for several decades (Gardner *et al*. 1987; Haines-Young and Chopping 1996; Uuemaa 2009). Numerous studies span the breadth of such inquiries, ranging from studies of land cover change (Linke *et al*. 2009) to assessing habitat suitability (McAlpine and Eyre 2002). While computing a suite of metrics is facilitated by readily accessible software (Baker and Cai 1992; McGarigal and Marks 1995; Vogt *et al*. 2007), many redundancies have been identified (Riitters *et al*. 1995) and limitations explicitly characterized that limit the inference that can be drawn from such metrics (Remmel and Csillag 2003). Scale effects have been tested by manipulating landscape extent and spatial resolution, but these generally rely on resampling or the use of multiple platforms of data (e.g. Saura 2002; Baldwin *et al*. 2004).

This study focuses on the characterization of planar shape, a single aspect of the broader collection of pattern metrics, with respect to scale. Shape provides the ability to characterize and discriminate individual landscape elements (Zhang and Atkinson 2016), rather than summarize landscape-wide patterns. Scaling analysis was applied to data obtained from a consistent platform, such that resampling or sensor substitution did not introduce spurious relationships and uncertainty that make interpretation more complex. By manipulating a single variable (spatial resolution) of a digital image while maintaining all camera, timing, and environmental conditions constant, we explored the effect of spatial resolution on the quantification of patch shape directly. This differs from Saura and Carballal (2004) where

scale was kept consistent and multiple indices were applied to measure shape (at the class level) to assess the indices' effectiveness.

Our research question asks: does spatial resolution have a significant effect on measuring patch shape? Obviously there is an effect once spatial resolution becomes coarse relative to the local detail being depicted, but we seek to characterize the functional relationship between spatial resolution and shape measurement for selected phenomena captured on images. We hypothesize that knowing the form of this function will allow improved selection of spatial resolutions for the detection, mapping, and analysis of shapes in landscapes.

2. Methods

The study was conducted at an urban parkland meadow consisting of grasses, shrubs, and scattered trees (approximate centroid: 45° 24' N, 75° 45' W) in Ottawa, Ontario, Canada. The site permitted easy access and an ability to perform field validation to ensure feature identification accuracy. The species and communities were not of interest as much as the ability to observe features at the ground level and also from airborne perspectives.

Data were acquired with a Tetracam ADC Snap multispectral camera sensitive to the green, red, and near infrared regions of the electromagnetic spectrum. The 90 g camera was nadir mounted to a Phantom 2 quad-copter unmanned aerial vehicle (UAV), and powered by the UAV battery (image at http://bit.ly/2aLycMl). A benefit of the fixed-rotor UAV is that it can be flown directly upwards by providing additional upward thrust. This allows vertical transects, rather than just horizontal, allowing us to manipulate spatial resolution while repeatedly imaging a common ground area. The flight platform is customized with a ground pilot display (iOSD mini) showing flight telemetry along with a real-time view of the image being acquired by the multispectral camera. The telemetry allowed us to position the flight platform directly above the object of interest and ensured that we acquired imagery at desired altitudes. We set the imaging interval at 3 seconds and controlled the rate of ascent such that we obtained multiple views of the same ground feature.

Prior to flying the vertical transects over prominent features in the study area, a series of polished metal salad bowls were placed upside-down in the field to serve as ground control points. Since the bowls did not move and were visible in each image, they were used to simplify the relative geo-registration of images centred over the feature of interest. The range of altitudes traversed resulted in images having spatial resolutions ranging from 5.6 mm (at 10 m altitude) through 5.1 cm (at 90 m altitude); the coarse maximum corresponds to the legal flying height restriction, not the limits of the imaging system.

In this initial study, we concentrated on a single feature of interest (shrubs), binary-classified (ISODATA clustering, 16 initial clusters, background image manually aggregated), at the centre of the field of view for each vertical transect. Here we present the results from the first vertical transect we have analysed, providing a suite of metrics that characterize the shape of the feature: area, perimeter, and corrected perimeter-to-area (cPA) ratio (relates the shape's perimeter:area to that of a perfect circle; Farina 2006: 321). The results are plotted relative to the spatial resolution to obtain functions relating scale to shape.

3. Results

Scaling relationships were markedly different at coarser vs. finer resolutions. At ultra-fine resolutions (less than 25 mm) there were strong scaling effects on the shape metrics (Fig. 1; full tabular results including image resolutions at http://bit.ly/2aMnHra). At elevations close to the ground (below 40 m altitude), and therefore the highest image resolutions, the detected shrub perimeter followed a fairly well defined linear scaling relationship, with higher perimeter detected at finer resolutions; however, the area of the

detected shrub fluctuated as contiguous portions of the shrub canopy as detected by ISODATA clustering became detached and re-attached (Fig. 2). As a result, the cPA shape index also had a linear trend towards higher ratios at finer resolutions, but with progressively increasing scatter. Once altitude reached about 40 m, with spatial image resolution increasing ~25 - ~55 mm, this scaling relationship levelled off. For at least the aspects of shape studied here, there appears to be no scaling relationship within that range of resolutions.

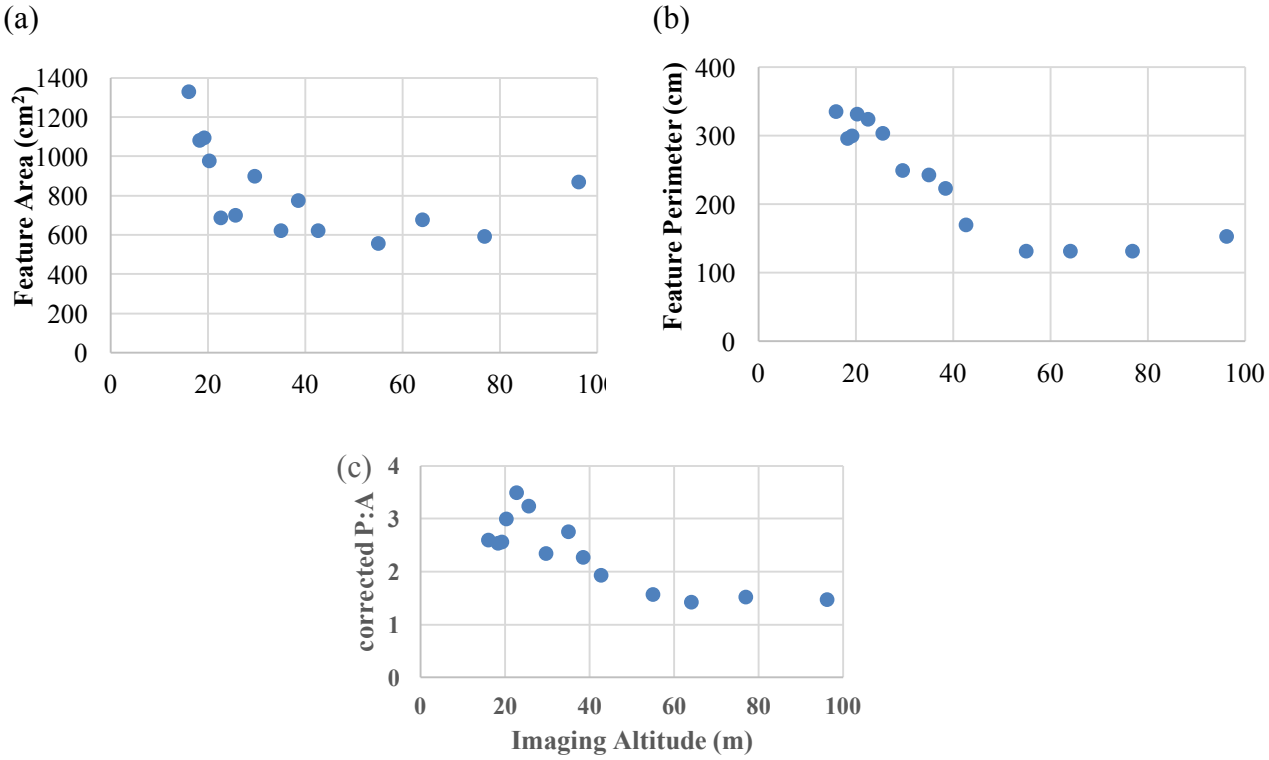


Figure 1 (a) shrub feature area, (b) perimeter, and (c) cPA, with respect to imaging altitude along a vertical transect.

Figure 2. Shrub representations at different spatial resolutions ranging from 9.13 mm (a) through 54.75 mm (n).

4. Discussion and Conclusions

Although these are clearly preliminary results, given the novel yet increasingly feasible range of applications provided by the rapidly developing UAV industry, we wish to rapidly disseminate these observations to help spur complementary research. The sharp changes in

the observed scaling relationships, and particularly the observation that shape descriptors remained stable through a range of resolutions, could have strong implications. Through spatial resolutions of ~9 to ~25 mm, the scaling behaviour of the shrub's perimeter is reminiscent of the infinite coastline example often used to illustrate fractal theory (Mandelbrot 1967). However, across a fairly large range (in terms of flight planning) of spatial resolutions (25-55 mm), there was little change in the calculated metrics.

The generality of these observations needs to be tested, both across other image objects within our study site, and in other contexts (are the slopes and asymptotes of the scaling relationships consistent?). We plan to repeat the analysis across other plant objects in the imagery to develop confidence intervals for our scaling relationships, and to compute *ShrinkShape* decompositions (Remmel 2015). The decomposition will conduct iterative shrinking of planar shapes interspersed with the measurement of the remaining area and perimeter until the shape becomes extinct due to shrinking, and extract MSPA morphological element summaries (Vogt *et al*. 2007). Conducting similar analyses using other physical environments, platforms, and extracted measures will help us understand the potential for exploiting the much higher, and controllable, spatial resolutions offered by UAV technology.

Acknowledgements

This work was funded by York University, Carleton University, and J. D. Barnes Ltd. (via a Mitacs Accelerate Grant).

References

Baldwin DJB, Weaver K, Schnekenburder F and Perera AH, 2004, Sensitivity of landscape pattern indices to input data characteristics on real landscapes: implications for their use in natural disturbance emulation, *Landscape Ecology* 19:255-271.

Farina A, 2006, Principles and methods in landscape ecology: toward a science of landscape. Springer, Dordrecht, The Netherlands.

Gardner RH, Milne BT, Turner MG and O'Neill RV, 1987, Neutral models for the analysis of broad-scale landscape pattern. *Landscape Ecology*, 1:19-28.

Haines-Young R and Chopping M, 1996, Quantifying landscape structure: a review of landscape indices and their application to forested landscapes, *Progress in Physical Geography* 20(4):418-445.

Linke J, McDermid GJ, Pape AD, McLane AJ, Laskin DN, Hall-Beyer M and Franklin SE, 2009, The influence of patch-delineation mismatches on multi-temporal landscape pattern analysis, *Landscape Eco*. 24:157-170.

Mandelbrot B, 1967, How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension, *Science* 156(3775): 636-638.

McAlpine CA and Eyre TJ, 2002, Testing landscape metrics as indicators of habitat loss and fragmentation in continuous eucalypt forests (Queensland, Australia), *Landscape Ecology* 17(8):711-728.

McGarigal K and Marks BJ, 1995, *FRAGSTATS: spatial pattern analysis program for quantifying landscape structure*. Gen. Tech. Report PNW-GTR-351, USDA Forest Service, Pacific Northwest Research Station.

Remmel TK and Csillag F, 2003, When are two landscape pattern indices significantly different? *Journal of Geographical Systems* 5(4):331-351.

Remmel TK, 2015, *ShrinkShape2*: a FOSS toolbox for computing rotation-invariant shape spectra for characterizing and comparing polygons. *The Canadian Geographer* 59(4):532-547.

Riitters KH, O'Neill RV, Hunsaker CT, Wickham JD, Yankee DH, Timmins SP, Jones KB and Jackson BL, 1995, A factor analysis of landscape pattern and structure metrics, *Landscape Ecology* 10(1):23-39.

Saura S, 2002, Effects of minimum mapping unit on land cover data spatial configuration and composition. *International Journal of Remote Sensing* 23(22):4853-4880.

Saura S and Carballal P, 2004, Discrimination of native and exotic forest patterns through shape irregularity indices: and analysis in the landscapes of Galicia, Spain, *Landscape Ecology* 19:647-662.

Uuemaa E, Antrop M, Roosaare J, Marja R and Mander Ü, 2009, Landscape metrics and indices: an overview of their use in landscape research, *Living Review Landscape Research* 3:1-28.

Vogt P, Riitters KH, Estreguil C, Kozak J, Wade TG and Wickham JD, 2007, Mapping spatial patterns with morphological image processing. *Landscape Ecology* 22:171-177.

Zhang C and Atkinson PM, 2016, Novel shape indices for vector landscape pattern analysis, *International Journal of Geographical Information Science* 10.1080/13658816.2016.1179313

Spatializing Global Urban Extent:

A Source Driven Approach

J. J. Moehl[1], A. N. Rose[1], E. A. Bright[1]

[1]Oak Ridge National Laboratory, Oak Ridge, TN 37831
Email: moehljj;rosean;brightea@ornl.gov

Abstract

"Urban" is something that intuitively feels very well defined; however, when it comes time to express this idea on a map, things get complicated. Definitions of "urban" vary globally, and as such there is not universal understanding of what makes a given place "urban". The common approach to spatially defining urban extents is through remotely sensed imagery. The alternative approach presented in this paper uses the percent of population in urban areas, which is a common macroeconomic (country-level) variable with the definition for urban generally defined by each country's statistical office, along with temporally-aligned high-resolution population data to spatially define urban extents for each country. Because the percent urban number is defined by the same producer as other urban/rural defined statistical data, such as household characteristics or birth rate, understanding the spatial aspect of urban from the same perspective is ideal for high resolution spatial modeling of these other phenomena.

1. Introduction

At Oak Ridge National Laboratory, we've been modeling population for nearly two decades. Whether using a top down approach such as the LandScan Global model (Dobson *et al*. 2000), or a hybrid top down/bottom up approach as is used to develop LandScan USA (Bhaduri *et al*. 2007), spatial data that aligns with tabular datasets is essential. Ensuring that national and subnational boundaries match the data from statistical organizations is extremely important for dasymetric mapping, as these boundaries play a large role in appropriately disaggregating zonal summaries. Many summary data values, such as population counts, birth rates, average household size, and age distributions are listed not only by administrative zone, but often times they are further listed by whether they occur in urban or rural areas. Spatial data for administrative boundaries are widely available, although of varying quality and level of detail. However, very rarely are spatial data available with urban area delineation from the perspective of the statistical organization.

Summary data designated as urban or rural for each administrative zone within the country is common. The idea of what is "urban" is defined and implemented by each country's statistical office. The UN's *World Urbanization Prospects: The 2014 Revision* provides an exhaustive comparison of these definitions which vary widely in their considerations (UN 2014). This may be an extremely detailed formal definition like the one provided by the US Census Bureau (US Census 2011) complete with spatial data freely available in multiple formats, or very generally defined with no publicly accessible spatial data. Administrative data and population density are the most common criteria used by countries to fit the urban definition, with 125 and 137 countries respectively using each criterion in part or whole (UN 2014). The HYDE datasets, which provide historical population and land use estimates, combine historical urban population estimates with population density in their models; however, these datasets are historical and at a coarser ~10 sq km resolution (Klein Goldewijk 2001, 2005).

2. Method

Our motivation for this study stems from a project where we are trying to model housing characteristics at a fine spatial resolution. These data are subnational and also include an urban/rural tag. This urban vs. rural designation often captures the variation in housing characteristics within the subnational zones. As such, having an urban extent is extremely important. We began exploring sources of spatial data for urban extent with global coverage. Potere and Schneider (2007) provide a useful comparison of these global urban area maps; the global urban area estimates ranged from 276,000 to 3.5 million sq. km. with variation coming from temporal, sensor, and methodological differences. While these datasets have been used to understand past and future urban expansion (Seto 2011), we felt these datasets were inadequate for our purpose. Our data summaries are provided by national statistical organizations whose decisions about which areas are urban generally have little to do with interpretations of remotely sensed data used to describe urban land cover. Thus, in many ways, it is irrelevant whether a particular spot in any given country could spectrally be described as urban.

Our goal is to accurately represent the urban area described by the summary data from the perspective of the producer of that statistical data. To model that extent, we combined other information from that producer, namely the percent of population living in urban areas, with high-resolution population data, LandScan Global. We are assuming that urban populations live in the most densely populated places and we otherwise know population density is often a factor in defining urban (UN 2014). With this scenario, it is possible to triangulate locally adaptive density thresholds for "urban" by summarizing the population, cell by cell, in each country from the most densely populated cell to the least. We use this population sum to determine the cumulative percentage of urban population at or above each density value. This percentage is then matched with the percent urban population provided by statistical organizations to determine the "break point" dividing urban and rural for each country. We use the break point to classify cells with greater or equal densities as urban for each country.

The first requirement for this process is a population dataset with high spatial resolution as well as global coverage. The LandScan Global dataset provides the granularity needed to calculate the most precise urban extent estimate possible. We developed an area grid of the same extent and resolution as the LandScan data in order to calculate population density. The next step in this process is to acquire urban population percentage data for the same time period as the LandScan data, 2014. There are several sources of this data. One option is to acquire this data from each national statistical organization. However, because this would largely repeat the processes employed by other organization whose mission is to acquire and assimilate such data, we feel it is sufficient to use data from these providers. Two producers, the CIA World Factbook (CIA 2016) and the United Nations Population Division (UN 2014), use similar methods; each acquires data from statistical organizations and adjusts them when necessary, for example when data are of poor quality. Figure 1 shows the CIA World Factbook versus the UN break point values; they are very similar and either is sufficient for our purposes. More importantly, Figure 1 shows the variation in the density break values. This illustrates the fact that "urban" means very different things statistically across the globe.

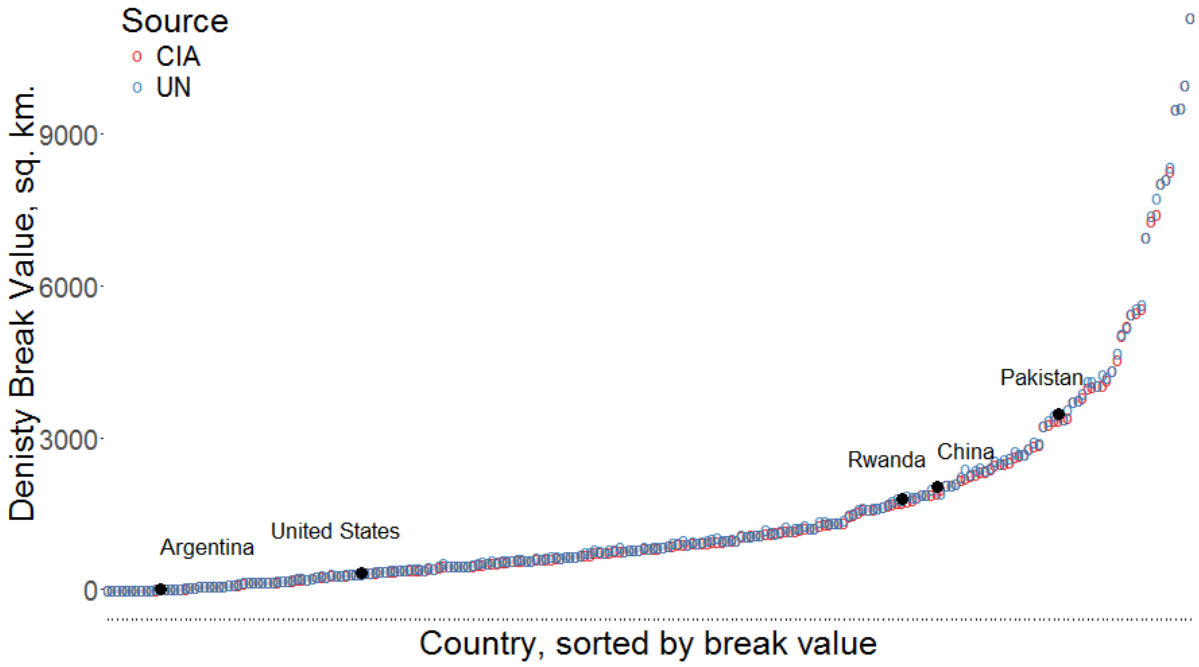


Figure 1. The population density value dividing urban and rural in our new method varies across the globe. Also, the CIA World Factbook and the UN Population Division data are extremely similar.

3. Results

Next we compared our urban extent with the GRUMP v1 urban extent layer (CIESIN 2011). GRUMP relies largely on nighttime lights and has been shown to estimate a larger urban area extent than other datasets (Potere and Schneider 2007). To calculate the total urban extent in GRUMP we used the same country boundaries and the same area grid and summarised the urban area in each country. Figure 2 shows a comparison of our new method and the GRUMP urban area extent. Each point in the chart represents the percentage of the total area, which is the sum of the GRUMP estimate and our method, accounted for by our new method. This allows for a relative examination of each country. The line at 0.5 means each country had 50% of the total, so they have equal area totals for the country. A break at 0.33 or 0.66 means one source has twice as much urban area in that country. The size of the points shows the urban area in sq. km. according to the new method. While most countries have much more urban area in GRUMP, there is no regional pattern to the variation, which shows the great difference in methodologies defining "urban" at the country level. For example, we estimate a much greater urban extent in Argentina than GRUMP due to the definition of urban in Argentina being "Localities with 2,000 inhabitants or more" (UN 2014). Figure 3 provides an example of the differences in spatial distributions between these models.

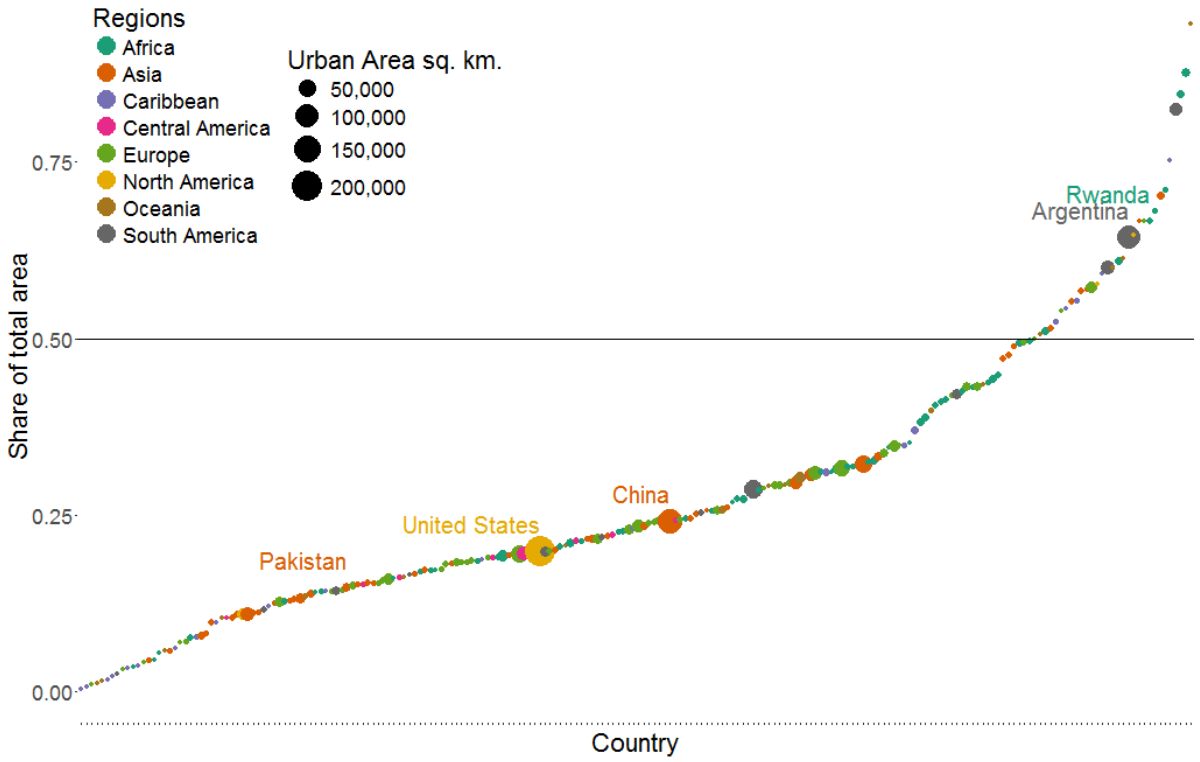


Figure 2: The new method urban area estimate shown as a percentage of the total area for the GRUMP estimate and the new estimate. Values above the line show where the new method estimates more urban area than GRUMP.

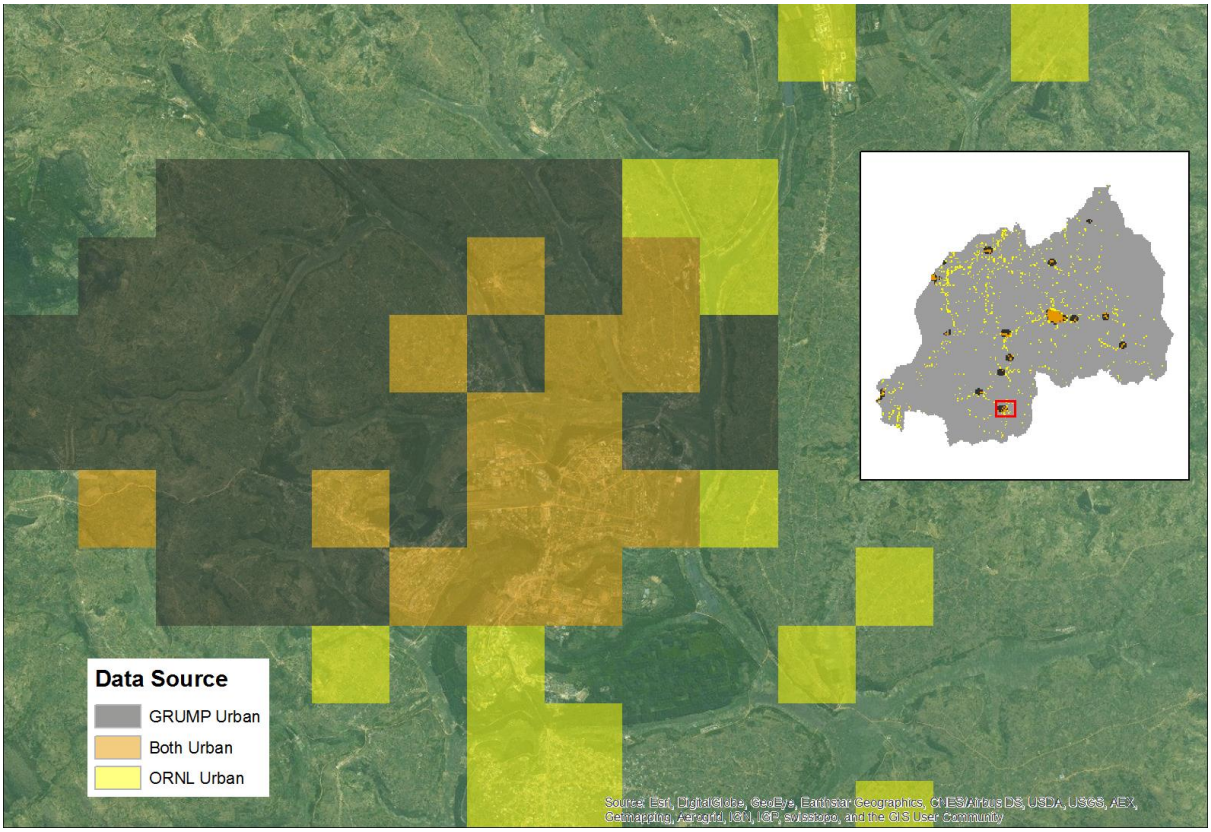


Figure 3: Close up and country level (inset) comparison of ORNL urban extent and GRUMP for Rwanda show how differences in methodology drive varied spatial distributions of urban and thus result in large differences in overall urban area between the models.

4. Conclusion

We conclude based on these results that it is appropriate to use data from the perspective of the statistical organization to calculate urban extent rather than to rely solely on remotely sensed imaging sources of "urban" data when trying to spatialize other phenomena from the same statistical organization. Further research is needed for a more in-depth comparison of urban extent as finer, city-level scales to fully test the veracity of our model. Additionally, explicitly including other data such as administrative boundary data may increase the accuracy of our model as these are often considered when defining "urban" at the local level.

Copyright

References

US Census Bureau, 2011. *Urban Area Criteria for the 2010 Census*. Department of Commerce, 76. *Federal Register* 164 (2011); 53030–53043.

Center for International Earth Science Information Network - CIESIN - Columbia University, International Food Policy Research Institute - IFPRI, The World Bank, and Centro Internacional de Agricultura Tropical - CIAT. 2011. Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Urban Extents Grid. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4GH9FVG. Accessed 28 April 2016.

Bhaduri, B., E. Bright, P. Coleman and M. Urban, 2007. LandScan USA: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal* 69(1): 103-117.

Bright, E. A., Rose, A. N., & Urban, M. L. (2015). *LandScan 2014* [digital raster data]. Retrieved from: http://www.ornl.gov/landscan/

Central Intelligence Agency, 2016. The World Factbook 2015-2016. Washington, DC: Central Intelligence Agency, 2016.

Dobson, J., Bright, E., Coleman, P., Durfee, R., & Worley, B., 2000. LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing*, 66(7): 849-857.

Klein Goldewijk, K., 2001. Estimating global land use change over the past 300 years: the HYDE database. *Global Biogeochemical Cycles*. 15(2): 417-434.

Klein Goldewijk, K. 2005. Three centuries of global population growth: A spatial referenced population density database for 1700 – 2000. *Population and Environment* 26(5): 343-367.

Potere, D., and Schneider, A., 2007, A critical look at representations of urban areas in global maps. *GeoJournal* 69(1):55-80.

Seto, K., Fragkias, M., Güneralp, B., Reilly, MK., 2011. A Meta-Analysis of Global Urban Land Expansion. *PLoS ONE* 6(8): e23777.

United Nations, Department of Economic and Social Affairs, Population Division, 2014. World Urbanization Prospects: The 2014 Revision, CD-ROM Edition.

A Dasymetric-Based Monte Carlo Simulation Approach to the Probabilistic Analysis of Spatial Variables

April Morton[1], Jesse Piburn[1], Ryan McManamay[1], Nicholas Nagle[2], Robert N. Stewart[1]

[1] Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831
Email: {mortonam; piburnjo; mcmanamayra; stewartrn}@ornl.gov

[2] University of Tennessee, Knoxville, Department of Geography, 1000 Phillip Fulmer Way, Knoxville, TN 37916
Email: nnagle@utk.edu

Abstract

Monte Carlo simulation is a popular numerical experimentation technique used in a range of scientific fields to obtain the statistics of unknown random output variables. Though Monte Carlo simulation is a powerful technique for the probabilistic understanding of many processes, it can only be applied if it is possible to infer the probability distributions describing the required input variables. This is particularly challenging when the input probability distributions are related to population counts unknown at desired spatial resolutions. To overcome this challenge, we propose a framework that uses a dasymetric model to infer the probability distributions needed for a specific class of Monte Carlo simulations dependent on population counts.

1. Introduction

Monte Carlo simulation is a numerical experimentation technique that has been widely used in a variety of scientific domains to obtain the statistics of unknown random output variables by repeatedly sampling values from a set of known input random variables and then feeding them through a computational model (Mahadevan 1997). Dasymetric mapping, on the other hand, has been widely used in the field of areal interpolation to disaggregate coarse resolution population data to a finer resolution through the use of ancillary data (Eicher and Brewer 2001).

Though Monte Carlo simulation is a powerful technique for the probabilistic understanding of many processes, it can only be applied if the probability distributions describing the required input variables can be inferred. Unfortunately, conventional inference methods cannot often be used to infer the probability distributions of population counts (i.e. counts of populations with specific characteristics) that are unknown at desired spatial resolutions. Fortunately, recent advancements in dasymetric mapping, which may not be well known to researchers utilizing Monte Carlo simulation in fields other than areal interpolation, provide novel methods for estimating the probability distributions of population counts. To highlight the potential link between dasymetric mapping and Monte Carlo simulation, we propose a framework that uses the penalized maximum entropy dasymetric model (PMEDM) proposed by Nagle *et. al* (2014) to learn the parameters of multinomial distributions describing population counts needed to complete a specific class of Monte Carlo simulations.

2. Methodology

Suppose we'd like to calculate, through Monte Carlo simulation, the sample mean $\bar{y}_t$ and sample standard deviation s_{y_t} of an output variable $y_t = f_t(a_t, x_t)$ for a set of non-overlapping regions $t \in \{1, \dots, T\}$ where $a_t = [a_{t1,\dots,}a_{tk}]$ and $x_t = [x_{t1,\dots,}x_{tk}]$ are vectors of random variables with

unknown and known probability distributions, respectively. Furthermore, assume a_{ti} represents the number of people or households with characteristic i in region t and assume x_{ti} represents some value conditioned on characteristic i in region t. For example, a_{ti} might represent the number of one bedroom households in region t while x_{ti} might represent the electricity consumption of a one bedroom household in region t.

Now, suppose that there also exist microdata, related to each of the k characteristics, for a given population survey containing n of N respondents sampled from region s, where region s is partitioned into the same $t \in \{1, \dots, T\}$ target regions. Furthermore, assume there exist summary count estimates and variances corresponding to each of the T target regions and k characteristics, and assume we know the prior probabilities q_{jt}, for all $j \in \{1, \dots, n\}$ and $t \in \{1, \dots, T\}$, that a person or household with the same k characteristics as respondent j lives in target region t. Given the preceding information, we can use the PMEDM to learn the actual probabilities p_{jt}, for all $j \in \{1, \dots, n\}$ and $t \in \{1, \dots, T\}$, that a person or household with the same k characteristics as respondent j lives in target region t. We can then simulate, for all j and t, several likely counts of each person j in target region t, from which we can compute several realizations of a_{ti}, or the total people in region t with characteristic i. Given these realizations of a_{ti}, we can complete the Monte Carlo simulation and compute the statistics of interest $\bar{y}_t$ and s_{y_t} to enhance our probabilistic understanding of the output variable y_t.

3. Application and Results

To illustrate the utility of the proposed framework, we use Monte Carlo simulation and the PMEDM to estimate the mean and standard deviation of the average aggregate monthly electricity consumption for all Census block groups t intersecting the Knoxville urbanized area defined by the Census in 2012 (Census 2012). More specifically, we compute the sample mean $\bar{y}_t$ and sample standard deviation s_{y_t}, for all t, of the average aggregate monthly electricity consumption given by

$$y_t = f_t(a_t, x_t) = \sum_{z=1}^{a_{t1r}} x_{t1z} + \sum_{z=1}^{a_{t2r}} x_{t2z} + \cdots + \sum_{z=1}^{a_{t8r}} x_{t8z} \quad (1)$$

where a_{tir} represents the r^{th} realization of the number of households with characteristic i in region t and x_{tiz} represents the z^{th} realization of the average monthly electricity consumption of a household with characteristic i. Out of the 8 characteristics i, the first 4 characteristics refer to the number of 1 through 4 or more bedroom detached houses in target region t while the last 4 characteristics represent the number of 1 through 4 or more bedroom non-detached houses in target region t. In this application "detached" house refers to all houses following the United States (US) Census' definition of "detached single-family housing units" and non-detached household refers to all other Census classifications for housing units (Census 2012). Also note that, due to limited sample sizes, studio apartments, or 0 bedroom houses, are grouped with 1 bedroom houses. Furthermore, from this point forward, the term "monthly electricity consumption" refers to the average monthly electricity consumption over a 12 month period.

3.1 Learning the Input Variable Probability Distributions

To learn the probability distributions of the random variables contained in a_t we collected all microdata variables, for all survey boundaries containing our study area block groups, matching the 8 categories defined above from the weighted 2008-2012 household-level Public Use Microdata Sample (PUMS) of the American Community Survey (ACS) (US Census 2012). Furthermore, we determined the summary count estimates and variances, related to the same characteristics, for all Census tracts and block groups, through the summary count estimates and 90% margins of error (MOEs) published in the 2008-2012 ACS summary tables (US Census 2012). In addition, we assumed each unique household had the same prior probability of belonging to each target region, and thus let $q_{jt} = \frac{w_j}{T \cdot \sum_{r=1}^{n} w_r}$, where w_j represents the weight of microdata respondent j. We then used the PMEDM to learn the probabilities p_{jt}, for all $j \in \{1, \ldots, n\}$ and $t \in \{1, \ldots, T\}$, from which we simulated several realizations of a_t.

To learn the probability distributions for the random variables in x_t we used the 2009 Residential Energy Consumption Survey (RECS) microdata, restricted to respondents living in Tennessee, published by the US Energy Information Administration (EIA) (Energy Information Administration 2009). More specifically, we assumed each of the 8 random variables x_{ti} followed a normal distribution and estimated the mean and standard deviation of the monthly electricity consumption of each of these categories using the annual kWh reported by Tennessee respondents belonging to each category.

3.2 Results and Discussion

To complete the Monte Carlo analysis we simulated, for all t, 30 sets of population counts for a_t and then computed 30 values of y_t, for each vector a_t, to obtain a total of 900 simulated values of each y_t. Figure 1 shows the mean monthly electricity consumption and standard deviation error bars for all household categories while figure 2 shows the median average monthly electricity consumption and standard deviation per household for all Census block groups intersecting the Knoxville urbanized area. As expected, the Census block groups closer to downtown Knoxville have a much lower median average consumption per household than the households in the wealthy suburban neighborhoods. This is likely due to the fact that the downtown block groups have a higher percentage of small apartments and student housing, which, according to figure 2, have a lower mean monthly electricity consumption than the wealthy suburban neighborhoods containing a higher percentage of large detached households. Though more difficult to interpret visually, the median average standard deviation per household within each block group varies according to the cumulative effect of the count and standard deviation of the mean electricity consumption coming from the mix of categories within each block group.

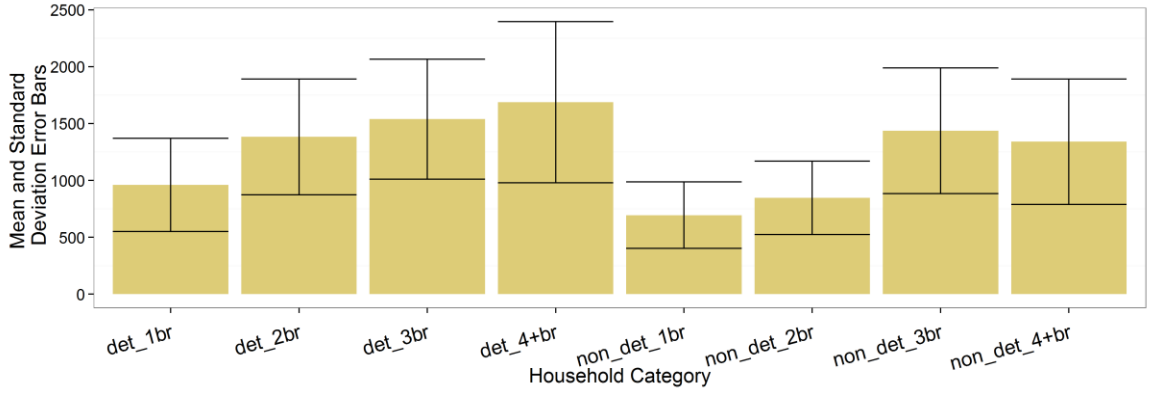


Figure 1. Mean monthly electricity consumption and standard deviation error bars (kWh) by household category

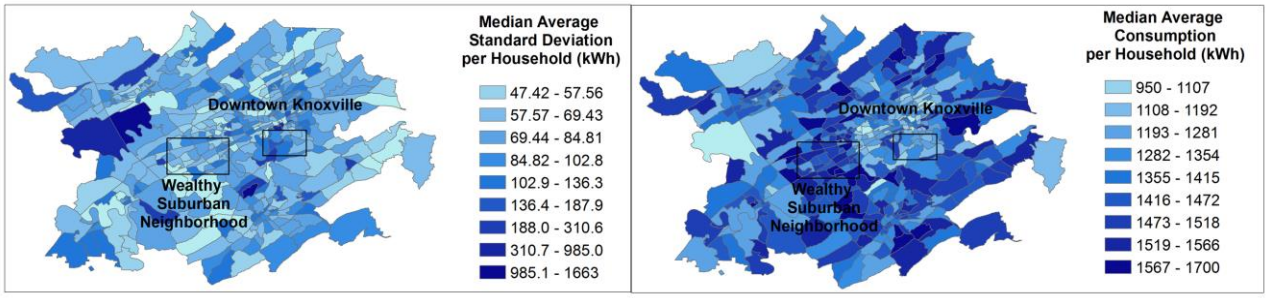


Figure 2. Median average monthly electricity consumption and standard deviation per household for all Census block groups intersecting the Knoxville urbanized area

In summary, this case study is one example of the potential usefulness of the proposed framework for completing Monte Carlo analyses which require probability distributions over population counts.

Copyright

Acknowledgements

The authors would like to acknowledge the financial support for this research from the U.S. Government for the development of a fine-resolution model of urban-energy systems' water footprint in river networks; Oak Ridge National Laboratory's Laboratory Directed Research and Development (LDRD).

References

Mahadevan, S (1997). Monte carlo simulation. Reliability-Based Mechanical Design: 123-146.

Eicher, C, Brewer C (2001). Dasymetric mapping and areal interpolation: Implementation and evaluation. Cartography and Geographic Information Science 28(2):125-138

Nagle, N, Buttenfield, B, Leyk, S, Spielman, S, (2014) Dasymetric modeling and uncertainty. Annals of the Association of American Geographers 104(1):80–95

US Census Bureau (2012), 2008 – 2012, American Community Survey microdata. US Census Bureau. http://factfinder2.Census.gov. Accessed 6 Jan 2015

US Census Bureau (2012), 2008 – 2012, American Community Survey summary tables. US Census Bureau. http://factfinder2.Census.gov. Accessed 6 Jan 2015

Energy Information Administration (2009) Residential Energy Consumption Survey. Energy Information Administration. http://www.eia.gov/consumption/residential/data/2009/index.cfm?view=microdata. Accessed 1 April 2016

TerraEx – a GeoWeb app for world-wide content-based search and distribution of elevation and landforms data

P. Netzel[1], J. Jasiewicz[1,2], T. F. Stepinski[1]

[1]Space Infromatics Lab, University of Cincinnati, 401 Braunstein Hall, Cincinnati, OH 45221-0131, US
Email: {netzelpl; stepintz}@mail.uc.edu

[2]Instititute of Geoecology and Geoinformation, Adam Mickiewicz University, Dziegielowa 27, Poznan, Poland
Email: jarekj@amu.edu.pl

Abstract

Terrain Explorer (TerraEx) is the first world-wide content-based search application for landscapes. Using 3" resolution world-wide DEM as an input it finds and displays, in the form of a similarity map, locations in the world where landscapes are similar to a user-selected query. TerraEx is a freely available, full service GeoWeb application. It also doubles as the most convenient distributor of global 3" DEM data, the global map of geomorphons, and the global map of terrain relief. TerraEx opens a possibility to utilize publicly available DEM archives in their entireties for world-wide exploration through content-based search.

1. Introduction

Digital Elevation Model (DEM) is one of the most useful data types in geosciences with applications in multiple disciplines from geomorphology through civil engineering to virtual reality and entertainment. Popularity and broad adoption of DEMs follows in part from the fact that medium and low resolution DEMs are freely available for download; examples include the Shuttle Radar Topography Mission (SRTM) world-wide DEM and US-wide National Elevation Dataset (NED). Users can access DEMs by selected region, coordinates, or administrative regions. However, until now, these vast resources could not be explored in their entirety due to the lack of search tools for content-based retrieval of topographic information.

In this paper we report on development of Terrain Explorer (TerraEx) – the first application for content-based retrieval of topographic information. TerraEx is a GeoWeb application (http://sil.uc.edu/webapps/terraex) that enables the world-wide search for topographic landscapes similar to a user-selected query landscape. It uses a world-wide 3" (90m) DEM (see next section for details) for elevation data. The search is not performed directly on the DEM; instead it is performed on the DEM classified into the ten landform elements using the geomorphon method (Jasiewicz and Stepinski 2013a). TerraEx also doubles as the most convenient source of world-wide, geographically registered topographic data. A user can navigate to a location of interest (with or without a prior exploration using the search) and download (in GeoTIFF format) a 3" resolution DEM, a map of geomorphons, and a map of terrain relief.

In the process of constructing TerraEx we developed the following new products and concepts: (a) a new, adaptively smoothed version of world-wide, 3" DEM that removes artifacts in the flat areas, and (b) a methodology for comparison of spatial patterns in a globe geometry

2. Content-based search

Input data is the world-wide 3" DEM (Ferranti, 2014) – a compilation of the SRTM DEM with DEMs from other sources for areas not covered by the SRTM. This dataset covers 215000 × 420000 16bit integer cells, in the WGS84 lat-lon projection with vertical resolution of 1m. We fill all the gaps in the Ferranti data, convert it to floating point, de-noise, and adaptively smooth while preserving elevation values in Ferranti DEM to within 1 m. This removes terrace-like artifacts in the flat areas (30% of lands surface) – an important improvement, especially when classifying DEM into landform elements. This new world-wide DEM is classified into ten landform elements yielding the 3" world-wide map of geomorphons which is used for landscape search.

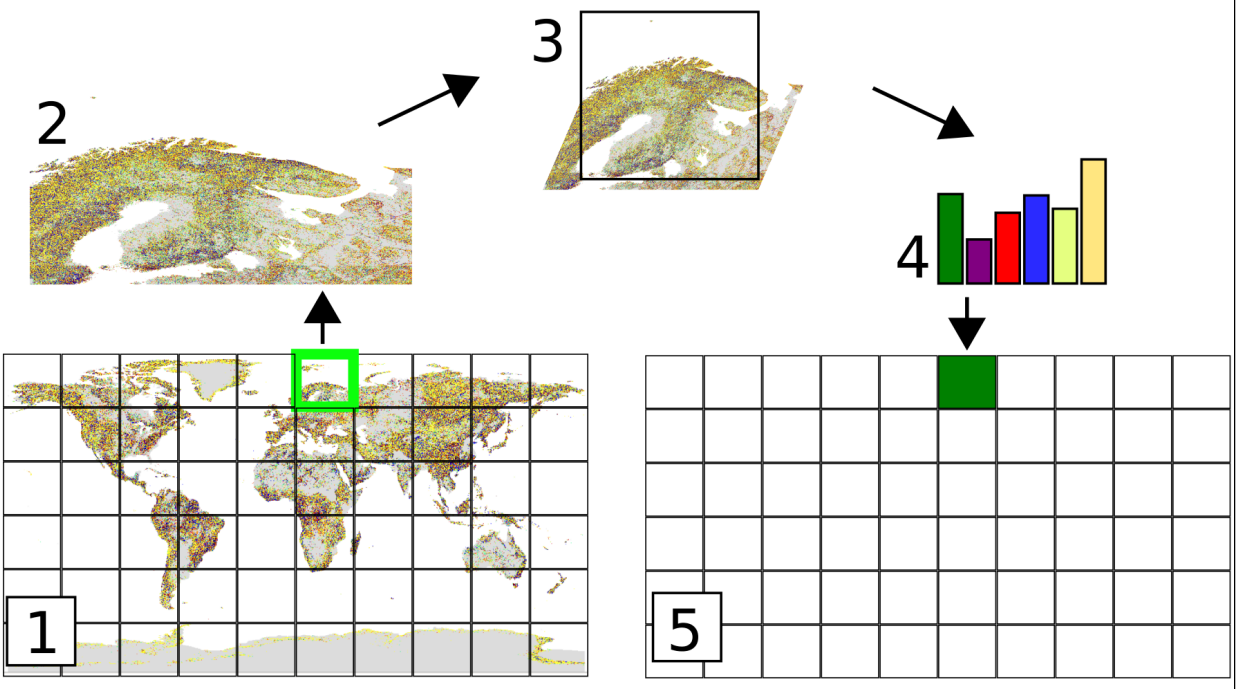


Figure 1: On-the-fly reprojection from WGS to UTM during grid calculation

The content-based search of topography is based on the same general principle as our earlier search for similar land cover patterns across the US (Jasiewicz and Stepinski, 2013b; Stepinski et al., 2014) but with modifications to accommodate the globe geometry. The geomorphons dataset is divided arbitrarily into a regular grid of local blocks of cells; each block (referred to as a motifel) contains a local pattern of the ten landform elements which serves as a proxy for the local landscape. For the purpose of calculating similarities between motifels (local landscapes) we represent them by the co-occurrence histogram of landform elements. In TerraEx a motifel has a size of 160 cells (~14km) and the global grid of all motifels has dimensions of 10800×5400 (motifels are arranged in a grid with a significant overlap).

The major difference from our previous US-wide search is that in the lat-lon grid of global extent the topography is severely distorted at high latitudes but search requires a comparison of undistorted landscapes. Fig.1 illustrated schematically how this problem is addressed in TerraEx: (1) lat-lon geomorphons dataset is divided into motifels; (2) for each motifel an appropriate

UTM zone is determined and a region of the geomorphon map large enough to cover a motifel after transformation to this UTM zone is selected; (3) the region is transformed to the local UTM; (4) co-occurrence histogram is calculated from the transformed map; (5) histogram is saved in the appropriate position in the lat-lon grid. The steps 1-5 are all calculated offline with the resultant grid of histograms saved in the server memory.

The landscape search starts with a user selecting a location; the histogram describing a local landscape in this location (a query) is set aside to be compared one-by-one with histograms from all motifels in the world-wide grid. We use the Ruzicka similarity measure (Deza and Deza 2014) between two histograms; it yields a value between 0 (not similar at all) to 1 (identical). The result of this comparison is the spatial layer containing values of similarities between a query and local landscapes. This layer – a similarity map – is displayed in TerraEx with a color gradient indicating similarity values. Visual examination of the similarity map provides information on where in the world are the landscapes similar to the query. The time from issuing a query to displaying a similarity map is about 10 sec. Such short wait time is achieved by an efficient computational engine (GeoPAT 2.0, a stand-alone extension of an original, GRASS GIS-based GeoPAT toolbox for pattern-based geoprocessing (Jasiewicz et al., 2015)), which is written in C and uses parallel computation based on the OpenMP library.

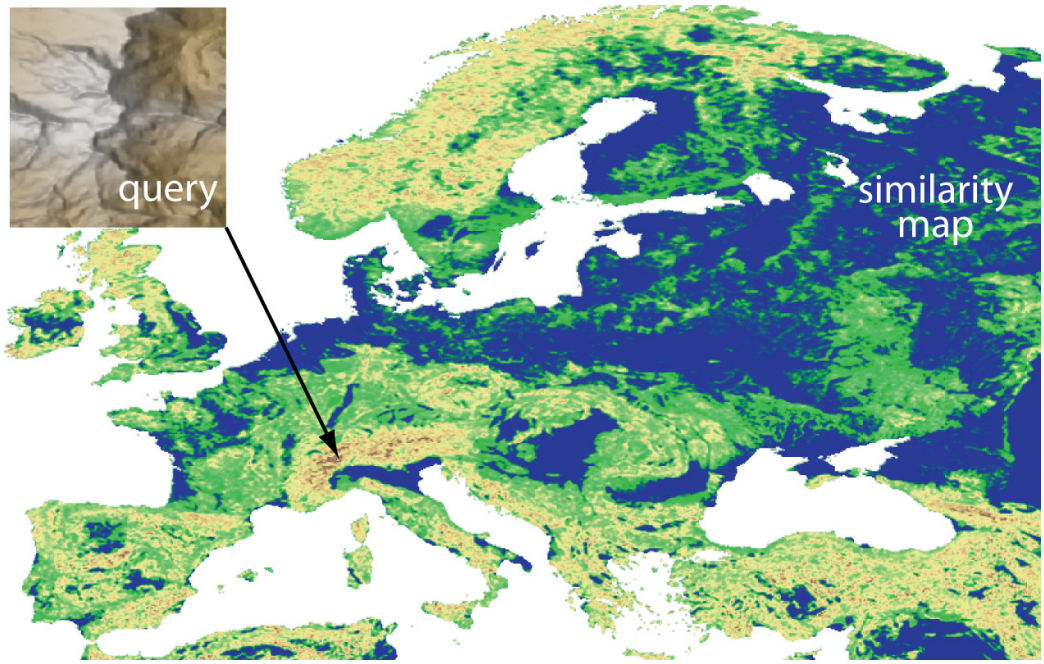


Figure 2: Example of landscape search using a specific alpine landscape as a query.

4. Example of landscape search

It is not possible to fully convey in the paper how TerraEx – an interactive, web-based application – works; one needs to use it to fully appreciate its functionality. Fig.2 illustrates

partial results of a particular search. As an example, we have chosen a location in the central Alps as a query. This location is indicated by an arrow in Fig.2 and a landscape in this location (spatially restricted to the scale of a motifel or ~14km) is shown in the inset in the form of a hillshade. The main panel of Fig.2 shows the similarity map for this query restricted to Europe (so some details could be seen). Brown colors indicate high similarity to a query; decreasing similarities to a query are shown by a yellow-green-blue gradient. As expected, TerraEx indicates that other locations in the Alps, as well as locations in Caucasus Mountains have largest similarities to a query, while central and eastern European planes have the smallest similarities to the query.

5. Conclusions and future work

TerraEx is a novel tool that enables, for the first time, content-based, world-wide exploration of topographic landscapes. Behind the scenes TerraEx is a modern, standards-compliant GeoWeb application, details of which could not be described here due to the lack of space. It also doubles as a very convenient distributor of world-wide 3" DEM and the map of geomorphons.

The major novelty of TerraEx is its ability of identifying places in the world having landscapes "similar" to a query. Note that what actually is calculated is similarity of patterns of landform elements as given by the geomorphon method. As with all content-based searches there is a possibility of the semantic gap – a difference between similarity as calculated by an algorithm and similarity as perceived by an analyst. Geomorphons are not the only way to define landform elements, co-occurrence histogram representation of landscape is one of many possible heuristics, and similarity between histograms can be calculated by a number of methods. In addition, TerraEx currently implements search of landscapes only on a single scale of ~14km. Future research will address these issues and future versions of TerraEx may provide more choices. For now TerraEx should be used as any other search engine, the results should be taken as suggestions and examined by an analyst for utility.

Acknowledgements

This work was supported by the University of Cincinnati Space Exploration Institute, by Grant NNX15AJ47G from NASA, and by the National Science Center (NCN) grant DEC-2012/07/B/ST6/01206.

References

Deza, M.M., Deza, E. (2014). *Encyclopedia of Distances, third edition*, Springer-Verlag, Berlin Heidelberg, p.325

Ferranti J. (2014) *DIGITAL ELEVATION DATA*, http://viewfinderpanoramas.org/dem3.html, internet source, accessed 2016-04-24

Jasiewicz, J. and Stepinski, T.F., 2013a. Geomorphons—a pattern recognition approach to classification and mapping of landforms. *Geomorphology*, 182, pp.147-156.

Jasiewicz, J. and Stepinski, T.F., 2013. Example-based retrieval of alike land-cover scenes from NLCD2006 database. *Geoscience and Remote Sensing Letters, IEEE*, *10*(1), pp.155-159.

Jasiewicz, J., Netzel, P., and Stepinski, T. (2015). *GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases*. Computers and Geosciences, 80, 62–73.

Stepinski, T. F., Netzel, P., and Jasiewicz, J., (2014). *LandEx - A GeoWeb tool for query and retrieval of spatial patterns in land cover datasets*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 7(1), 257–266.

Unsupervised Delineation of Urban Structure Types Using High Resolution RGB Imagery

J. Niesterowicz[1], T. F. Stepinski[1], J. Jasiewicz[1,2]

[1]Space Infromatics Lab, University of Cincinnati, 401 Braunstein Hall, Cincinnati, OH 45221-0131, US
Email: {niestejk; stepintz}@mail.uc.edu

[2]Instititute of Geoecology and Geoinformation, Adam Mickiewicz University, Dziegielowa 27, Poznan, Poland
Email: jarekj@amu.edu.pl

Abstract

We present a method for delineating Urban Structure Types (USTs) using only high resolution RGB images. As the method is unsupervised, it does not require training; the interpretations of delineated USTs are assigned a posteriori. The method utilizes freely available software and performs delineation in a short time even for very large images. A 1-meter resolution image of the entirety of Los Angeles is delineated as an example. We have found seven distinct USTs which were given interpretations based on examination of their patterns. These interpretations are validated by population statistics. The method aims at broaden the usage of USTs delineations for applications in urban and social studies.

1. Introduction

Urban Structure Type (UST) is a distinct spatial pattern of the urban structure at the neighborhood scale, which can be interpreted in terms of the type of activity or of residential pattern. Classification of a city into USTs complements standard land cover/land use classification by working at a scale that is significantly coarser than an individual pixel. Fairly extensive literature exists on how to delineate USTs from remotely sensed data (for example, see Heiden *et al.* 2012), but, because these works focus on supporting effective urban planning, they use multisource data and supervised learning. This means that they are restricted to very few places where this data exists and where the significant cost of supervised analysis is justified by the need. There also exists extensive literature on using single-source data (RGB or multispectral images) but only in the context of separating two specific types of USTs – formal from informal settlements (slums); for example see Graesser *et al.* (2012). Algorithms proposed there are restricted to this single purpose; they also are predominantly based on supervised learning.

In this paper we present an approach to delineation of USTs that uses only RGB images (many of which are freely available online) as input, delineates an exhaustive set of USTs, is based on training-free, data-driven unsupervised principles, and can process very large input data in a reasonable time. In addition, our method relies only on existing public domain software. Our motivation is to make the delineation of USTs more broadly accessible to analysts from different disciplines. The methodology is described and applied to a ~2 billion pixel 1 m-resolution image covering the greater Los Angeles area.

2. Methodology

Our method is based on the concept of Complex Object-Based Image Analysis (COBIA) (Vatsavai 2013; Stepinski *et al*. 2015). In COBIA a raster (not necessarily restricted to an image) is divided arbitrarily into a grid of local blocks of cells. We refer to these blocks as motifels –

they encapsulate a local pattern (motif) of raster variable. Motifels are much larger than pixels and they are also much more complex; motifels are used as elementary units for further analysis. COBIA is a well-suited approach to the problem of delineation of USTs from an RGB image. Individual pixels in the RGB image may not have an unambiguous interpretation, but we expect that the composition and arrangement of different colors within a motifel can be associated unambiguously with the specific UST. We utilize the GeoPAT toolbox (Jasiewicz *et al*. 2015) – an open source collection of GRASS-GIS modules for pattern-based geoprocessing – to implement COBIA. GeoPAT works with categorical rasters, thus, as the preprocessing step, we first quantize an RGB image to obtain a raster of class color labels.

We encapsulate the complexity of the motifel structure by a color co-occurrence histogram and calculate the degree of dissimilarity between two motifels using the Jensen-Shannon Divergence (JSD) (Lin, (1991)) between their corresponding histograms. A grid of motifels having dimensions much smaller than an original image, is first segmented into local areas of homogeneous patterns. Segmentation provides enhanced spatial cohesion to the final UST classification and reduces the dimensionality of the subsequent clustering step. As spatially distinct segments may have very similar patterns, in the final step we cluster the segments into a small number of USTs. Both segmentation and clustering steps utilize JSD as the measure of distance. We use a segmentation algorithm which is custom-designed for grids of motifels. This algorithm is a part of the GeoPAT toolbox and is described in Jasiewicz *et al.* (2016) (see also Jasiewicz *et al.* contribution to this conference); its only parameter is the motifel's size. We use R implementation of the Partitioning Around Medoids (PAM) algorithm to cluster the segments.

3. Delineating USTs in the Los Angeles area

To demonstrate our methodology we delineated USTs in the greater Los Angeles area (see inset in panel A of Fig.1) using 1-meter resolution RGB (3-band) aerial imagery freely distributed by the U.S. Geological Survey (http://viewer.nationalmap.gov/launch). We downloaded 36 individual images (available in jpeg2000 format) from USGS and mosaicked them into a single image (shown in Fig.1 panel A) having dimensions of 41600×50200 pixels. This image was quantized into 27 color class labels to prepare it for GeoPAT processing. The quantization procedure first converts RGB color into the CIE L*a*b* color space, in which the numerical distance between two colors corresponds to the perceptual distance between them. Then, the CIE L*a*b* color space is divided into cubic blocks of a selected size. Non-empty blocks are the quantized colors; 27 colors were obtained with the size of the block equal to 25 units. Next, we have chosen the size of the motifel to be 200 pixels (meters) which roughly corresponds to the size of a city block. With such a choice the image was transformed into a grid of 208×251 motifels. The segmentation step resulted in 4841 segments which were clustered into 7 USTs. The number of USTs was determined experimentally so the homogeneities of patterns in USTs are maximized while the number of USTs is kept to the minimum.

Because our method is unsupervised, an interpretation of each UST must be made a posteriori. To arrive at an interpretation we constructed a synthetic image for each UST. A synthetic image of an UST consists of 400 motifels selected randomly from its extent and organized into 20×20 array. Smaller versions of synthetic images for all USTs are shown in the last two rows in Fig.1. Based on our examination of these synthetic images we gave the USTs interpretations as listed in the legend in Fig.1. Two of the USTs (labeled as 5 – green areas and 6 – forest/shrub) are interpreted as undeveloped and uninhabited areas. Two others (labeled as 1 – sparse residential and 2 – dense residential) are interpreted as residential areas with detached

housing. Two UST classes (labeled as 3 – dense urban and 4 – commercial/ind) are characterized by their high percentage of impervious surfaces with class 3 appearing to be a mixture of multi-level apartments, shopping centers, and urban infrastructure; class 4 appears to contain purely commercial structures. The seventh UST class (labeled as 7 – others) consists of large construction areas, barren land, and partially, of sparsely populated barren land.

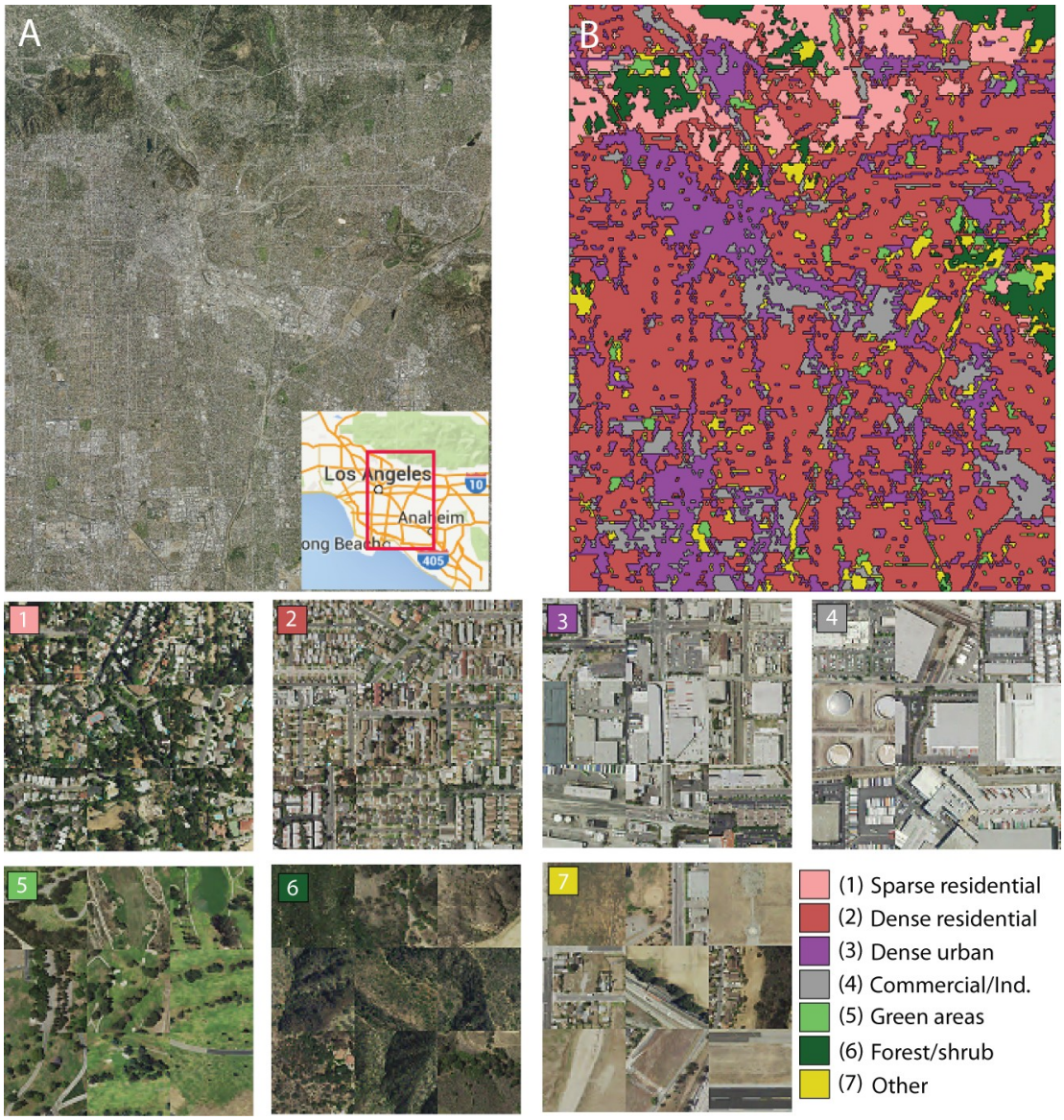


Figure 1: Unsupervised delineation of USTs in the Los Angeles area. (A) Original 1-meter resolution image (inset shows the location of the image). (B) Map of seven USTs found in this area. The two lowest rows show a series of seven synthetic images each consisting of 9 randomly selected motifels from each UST. The legend to USTs is given in the lower right corner.

In order to validate our interpretations we used the newly available online resource SocScape (available at http://sil.uc.edu/webapps/socscape_usa/) which provides high resolution (30-meters) gridded population data for the entire U.S. Unlike census data, which gives population counts aggregated to areal units, SocScape data distinguishes between inhabited and uninhabited

areas and can be used to calculate population density and the percentage of inhabited area in each UST class. The results are given in Table 1 and confirm our interpretations.

Table 1. Population statistics for the UST classes.

	1	2	3	4	5	6	7
Population density (people per ha)	20.0	50.2	40.3	8.9	2.9	0.6	10.8
% inhabited area	82.4	86.3	55.2	22.3	10.3	14.2	35.4

4. Conclusions

We presented a method for fast, unsupervised delineation of USTs from high resolution RGB images. We also demonstrated how this method works using an image of the Los Angeles area as an example. For this large image the processing time on the 8-core I7 computer was ~30 min, of which ~20 min was color quantization preprocessing. The results indicate that the method gives reasonable and valuable results using only an easy-to-obtain RGB image and our software which is in the public domain. The method works even if an image is in a compressed jpeg2000 format. This is in contrast to other works on delineation of USTs which either use multisource, difficult to obtain data or concentrate on the extraction of a single UST (slums), in both cases using proprietary software. We noticed that some USTs are more difficult to distinguish from an RGB image than others when using our method. In particular, dense urban environments mix with some (but not all) infrastructure and light commercial environment (class 3). Increasing the number of clusters does not help to resolve this problem because these environments are indeed characterized by similar patterns in an image. As 1-meter RGB images are available from the USGS for the entire U.S., our method can be used for comparative studies among major U.S. metropolitan areas.

Acknowledgements

This work was supported by the University of Cincinnati Space Exploration Institute, and by Grant NNX15AJ47G from NASA.

References

Graesser, J., Cheriyadat, A., Vatsavai, R.R., Chandola, V., Long, J., Bright, E., 2012. Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 5(4), pp. 1164–1176

Heiden, U., Heldens, W., Roessner, S., Segl, K., Esch, T., Mueller, A., 2012. Urban structure type characterization using hyperspectral remote sensing and height information. *Landscape and Urban Planning*. 105, pp. 361–375

Jasiewicz, J., Netzel, P. & Stepinski, T., 2015. GeoPAT: A toolbox for pattern-based information retrieval from large geospatial databases. *Computers & Geosciences*, 80, pp.62–73.

Lin, J. 1991, Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Stepinski, T.F., Niesterowicz, J. & Jasiewicz, J., 2015. Pattern-based Regionalization of Large Geospatial Datasets Using Complex Object-based Image Analysis. *Procedia Computer Science*, 51, pp.2168–2177.

Vatsavai, R.R., 2013. Object based image classification: state of the art and computational challenges. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. pp. 73–80.

Modelling Vague Shape Dynamic Phenomena from Sensor Network data using a Decentralized Fuzzy Rule-Based Approach

Roger Cesarié Ntankouo Njila[1], Mir Abolfazl Mostafavi[1], Jean Brodeur[1]
[1]Centre de Recherche en Géomatique, 2314 Pavillon Casault; Université Laval, Québec, Canada, G1K 7P4
Email: roger-cesarie.ntankouo-njila.1@ulaval.ca, Mir-Abolfazl.Mostafavi@scg.ulaval.ca, recherchegeosemantic@videotron.ca

Abstract

Modelling dynamic phenomena of vague shape from sensor data is still a challenging problem for many applications. In this paper, we propose a decentralized fuzzy rule-based approach based on fuzzy object model to build a more realistic spatiotemporal representation for such phenomena. This approach has been successfully implemented in a simulation case of bushfire monitoring, showing advantages for spatial decision making in a disaster management context.

Keywords: *sensors, sensor data, fuzzy objects, disaster management, spatial decision support*

1. Introduction

Extracting geospatial information from geosensor data can help to better understand a complex phenomenon for real time decision making process (Sadeq *et al*. 2013). Several approaches are used for the extraction of geospatial information from sensor data. Many of these approaches are developed based on the assumption that monitored phenomena are of crisp shape with well-defined boundaries. However, many dynamic phenomena have vague spatial boundaries, and their accurate detection and extraction from sensor data is a challenging problem.

In this paper we propose a decentralized fuzzy rule-based approach to address this problem. In the proposed method, sensors detect vague shape phenomena using a fuzzy logic reasoning approach and collaborate with their neighboring sensors to infer vague spatial extent of the phenomena and its dynamics. We adopt Crisp-Fuzzy objects (Pauly and Schneider 2008) as a more realistic model for large scale and vague shape dynamic phenomena.

This paper is organized as follows. After a brief background presented in section 2, section 3 describes the proposed approach implemented in section 4 for a bushfire simulation case showing it applicability for real time spatial decision process in disaster management. Section 5 presents conclusions and future works.

2. Background

Spatial computing can be undertaken in a sensor system following a centralized (all data are sent to process center), a decentralized (located at sensor site or other site) or a hybrid approach (Chong and Kumar 2003). Crisp vector objects are extracted in the existing approaches using statistics or filters (Chintalapudi and Govindan 2003) or qualitative reasoning (Guan and Duckham 2009). Real-world phenomena are inherently uncertain (Carniel *et al*. 2015).

Crisp-Fuzzy objects model (Pauly and Schneider 2008) is an interesting candidate for the representation of such phenomenon. In this model the geometry of object is composed of a kernel and a conjecture part, the kernel part belongs definitely and always to the vague object but one can't say with certainty whether conjecture part considered as the broad boundary belongs to the vague object.

Fuzzy models are defined based on Fuzzy Set Theory that deals with inherent fuzziness and uncertainty of a phenomenon in real world through a membership function (***MF***) (Ross 2010) determining the degree of belonging of an element to a set. ***MF*** is essential for defuzzyfication, it helps in reducing a fuzzy set to a crisp single-valued quantity, or to a crisp set (Ross 2010), thus given sensor data values, ***MF*** is used to infer the belonging of sensor positions to parts of the vague object (***Vobj***).

The fuzzy spatiotemporal representation of a dynamic phenomenon (forest fire) extracted from sensor data represents an interesting tool for decision making. This can help for example to manage people evacuation to safe place or efficiently deploy resources for firefighting. This can also help individual users to answer to the questions such as: Is my location surely in, out, near or far from fire zone? Complete answer to such questions requires collaboration among sensors.

3. Presentation of the proposed approach

Sensors perform measurements over time, from the set of collected data at a given time, spatial computing can be undertaken in the sensor network (SN), building temporal spatial view of the phenomenon. This can be undertaken over time with changes detection to understand the dynamics of the phenomenon. This section presents the two reasoning stages implemented by sensor nodes to build a fuzzy spatiotemporal representation of a phenomenon from sensor data.

3.1 Stage-1: Local detection of phenomenon using three-valued logic

Using a built-in reasoning engine, each sensor evaluates its membership to different parts of the spatial extend of a dynamic phenomenon using a *MF* and the observed data at a given time. *MF* definition should cope with the semantics of sensor data and phenomenon (Ross 2010). For illustration, let's consider a sensor network recording temperature in an area under bushfire, considering the ontology of sensor data for fire monitoring (Gao et al. 2014), *MF* for such case can be of S shape as illustrated in Figure 1.

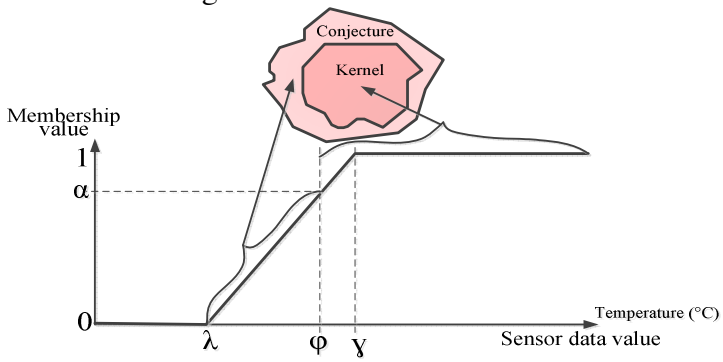


Figure 1 : S shape membership function for bushfire detection from sensed temperature.

Considering Sr as record value of sensor s at location $Loc(s)$ and time t, the membership function f defines μ the membership value of $Loc(s)$ as follows.

$$\mu = f\{Sr(Loc(s),t)\} : Sr \rightarrow [0,1] \quad (1)$$

Considering the Crisp-Fuzzy object $(Vobj)$ representing the phenomenon, we use three-valued logic to define IF-Then rules used for defuzzyfication, these rules can be defined as follows:

$$IF\ \mu \geq \alpha\ Then\ (Loc(s),t) \in Kernel \quad (2)$$

$$IF\ 0 < \mu < \alpha\ Then\ (Loc(s),t) \in Conjecture \quad (3)$$

$$IF\ \mu = 0\ Then\ (Loc(s),t) \in Outside \quad (4)$$

At this stage, sensors are aware of the presence or absence of phenomenon at their location but no information about the spatial extent of the phenomenon is inferred.

3.2 Stage-2: collaborative spatial reasoning for the extraction of vague object representing the phenomena

A sensor detecting the phenomenon (Kernel or Conjecture) sends queries to its one hop (directly linked) neighbors named N(s) through the communication mesh (Gabriel Graph) materialized by links between nodes, as shown in Figure 2. Nodes sending queries provide each of their respondents with the required information: Identification, location, detection value at given time. The general form of query is as follows:

$$Type\_Query : My\_Id = ii, My\_loc = xy, My\_detection = type, Time = tt, ?Your\_Id, ?Your\_loc, ?Your\_detection \quad (5)$$

Two types of queries are propagated in the network, ***KQuery*** and ***CQuery*** for query with ***My _detction = Kernel*** and ***Conjecture*** respectively. Each node not detecting the phenomenon (outside) by receiving even one query (***KQuery*** or ***CQuery***), holds an outer-bordering position as shown in Table 1. Querying node infers on it possible inner-bordering position from the set of received answers, then determines closest vertices halfway to linked outer-bordering nodes. Joining the set of identified vertices can be undertaken by leads nodes, to build kernel and conjecture boundaries at a given time. Nodes are labeled for easy reading/computing with 2 characters denoting phenomenon detection and bordering position respectively.

Table 1 : Relative position of nodes according to parts of the monitored phenomenon.

Phenomenon part detection	Type of query or answer received	Relative position	Label-value
Kernel	Only ***KQuery***	*Kernel - inner*	1 – 0
Kernel	Even one ***CQuery*** or one answer from Outer node	*Inner - Kernel - boundary*	1 – 1
Conjecture/Outside	Even one ***KQuery***	*Outer - Kernel - boundary*	1 – 2
Conjecture	Only ***CQuery***	*Conjecture - Inner*	2 – 0
Conjecture	Receiving answer from an Outer node	*Inner - Conjecture - boundary*	2 – 1
Outside	Even one ***CQuery***	*Outer - Conjecture - boundary*	2 – 2
Outside	No query	*Outer*	0 – 0

In dynamic phenomena, occurring changes can modify the relative position of nodes over time, thus changing the spatial extent of the phenomenon (expansion, narrowing...).

4. Modeling a monitored bushfire for spatial decision support

Here we present a bushfire case study to illustrate the applicability of the proposed method for the extraction of a fuzzy spatiotemporal representation of the phenomena using sensor data. As presented in Figure 2, The monitored phenomena is made of Kernel (pink) and Conjecture (yellow) parts in an area with road and water networks (magenta and blue respectively). Sensors are represented by dot and are linked together by a communication mesh.

Information on the spatial extent of kernel and conjecture parts or about their spatial evolution can help in managing the resource (human/material) to be mobilized while fighting against fire. Sensors are also labeled with information which can be used in strategic decision making process as evacuation strategy of endangered population by selection the appropriate road linking to safe area (no label for $0-0$ as mentioned in Table 1). Also, $2-2$ labeled positions which are out of the fire area express the proximity to area where one may meet fire.

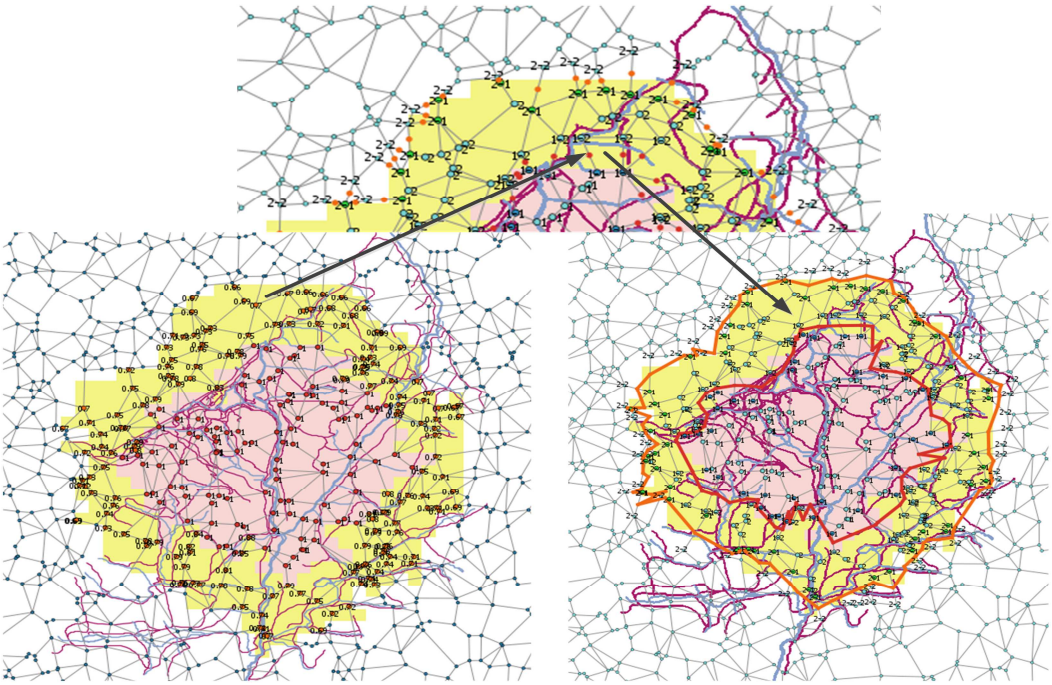


Figure 2 : Extracting fuzzy spatiotemporal representation of bushfire.

5. Conclusions and future works

In this paper we have presented a fuzzy rule-based approach that integrates sensor data describing phenomena of vague shape. Semantic information on sensors are used to establish fuzzy rules used by sensor to reason on their membership value and collaboratively infer a more realistic spatiotemporal representation of a vague dynamic phenomenon. We have also presented how these representations can be used for spatial decision making process for a disaster management case. For future works, we intend to consider heterogeneities in sensor data and observations context in the reasoning approach.

Acknowledgements

The authors would like to thank all the committee members of PEFOGRN-BC for their support.

References

Carniel, A.C., Schneider, M. and Ciferri, R.R., 2015. FIFUS: A Rule-Based Fuzzy Inference Model for Fuzzy Spatial Objects in Spatial Databases and GIS. In *ACM-SIGSPATIAL, Seattle, WA, USA*.

Chintalapudi, K.K. and Govindan, R., 2003. Localized edge detection in sensor fields. *Ad Hoc Networks*, 1(2-3), p.273–291.

Chong, C.-Y. and Kumar, S.P., 2003. Sensor networks: evolution, opportunities, and challenges. *Proceedings of the IEEE*, 91(8), p.1247–1256.

Gao, L., Bruenig, M. and Hunter, J., 2014. Estimating fire weather indices via semantic reasoning over wireless sensor network data streams. *International Journal of Web & Semantic Technology (IJWesT)*, 5(4), p.20.

Guan, L.-J. and Duckham, M., 2009. Decentralized computing of topological relationships between heterogeneous regions. In *Proc. 10th International Conference on GeoComputation, Sydney, Australia*.

Pauly, A. and Schneider, M., 2008. Vague Spatial Data Types. In *Encyclopedia of GIS SE - 1434*. Springer US, p. 1213–1217.

Ross, T.J., 2010. *Fuzzy logic with engineering applications* 3rd ed., John Wiley & Sons.

Sadeq, M.J., Duckham, M. and Worboys, M.F., 2013. Decentralized detection of topological events in evolving spatial regions. *The Computer Journal*, 56(12), p.1417–1431.

Managing and Updating Geographical Data: Issues Along the Hierarchical Chain?

K. Ooms[1], J. Crompvoets[2], P. De Maeyer[1], P. Lambert[3], E. Mannens[3], N. Van de Weghe[1], S. Verstockt[3], P. Viaene[1]

[1]Department of Geography, Ghent University, Krijgslaan 281 (S8), 9000 Ghent, Belgium
Email: {kristien.ooms;philippe.demaeyer;nico.vandeweghe;pepijn.viaene}@ugent.be

[2]KU Leuven Public Governance Institute, Parkstraat 45 (box 3609), 3000 Leuven, Belgium
Email: joep.crompvoets@soc.kuleuven,be

[3]Data Science Lab, Ghent Univeristy - iMinds, Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
Email: {peter.lambert;erik.mannens;steven.verstockt}@ugent.be

Abstract

This article describes the needs and challenges (technical, juridical and governance) related to exchanging, managing and updating a large scale geographic reference dataset. The focus is placed on a specific case, namely the large scale reference frame of Flanders, the GRB. Furthermore, new challenges and needs for the future are considered.

1. Introduction

Nowadays most countries have a detailed digital geographic dataset covering their territory (Carpenter & Snell 2013). To manage this dataset, it is integrated in the country's hierarchical administrative structure, with specific rules regarding who can use, manage and update the information. This is embedded in framework, which is often referred to as a Spatial Data Infrastructure (SDI) (e.g. Crompvoets et al. 2004). These SDIs have proven to be especially challenging when considering the interplay between the different governments and other actors within a country, each with their own authorities and responsibilities (Jacoby et al. 2002; Warnest 2005). Nevertheless, the development of (national) SDIs is not new, but a shift has been noticed in its main goal from data to data use (Williamson et al. 2006).

1.1 GRB – the Large Scale Reference Frame in Flanders

In Flanders, the dataset which serves as the large scale reference frame is called GRB (*Grootschalig Referentiebestand*). Its specifications are imposed by the Flemish Government (objects to be included, metadata, finances, use, management, maintenance, etc.) (Ministerie van de Vlaamse Gemeenschap 2004; AGIV 2014), and is managed by *Informatie Vlaanderen* (IV), a governmental institution. Recently, access to this detailed dataset has been opened up to a wider public, including architects, notaries, surveyors, intercommunal, federal government, etc. (Informatie Vlaanderen 2015). Since January 1st 2015, the GRB became obligatory as a reference frame for its users to facilitate the exchange of large scale geographical data. This means that the users of the reference frame will superimpose their own (thematic) data layers on top of this reference frame.

1.2 Top Level Updates vs Low Level Mismatches

When dealing with geographic information that covers a territory, keeping it up-to-date across all related instances can be problematic (Jing et al. 2014). The Flemish Government is responsible for maintaining the dataset of the GRB. Updates can be divided into two groups: (1) a change in the real situation (splitting or merging a parcel, change in road infrastructure, adding a building, etc.); (2) quality improvements (e.g. more accurate measurements,

correction of errors) with no change in the real situation. The latter is not linked to any official documentation (e.g. from a notary); potentially creating inconsistencies between the reference data and the 'local' layers. These mismatches are dependent on how these data layers are linked and are, at the moment, corrected manually, demanding a lot of effort and time from the GRB users. These incoherent geographical data bases introduce a level of uncertainty towards citizens who request information from the local government.

2. Creating a 'smart' framework for data exchange?

2.1 Technical challenges

The technical challenges are related to three dimensions; finding appropriate data structures, matching processes and migration processes. As a first step, existing structures and processes for data exchange need to be reviewed (Shi & Walford 2012). In this context, the use of ontologies and semantic geodata is crucial (Pauwels et al. 2009). This is already being implemented in the context of INSPRIRE, ISA, core vocabularies, etc. (e.g. Masser 2007; de Vries et al. 2011) Furthermore, the implementation of 'linked data' structures (e.g. using a Resource Description Framework) seems to hold promises in updating data in this complex framework (Geiger & von Lucke 2012; Kuhn et al. 2014). Because the data is used by different types of actors, who superimpose their own (local) datasets, interoperability is of utmost importance (Bishr 1998; Stoimenov & Đorđević-Kajan 2002)

2.2 Juridical challenges

Because of the type of data – large scale governmental data, including parcels, buildings and their (legal) characteristics – juridical challenges should be considered in this framework (Onsrud 2004; Janssen & Crompvoets 2012). Two key elements are at play here: *responsibility* (Zevenbergen et al. 2016) and *liability* (Cho 2012). In this context the following questions arises: How precise and accurate should the provided data be to avoid legal issues? Who is responsible and liable when this information is wrong or inaccurate?

2.3 Governance challenges

Besides the technical and juridical challenges, an efficient workflow throughout the hierarchical chain is a key factor, aligning the framework and the decision making process. This is closely linked with the technical (distributions of the changes throughout the framework) and legal implications (Who is responsible?). Collaboration, motivation, and trust are the key elements to create an operational framework (Harvey 2003; Craig 2005; Warnest 2005). Nevertheless, the needs of the end users should not be neglected in this process.

3. Conclusion

In Flanders, the GRB is used as the large scale reference frame by governmental and other institutions, which is managed by the Flemish Government. Corrections in this dataset can create a conflict with existing datasets created by other (governmental) institutions, resulting in an uncertain data source. Nevertheless, significant challenges – technical, juridical and governance – need to be tackled to be able to provide reliable geographic information to the end user.

Acknowledgements

The work presented in this paper is initiated at the request of several governmental institutions in Flanders, Belgium. The authors therefore would like to acknowledge their valuable input on this topic, specifically Ward Van Hal (Flemish Association for Cities and Municipalities); Peter

Bogaert (City of Ghent); Ruben Maddens (City of Ostend); Dirk Goeminne (Province of East-Flanders); Hendrik van Hemelryck (Flemish Government).

References

AGIV. (2014). Het Grootschalig Referentiebestand.

Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability. *International Journal of Geographical Information Science, 12*(4), 299-314.

Carpenter, J., & Snell, J. (2013). *Future trends in geospatial information management: The five to ten year vision*.

Cho, G. (2012). Geographic Data and Legal Liability Issues. In K. Janssen&J. Crompvoets (Eds.), *Geographic Information and the Law – Defining New Challenges* (pp. 109-122). Leuven: Leuven University Press.

Craig, W. (2005). White knights of Spatial Data Infrastructure: The role and motivation of key individuals. *URISA journal, 16*(2), 5-13.

Crompvoets, J., Bregt, A., Rajabifard, A., & Williamson, I. (2004). Assessing the worldwide developments of national spatial data clearinghouses. *International Journal of Geographical Information Science, 18*(7), 665-689.

de Vries, W., Crompvoets, J., Stoter, J., & VandenBerghe, I. (2011). Atlas of INSPIRE: Evaluating SDI Development through an Inventory of INSPIRE Experiences of European National Mapping Agencies. *International Journal of Spatial Data Infrastructures Research, 6, 2011*.

Geiger, C. P., & von Lucke, J. (2012). Open government and (linked)(open)(government)(data). *JeDEM-eJournal of eDemocracy and Open Government, 4*(2), 265-278.

Harvey, F. (2003). Developing geographic information infrastructures for local government: The role of trust. *The Canadian Geographer/Le Géographe canadien, 47*(1), 28-36.

Informatie Vlaanderen.(2015). Vlaamse Regering maakt van GRB open data. Retrieved 01/05/2016, 2016

Jacoby, S., Smith, J., Ting, L., & Williamson, I. (2002). Developing a common spatial data infrastructure between State and Local Government--an Australian case study. *International Journal of Geographical Information Science, 16*(4), 305-322.

Janssen, K., & Crompvoets, J. (2012). *Geographic Information and the Law – Defining New Challenges*. Leuven: Leuven University Press.

Jing, Y., Rohan, A. P., & Zevenbergen, J. (2014). Up-to-dateness in land administration: setting the record straight. *Coordinates*, 37.

Kuhn, W., Kauppinen, T., & Janowicz, K. (2014). Linked data-A paradigm shift for geographic information science *Geographic Information Science* (pp. 173-186): Springer.

Masser, I. (2007). *Building European spatial data infrastructures* (Vol. 380): Esri Press Redlands, CA.

Ministerie van de Vlaamse Gemeenschap (2004). Decreet houdende het Grootschalig Referentie Bestand (GRB).

Onsrud, H. J. (2004). Geographic information legal issues. *Encyclopedia of Life Support Systems (EOLSS), Developed under the auspices of the UNESCO*.

Pauwels, P., Verstraeten, R., De Meyer, R., & Van Campenhout, J. (2009). Semantics-based design: can ontologies help in a preliminary design phase? *Design principles and practices. An international journal, 3*(5), 263-276.

Shi, S., & Walford, N. (2012). Automated geoprocessing mechanism, processes and workflow for seamless online integration of geodata services and creating geoprocessing services. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 5*(6), 1659-1664.

Stoimenov, L., & Đorđević-Kajan, S. (2002). Framework for semantic GIS interoperability. *FACTA Universitatis, Series Mathematics and Informatics, 17*(2002), 107-125.

Warnest, M. (2005). A collaboration model for national spatial data infrastructure in federated countries.

Williamson, I. P., Rajabifard, A., & Binns, A. (2006). Challenges and issues for SDI development.

Zevenbergen, J., de Vries, W., & Bennett, R. (2016). *Advances in Responsible Land Administration*: CRC Press.

Spatially situating remote users: An examination of immersive technology on presence and team participation

D. Oprean[1], M. B. Simpson[2], A. Klippel[2]

[1]Stuckeman Center for Design Computing
The Pennsylvania State University
Email: dxo12@psu.edu

[2]Department of Geography
The Pennsylvania State University
Email: {marksimpson;klippel}@psu.edu

Abstract

As today's workforce becomes more distributed, technology provides a means to communicate over long distances. This technology however offers limited forms of communication which can lead remote participants to not feel fully participant in collaborative efforts. With the renewed interest in immersive virtual reality (iVR) technology, the promise of more affordable and better capabilities is better than ever. iVR technology can allow remote collaborators to become spatially situated in the content of a collaboration. For this work-in-progress we examine levels of immersion to understand the role of co-presence for a remote participant on team membership and participation. Our experimental set up and preliminary results will be discussed.

1 Introduction

In today's workforce, technology has enabled communication and work to occur at a distance. Remote parties can easily join in meetings virtually through the use of various collaborative equipment. However, traditional audio and video technologies cannot create the full experience of being co-present in a meeting space with collaborators.

With the increased capabilities, affordability, and renewed interest in immersive virtual reality (iVR) technology, the potential to spatially situate remote parties into a meeting is higher than ever. iVR technology has been a communication tool for several decades, but recent breakthroughs make the hardware and software to create iVR experiences easier to use than ever before. Past research using iVR systems have examined their ability of spatially situating individuals into task spaces with great success. However, these studies have examined such technology through the limited scope of testing a single system. This 'black-box' approach to VR systems research is a limiting factor in the literature. We believe that comparing multiple iVR system aspects will allow for better connections to spatial experiences that will be generalizable to a broader scope of technology.

We used our work-in-progress to address the question of how spatially present a remote collaborator feels when experiencing different levels of visual immersiveness. Specifically, this study seeks to examine the relative impact of different levels of immersive technology on spatial co-presence in a meeting with one remote participant. To do this, our study built on the theoretical framework and methodological approach of Balakrishnan, Oprean, Martin, and Smith (2012) to identify levels of immersiveness. Our methodological approach allowed us to address our question through examining technology affordances—the attributes of technology that allow for actions or perceptions of actions by the appropriate entity (Greeno 1994).

1.1 Presence and Immersion

The nature of presence, the feeling of being 'present' within a given medium (Steuer 1992), in and of itself is highly subjective, making it hard to distinguish from concepts like immersion,

an objective measure of a medium providing sensory engagement (Slater 1999). Our focus was on a theoretical framework forming sense of presence in various forms (spatial and co-) through more engaging, 'immersive', technology.

We focused on co-presence, which should be distinguished from *telepresence*. Whereas telepresence can occur without others involvement, co-presence depends on having another human to connect with through a communication medium (Nowak and Biocca 2003), a human-human relation (Zhao 2003). The sense of presence felt by a remote participant may have implications for their participation in team activities. Sense of presence has been proven to be impacted by immersive capabilities of technology, specifically iVR technology (Balakrishnan and Sundar 2011).

1.2 Exploration of Technology to Spatially Situate Users

Social co-presence is important, but we believe there are other influences on a user's sense of participation, particularly when it comes to real-world team problems like distant field sites. By contextually placing remote participants into spaces through more immersive technology, we sought to engage the user's senses, motivating them to become more involved. We built off the education domain's use of situated learning, where virtual environments (VE) spatially situate learners (Wilson and Myers 2000). Learning through a VE raises two challenges: teaching users how to communicate with a medium and teaching users the content presented. This two-step process applies to collaborative systems in distributed workforces as well as education (Robey *et al.* 2000). As technology becomes easier to use, we believe that features meant to engage while reducing difficulty in use will enable remote collaborators to communicate better.

2 Methods

To address our overall inquiry, we aligned our selected technology on a spectrum where we considered each of the attributes for comparison as conceptualized variables. This process followed the adapted variable-centered approach utilized by Balakrishnan and Sundar (2011) to look at technology affordances. Immersion as distinguished by Slater (1999) is the objective output of the technology engaging a user's senses. Bowman and McMahan (2007) additionally used a taxonomy of display attributes for virtual technology. Through the adapted variable-centered approach, we applied this taxonomy across different display technology. Specifically, we aimed to address and incrementally measure aspects of visual immersion on co-presence and team membership, allowing for results that are more generalizable across a broader scope of iVR technology.

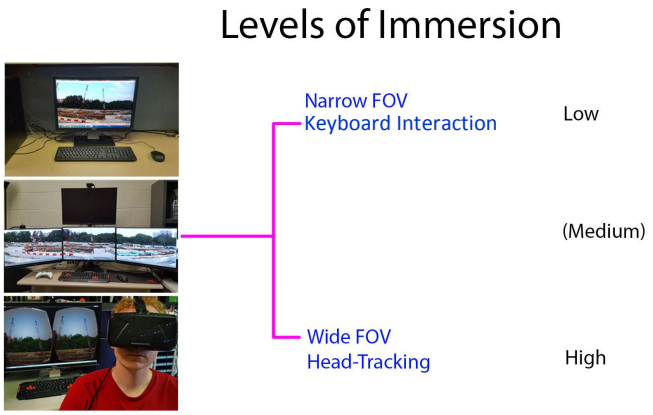


Figure 1. Three levels of immersion (single-screen, multi-screen, Oculus Rift).

For this work-in-progress, we looked at the relative impacts of Field of View (FOV) and basic changing viewpoint interaction, classified as "passive" input (Bowman, Kruijff, LaViola, and Poupyrev 2005: 89). To represent these technology affordances, we used three setups (Figure 1): 1) standard desktop (narrowest FOV and keyboard interaction), 2) multi-screen desktop (wider FOV and keyboard interaction), and 3) Oculus Rift (wider FOV and head-tracking interaction). The same content was provided using a 360° camera (Ricoh Theta S) to help remove any influence of video quality.

We used a (3x2) experimental design with three levels of immersiveness examined across two environments: 1) conference room and 2) outdoor 'field site'. As presence was a subjective measure, we adapted items of spatial presence (Vorderer *et al.* 2004) and co-presence (Slater, Sadagic, Usoh, and Schroeder 2000). Additionally, we looked at team membership and participation through items for team virtuality (Schweitzer and Duxbury 2010) and coding of communication (Gabriel and Maher 2002).

Our collaboration setup used two students with a semi-scripted discussion on placing recycling bins outside of a building on a university campus. The collaboration utilized a combination of a turn-based decision-making with open discussion to attempt to reach consensus. The study site, Figure 2, remained the same for all conditions.

Figure 2. Site for the Collaboration on recycling bin placement.

The collaborators were positioned either in a conference room with PowerPoint photos or outside of the building at the location being discussed. The setup for each location consisted of the Ricoh Theta S, spherical 360° camera, and a phone running Skype for audio over WiFi, Figure 3. After consent was obtained, participants were randomly assigned to one of the three immersive conditions.

Figure 3. Conference Room and Field Site Setup.

3 Results

Our preliminary analysis looked at the relationship between presence measures based on which of the three immersive setups a user experienced. All of our survey items were tested and found reliable. A total of N=90 participants (45 male) from a northeast university acted as a remote third member of a collaboration. With further analysis, we aim to better inform the way remote

collaborators participate in meetings virtually, through an increased understanding of iVR technology solutions. The study will provide important insights into the benefits of currently hyped immersive VR technologies for developing a sense of presence that has the potential to improve both iVR applications and ecological validity of psychological experiments.

Acknowledgements

The authors would like to acknowledge the generous support of the LMI Research Institute's Academic Partnership Program.

4 References

Balakrishnan B and Sundar S, 2011, Where am I? How can I get there? Impact of navigability and narrative transportation on spatial presence. *Human-Computer Interaction*, 26(3): 161-204.

Balakrishnan B, Oprean D, Martin B and Smith M, 2012, Virtual reality: factors determining spatial presence, comprehension and memory. *Proceedings of the 12th International Conference on Construction Applications of Virtual Reality (CONVR), Taipei, Taiwan,* 451-459.

Bowman D A, Kruijff E, LaViola Jr J J and Poupyrev I, 2005, *3D user interfaces: theory and practice*. Addison-Wesley: Boston, MA.

Bowman D A and McMahan R P, 2007, Virtual reality: how much immersion is enough?. *Computer*, 40(7): 36-43.

Gabriel G C and Maher M L, 2002, Coding and modelling communication in architectural collaborative design. *Automation in Construction*, 11: 199-211.

Greeno J G, 1994, Gibson's affordances. *Psychological Review*, 101(2): 336-342.

Nowak K L and Biocca F, 2003, The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments, 12*(5): 481-494.

Robey D, Khoo H M and Powers C, 2000, Situated learning in cross-functional virtual teams. *Technical Communication*, *47*(1): 51-66.

Slater M, 1999, Measuring presence: A response to the Witmer and Singer presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 8(5): 560-565.

Slater M, Sadagic A, Usoh M and Schroeder R, 2000, Small-group behavior in a virtual and real environment: A comparative study. *Presence*,*9*(1): 37-51.

Steuer J, 1992, Defining virtual reality: Dimensions determining telepresence. *Journal of communication*, 42(4): 73-93.

Schweitzer L and Duxbury L, 2010, Conceptualizing and measuring the virtuality of teams. *Information Systems Journal*, *20*(3): 267-295.

Vorderer P, Wirth W, Gouveia F R, Biocca F, Saari T, Jäncke F, …. Jäncke P, 2004, MEC spatial presence questionnaire (MEC-SPQ): Short documentation and instructions for application. Report to the European Community, Project Presence: MEC (IST-2001-37661). Retrieved from http://www.ijk.hmt-hannover.de/presence.

Wilson B G and Myers K M, 2000, Situated cognition in theoretical and practical context. In: D H Jonassen and S M Land (eds), *Theoretical foundations of learning environments*. Mahwah, N.J: L. Erlbaum Associates: 57-88.

Zhao S, 2003, Toward a taxonomy of copresence. *Presence: Teleoperators and Virtual Environments*, *12*(5): 445-455.

A Closer Examination of Spatial-Filter-Based Local Models

Taylor Oshan, A. Stewart Fotheringham

Arizona State University, Tempe, Arizona 85281
Email: {toshan; stewart.fotheringham}@asu.edu

Abstract

Local modeling is a spatial analysis technique that explores spatial non-stationarity in data-generating processes. Geographically weighted regression (GWR) is one method that has been widely applied across domains, which has helped uncover local spatial relationships and generally increases model goodness-of-fit. Despite this, some have criticized GWR for being extra susceptible to issues such as multicollinearity and spatial autocorrelation. Spatial-filtering-based local regression (SFLR) has been suggested as a solution. While SFLR is claimed to be approximately equivalent to GWR, it is also touted as superior. Therefore, it is of interest to compare the output from these two techniques. We do this by examining how well both techniques replicate the known coefficient values derived by simulating data that is representative of spatially varying processes. The results indicate that the original SFLR specification is prone to overfitting, while an alternative specification that minimizes the mean square error produces coefficients similar to GWR.

1. Introduction and Background

Non-stationarity in data-generating processes goes largely undetected in traditional global models. Hence, local models, which explicitly allow regression coefficients to vary over space, are necessary to capture such heterogenous processes. One local modeling technique that has become particularly popular is geographically weighted regression (GWR) (Fotheringham *et al.* 2002). Despite its usefulness, GWR has been critiqued as being highly susceptible to multicollinearity (Wheeler & Tiefelsdorf 2005), whereby multicollinearity amongst explanatory variables causes intolerable levels of correlation amongst GWR coefficients. However, results show that when the sample size is large, GWR is robust to even remarkably high levels of multicollinearity (Páez *et al.* 2011; Fotheringham & Oshan Submitted). Still, the work of Wheeler and Tiefelsdorf (Wheeler & Tiefelsdorf 2005) sparked much subsequent critiques of GWR. Specifically, spatial-filter-based local regression (SFLR) has been suggested as a superior alternative to GWR (Griffith 2008). While Griffith (2008) posits that SFLR and GWR are approximately equivalent, he also points out that local coefficients produced from GWR and SFLR are minimally correlated, suggesting that the two models are producing much different results, thereby implying a contradiction within the SFLR method. To the knowledge of the authors, the SFLR method has not been applied outside its original conception (Griffith 2008), while GWR has produced agreeable results in many studies. Therefore, the primarily goal of this paper is to employ simulated data in order to test which of the techniques can more reliably estimate the true coefficients of non-stationary processes. Some modifications of the SFLR routine are investigated and several issues of the SFLR framework are highlighted.

A basic GWR model may be specified as

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + \varepsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where y_i is the dependent variable at location i, β_{i0} is the intercept coefficient at location i, x_{ik} is the kth explanatory variable at location i, β_{ik} is the kth local regression coefficient for the kth explanatory variable at location i, and ε_i is the random error term associated with location i. A kernel function is applied at each location, which is chosen either using cross-validation or maximizing a model fit criterion, in order to weight nearby observations based on their proximity to the calibration location. Consequently, GWR can generate an ensemble of models by creating local subsets of data.

In contrast, SFLR is based on the interpretation that the eigenvectors of a modified connectivity matrix are the set of possible orthogonal and uncorrelated map patterns (i.e., degree of spatial autocorrelation) (Griffith 1996). Griffith proposes that interaction terms between each explanatory variable and each eigenvector can be utilized to create local coefficients that are approximately equal to those from GWR. Therefore, the following specification is derived

$$y_i = \beta_0 1 + \sum_{p=1}^{P} X_p \cdot 1\beta_p + \sum_{k=1}^{K} E_k \beta_{E_k} + \sum_{p=1}^{P} \sum_{k=1}^{K} X_p \cdot E_k \beta_{pE_k} + \varepsilon \tag{2}$$

where $\beta_0 1$ is the intercept, $\sum_{p=1}^{P} X_p \cdot 1\beta_p$ represents the explanatory variables and their coefficients, $\sum_{k=1}^{K} E_k \beta_{E_k}$ denotes a subset of the possible eigenvectors and their coefficients and $\sum_{p=1}^{P} \sum_{k=1}^{K} X_p \cdot E_k \beta_{pE_k}$ are a subset of the possible interaction terms between explanatory variables and eigenvectors and their coefficients. An important step in the spatial-filter-based framework is selecting a subset of the eigenvectors and interaction terms. The methodology entails a forward stepwise regression variable selection algorithm amongst all interaction terms and eigenvectors based on a statistical significance criterion. For each iteration of the routine, each candidate variable is tested in the model and the one that produces the smallest p-value is selected. At the end of each iteration, any variable that produces a p-value greater than the threshold of 0.1 is removed. The algorithm continues until no variable can be added that produces a p-value less than the threshold, at which point the routine terminates and the local coefficients can be processed.

2. Methods and Results

A dataset with known properties was derived by simulating 2,500 observations for the cells of a 50 by 50 grid. The observations within the cells were generated using the following equation

$$Y = B_{0i} + B_{1i} X_{1i} + B_{2i} X_{2i} + \varepsilon_i \tag{3}$$

where Y are the generated observation, B_0, B_1, and B_2 are *known* locally varying coefficients, X_1 and X_2 are variables drawn form random normal distributions, ε is a random normal error term, and i is the index for each of the 2,500 locations. The spatial distributions of B_0, B_1, and B_2 llustrated in figure 1A display varying levels of non-stationarity, which were created by drawing random values from normal distributions that vary over space and, therefore, are not a result of either the GWR or SFLR specifications. Since these surfaces are based on local processes, they result in MC values that signify very strong spatial autocorrelation (0.982, .947, and .912 for B_0, B_1, and B_2, respectively).

Griffith (2008) previously insinuated that the SFLR framework was superior to GWR based on the fact that the SFLR estimated coefficients "remain unbiased, yield a better global model fit, are polluted with considerably lower levels of spatial autocorrelation, and, for the most part, display little relationship to the GWR coefficients". Therefore, SFLR and GWR were both estimated using the simulated data to compare the resulting coefficient surfaces which can be seen in figures 1B and 1C. It was noticed in figure 2 that the mean square error (MSE)

Table 1: Comparison of MSE, MC, and model fit for models.

	MSE b_0	MSE b_1	MSE b_2	MC b_0	MC b_1	MC b_2	adj. R^2	AIC
GWR	3.239	0.00129	0.00401	0.979	0.994	0.994	0.888	21164.561
SFLR	72.725	0.0130	0.0123	0.651	0.809	0.905	0.946	20116.961
SFLR*	7.498	0.00170	0.00380	1.0132	1.0097	1.00210	0.906	21250.258

of the SFLR coefficients and the known coefficients decreased after about the first 30 terms were selected. However, even as model fit increased, additional model terms increased the MSE, indicating overfitting. Therefore, a new MSE-minimizing stepwise selection criterion (SFLR*) was employed that sought to terminate the algorithm before overfitting occurred. The resulting coefficients are displayed in figure 1D.

Overall, the model fit, coefficient accuracy, and spatial autocorrelation statistics provided in table 1 indicate that SFLR is not approximately equivalent to GWR and that GWR produces less error in estimated coefficients compared to SFLR. In contrast, the SFLR* model produces extremely similar model fit, coefficient estimates, and spatial autocorrelation statistics compared to GWR, though there is no indication of superiority. In fact, model fit statistics that account for model complexity, such as the AIC values presented in table 1 tend to favor GWR (i.e., lower values), since it utilizes far fewer covariates.

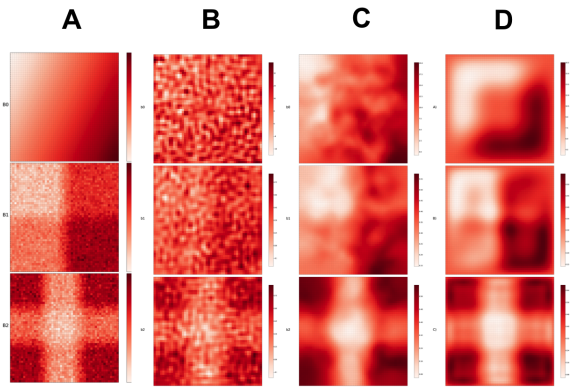


Figure 1: Coefficient surfaces for A) simulated data B) SFLR estimates; C) GWR estimates; D)SFLR* estimates. Top row corresponds to B_0, middle row to B_1 and bottom row to B_2. Low coefficient values are shaded lighter while high values are shaded darker.

3. Discussion & Conclusion

In its original conception, SFLR does not seem to be a competitor of GWR. Using Griffith's (2008) stepwise selection routine produces overfitted models where a severe loss of coefficient accuracy occurs. Griffith was correct in that the SFLR framework and GWR do not produce similar results, though this is only due to an inherent flaw in his methodology that

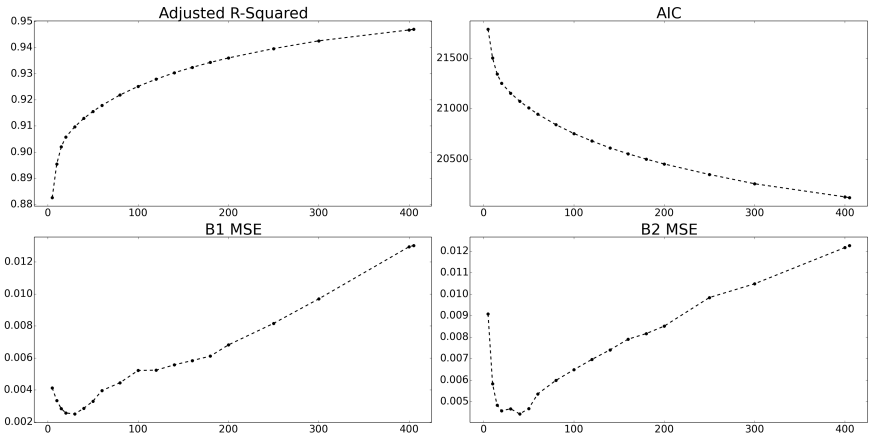


Figure 2: Adjusted R-squared, AIC, and MSE for SFLR model with increasing stepwise selection iterations.

causes overfitting, and can require lengthy compute times of up to several days on a standard notebook computer. It turns out that the SFLR framework *can* be approximately equivalent to GWR when a MSE-minimizing stepwise selection routine (SFLR*) is employed. By doing so, the SFLR* framework produces coefficient estimates that are similar to those from GWR in magnitude, overall accuracy, spatial autocorrelation, and that yield a similar model fit. The computation time is also drastically reduced from days to only minutes, which is comparable to GWR. It seems then that the SFLR* framework provides a sort of discrete spatial weighting mechanism in the form of a subset of eigenvectors and interaction terms, somewhat akin to GWR's continuously defined kernel function weighting mechanism. Given the potential of both GWR and SFLR* to produce such similar results, this provides strong evidence that there is no *a priori* disadvantage to local coefficients displaying strong spatial autocorrelation. Despite the promising results from the SFLR* specification, it is uncertain how to deploy the method when there are no known coefficients because their MSE cannot be computed. Other drawbacks to the SFLR framework include the lack of a means of testing for statistical significance of the estimated coefficients, potential issues of replicability of the methodology, and uncertainty of its robustness to different types of spatial weight matrices.

References

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons.

Fotheringham, A. S., & Oshan, T. (Submitted). Geographically weighted regression and multiocollinearity: dispelling the myth. *Journal of Geograpical Systems*.

Griffith, D. A. (1996). Spatial Autocorrelation and Eigenfunctions Of The Geographic Weights Matrix Accompanying Geo-Referenced Data. *Canadian Geographer / Le GÃl'ographe canadien*, *40*(4), 351–367.

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A*, *40*(11), 2751–2769.

Páez, A., Farber, S., & Wheeler, D. (2011). A Simulation-Based Study of Geographically Weighted Regression as a Method for Investigating Spatially Varying Relationships. *Environment and Planning A*, *43*(12), 2992–3010.

Wheeler, D., & Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, *7*(2), 161–187.

A Function-based model of Place

E. Papadakis[1*], B. Resch[1], T. Blaschke[1]

[1]Dept. of Geoinformatics – Z_GIS, University of Salzburg, Schillerstr. 30, 5020 Salzburg, Austria
Email: {emmanouil.papadakis; bernd.resch; thomas.blaschke}@sbg.ac.at

Abstract

People assign context to space by defining places. Formalizing place enables digital systems to provide a human-centred representation of the geographical world. In this paper we propose a multi-dimensional definition of place including spatial properties, composition and functionality. These dimensions define place as a set of functions, which entail a spatial structure expressed in patterns of spatial descriptions. Relying on this model, it is possible to define places as space infused with functional context by converting geometries to interrelated components that support certain functions.

1. Introduction

There is a long standing conflict between the representation of the geographical world, as it is depicted in the digital and the human world. On the one hand, digital systems understand and analyse space, whereas on the other hand, people perceive and refer to places. Attempting to reconcile these disparate views requires methods that represent the vague and individual-driven concept of place in the rigid and strict formulations of space, and vice versa.

There have been attempts to soften these sharp differences guided by Tuan's definition of place as "space infused with human meaning" (Tuan 1977). There are two general directions in defining the notion of place, either by infusing spatial representations with semantics or by projecting semantics on space.

A leading approach of augmenting space with semantics is the objectification of space (Smith and Mark 2003). According to this, spatial structures are converted into sophisticated objects with ascribed properties, attributing a context to them.

In the opposite way, digital gazetteers (Goodchild 2011) offer a linkage between place names and semantics to spatiotemporal footprints. Finally, the affordance-based model of place (Jordan et al. 1998) focuses on annotating space with context derived from people's actions. Particularly, space, expressed as a set of affordances, is imbued with meaning expressing the ability to serve human intentionality on achieving a final goal.

Most of the aforementioned methods do not fully utilize the expressive power of place. The first two methods associate space with semantics, although it does not always define a human context. On the other hand, the affordance-based model approximates a context; however, affordances provide limited and individual-driven knowledge, defining whether space affords a final goal, which impedes the model's operationalization.

We propose an alternative approach of defining place, adopting the concept of functions and considering that place is space that supports certain functionality. In this way, operationalization is achievable because functions provide an understanding of how and why places operate given a set of properties, avoiding questions of spatial perception. To illustrate this, consider the following example. Suppose we have access to an urban map either without annotations or ones that are incomprehensible due to language barriers. Would it be possible to define the potential location of a shopping centre, using exclusively lines and polygons? Answering this would require to interrelate the spatial organization of places with their functionality.

2. Defining Place

2.1 Background

Place is a product of human thinking, consequently, it can be rightfully considered as an object of discourse, as it was developed by (Couclelis 2010). An object of discourse is whatever we can talk about, regardless of its nature. Four main dimensions describe an object of discourse. The formal dimension focuses on the properties that determine the object's category. The constitutive dimension analyses the composition of the object in terms of associated parts (material or abstract). The agentive dimension identifies the functions of the object relative to a final goal. Finally, the telic dimension refers to the reasons of the object's existence.

Each dimension reveals a semantic resolution level, as it is derived from the semantic contraction process (Couclelis 2010), outlining what is possible to say about the object under consideration. With respect to the notion of place, the derived semantic levels describe its categorization, composition, function and purpose, respectively. Considering that place can be projected on space, the aforementioned levels can be extended with even more coarse-grained semantics, revealing low cognition descriptions such as clarification, perception or even awareness. The aforementioned layers can be viewed as a gradual description of the "platial" structure, from spatial existence and properties to more complicated notions, such as intentionality.

2.1 Function-based Place

Relying on the aforementioned set up of semantic layers we propose a model of place that conforms to the following assumptions. First, our model represents only places that exist in the real world. Second, our approach focuses only on places marked by human intervention and designed for certain goals. Third, we only consider those semantic levels whose information is gained through cognitive process, skipping primitive steps such as perception and awareness of existence.

We propose a multi-faceted definition of place incorporating the dimensions of spatial properties, composition and functions. The dimension of spatial properties reflects the semantic level of classification, describing place as a set of properties (e.g. geometries). The dimension of composition illustrates the constitution level, representing the spatial organization of a place as a composite object formed by simple interrelated components. Finally, the functional perspective points to the semantic level of functions that provides a sense of context by depicting the set of operations that the place supports.

We assume that place is space infused with functionality. The underlying functions suggest a context by describing how a place operates based on its spatial structure. The unique feature of the functional perspective is the implied objectiveness that stems from the dimension of composition. Particularly, it is considered a blueprint of place; it describes how the spatial objects that form a place should be organized in order to enable a desired function. This emphasizes the crucial difference between affordances and functions. Affordances refer to individual perceptions, giving answers on whether an operation can be provided by a place or not. On the contrary, functions driven by the composition process justify why a place supports a specific operation, in terms of strict spatial rules. For instance, in order for an industrial area to be functional, it has to comply with a set of standards. These standards are not products that support affordances, such as characteristics that allow an area to be perceived as industrial. Instead they are strict rules, such as topology, inclusion, and so on, introduced by urban planners and architects, which suggest a spatial design plan that will make an area to operate as industry. The lack of functions' subjectivity facilitates formalization, precision and a degree of inter-subjectivity. It is worth noting that there are

additional perspectives that determine the context of place, such as purposes, feelings or meaning. However, these are out of the scope of the work presented in this paper, but constitute future developments of our approach, as mentioned in the next section.

A place can provide several functions. With respect to the motivating example, a shopping center mainly facilitates shopping and allows re-supplying operations, while it might also provide leisure facilities. Based on our model, we propose that a set of functions, which are supported by a certain type of place, composes a functional class. This class categorizes places according to their major and minor functions, forming its functionality.

In order for the proposed model of place to support domain independence, its formalization should be flexible, reusable and extensible. An inspiring solution is the ontology design pattern (Gangemi 2005), which introduces the model of place as a self-contained building block able to be integrated into other ontologies. As stated by Gangemi (2005), an ontology pattern is discovered and refined through a set of competency questions. In our case, these questions should indicatively focus on function-based search of places, semantic annotation, reasoning, and so on. The complete definition of the ontological design pattern is currently under preparation. An initial version of the model of place is shown in Figure 1.

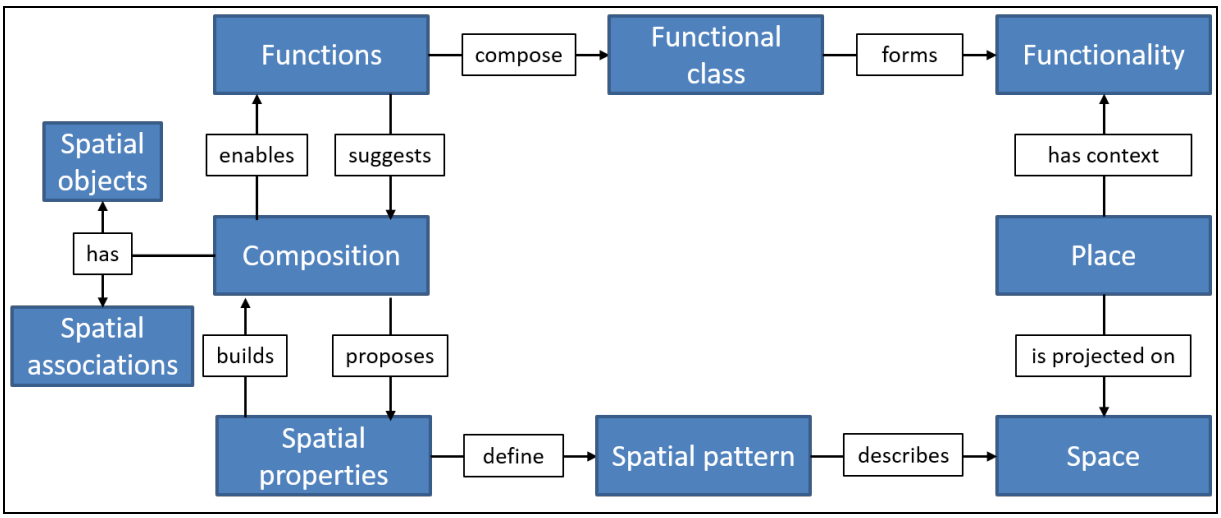


Figure 1. Function-based model of Place.

In the sequel, we provide an example of spatial design and function-based search of place that illustrates the operationalization of the proposed model. Spatial design is the process of defining an archetypal composition and hence a set of spatial properties, which shape a spatial pattern based on a set of functions. Based on our motivating example, the shopping center should be located away from unpleasant industrial areas and close to major roads, facilitating re-supplying operations. Furthermore, it should occupy a minimum profitable space that allows shoppers to move freely and also provides separate places for its re-supply operations. Optionally, it should include a parking area along with other facilities such as restaurants and cinemas that attract customers.

Rule sets such as the aforementioned one suggest archetypal compositions for places that are described by the corresponding functional class. Such compositions reveal spatial patterns by analyzing the comprised components and their associations into geometries and hence spatial properties, as shown in Figure 2a. Having obtained these spatial patterns we can answer the question of the motivating example as follows. Space can be infused with a functional context by converting spatial properties to interrelated components that conform to a composition, which supports shopping center's functionality, as shown in Figure 2b.

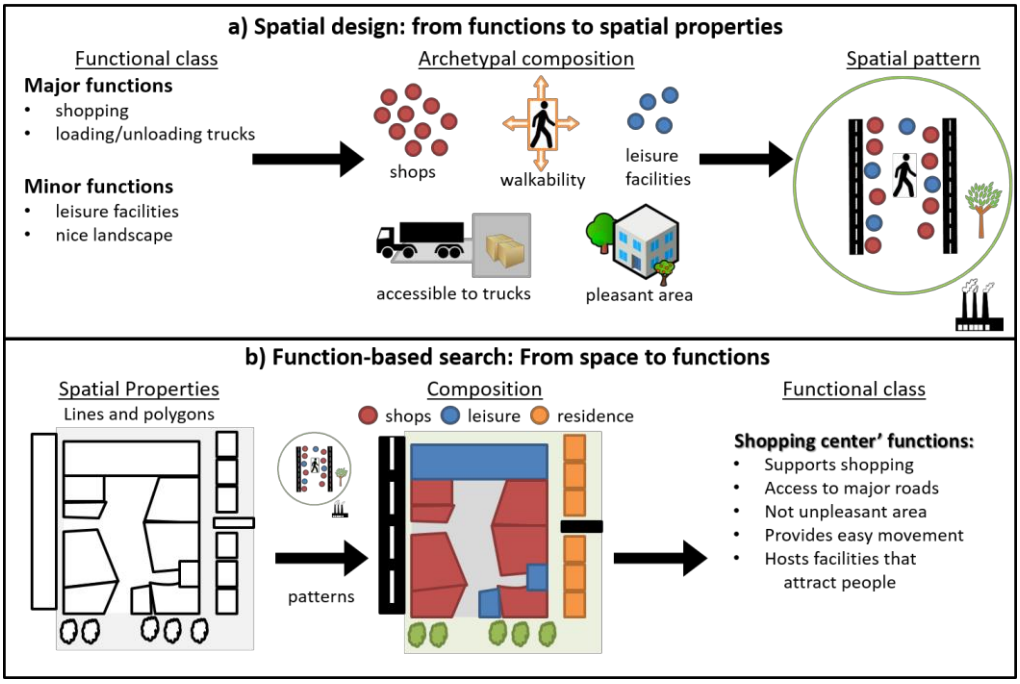


Figure 2. Interrelating functions and space.

3. Conclusions and Future work

In this work we provide an ontological model of place, based on the notion of functions. We propose a multi-faceted definition of place including the dimensions of spatial properties, composition and functions. These dimensions define place as a set of functions that entail an archetypal spatial structure. This model introduces the extraction of spatial patterns of places and the infusion of space with a functional context by converting geometries to interrelated components that support certain functionalities.

Future research includes the formalization of the proposed model by defining a representative ontological design pattern of place, and then evaluating its applicability on spatial design and function-based search. Additionally, we plan to extend the existing definition of place with the perspective of intentionality describing how places' functionality serves humans' purposes. Furthermore, we will also focus on the sentimental perspective describing how people associate places with emotions. The integration of these perspectives will provide insights on adding an element of subjectivity in the current definition of place.

Acknowledgements

The presented work is framed within the Doctoral College GIScience (DK W 1237N23), funded by the Austrian Science Fund (FWF).

References

Smith B, Mark, D M, 2003, Do mountains exist? Towards an ontology of landforms. In: Environment and Planning B: Planning and Design, 30(3):411-427

Couclelis H, 2010, Ontologies of geographic information. *International Journal of Geographical Information Science*, 24(12):1785-1809.

Goodchild M F, 2011, *Communities, neighbourhoods, and health: expanding the boundaries of place*, Springer, 21-33.

Jordan T, Raubal M, Gartrell B and Egenhofer M, 1998, An affordance-based model of place in GIS. In: *Proceedings of the 8th International Symposium on Spatial Data Handling*, 98:98-109.

Tuan Y-F, 1977, *Space and place: the perspective of experience*. University of Minnesota Press.

Gangemi A, 2005, Ontology design patterns for semantic web content. In: International semantic web conference. Springer, 262-276.

Dynamic Optimization of Phenomena Mapping in GIS

H. Pham[1], A. Ruas[1], T. Libourel[2]

[1] IFSTTAR, France
Email: {ha.pham; anne.ruas}@ifsttar.fr

[2] University of Montpellier, France

Abstract

Pollutions and urban climate are becoming a global issue. However the cartographic quality of related maps is not as expected. The visualization of phenomenon without its context or with inappropriate density might be incomprehensible for non-experts. The purpose of our research is to propose methods for phenomena mapping to facilitate map interpretation. In this paper, we generate adapted data for each level of detail (LoD) from raw data to visualize phenomena at different visual scales, add expert's knowledge into the visualization and readjust phenomenon data according to view conditions for a better visual perception.

1. Introduction

Pollutions and urban climate are becoming a global issue. However pollution and temperature are fields of values that are complex to estimate and to represent. In most cases, phenomena mapping has two deficiencies: 1- the phenomenon is not contextualized or overlaps its context 2- the information density is never adapted to user requirement. The aim of our research is to propose a better visualization of these phenomena, firstly by mapping phenomena with contextual (urban) data to estimate their impact or dangerousness. As phenomena are per-se complex, they should be viewed at different LoDs, from different angles. The understanding of phenomena requires exploration capacities. This research thus completes research in visual analytics (Andrienko *et al*. 2013).

We propose two ways to represent a phenomenon (Pham *et al*. 2015):

- With grids where a grid is composed by a set of punctual or area symbols. Grids allow seeing the phenomenon with its context because the symbols do not cover all the space.
- With plan, composed by a set of cells, each cell being a polygon represented by a color. Plan is used for small scale to give an overview of the phenomenon.

We propose to change the data density according to the zoom and to readjust the symbolization to the contextual information such as buildings. In order to do that, we create data for mapping by adjusting grid step, type of portrayal (grid or plan), grid orientation and we optimize symbolization. In the following section we propose a data model for phenomena mapping. Then we detail our optimization process in section 3 and conclude in section 4.

2. Model to facilitate the representation of data fields

Figure 1 presents a data model to facilitate the phenomenon visualization. In this model we distinguish *raw data* from *data for mapping* which are computed from raw data.

A *phenomenon* is represented by a series of *phenomenon episodes*. An *Episode* is described by a set of *value fields* from an initial time to a final time. For example we describe a Pollution Episode in London from 1.1.2016 to 1.3.2016, with one value field every hour. A *value field* is described by a set of values. Each *value* is associated to a *node* that belongs to a *grid*. In order to improve the visualization, we propose to generate a set of ***value fields for mapping*** because raw data is not suitable for all LoD. Each *value field for mapping* (vfm) is characterized by an area, a LoD, a color family, a cell size, - generated by generalization or

interpolation and - mapped in a resolution range. A vfm is represented on an ***adapted geometric structure*** which can be a plan (LoD1) or a grid according to requirements and zooming. In our research the LoD depends on the area extend and the thematic data granularity (the raw data cell size) on one side and, on the other side, the smallest reasonable cell size to explore the data. For air pollution, the smallest LoD would allow seeing Paris and the largest would allow navigating in a street (see figures 3 to 5)

To readjust the visualization to a view condition, we create a *temporary adapted value field* (t-avf) extracted from a vfm. This t-avf is automatically modified every time we zoom in/out or change observer's position.

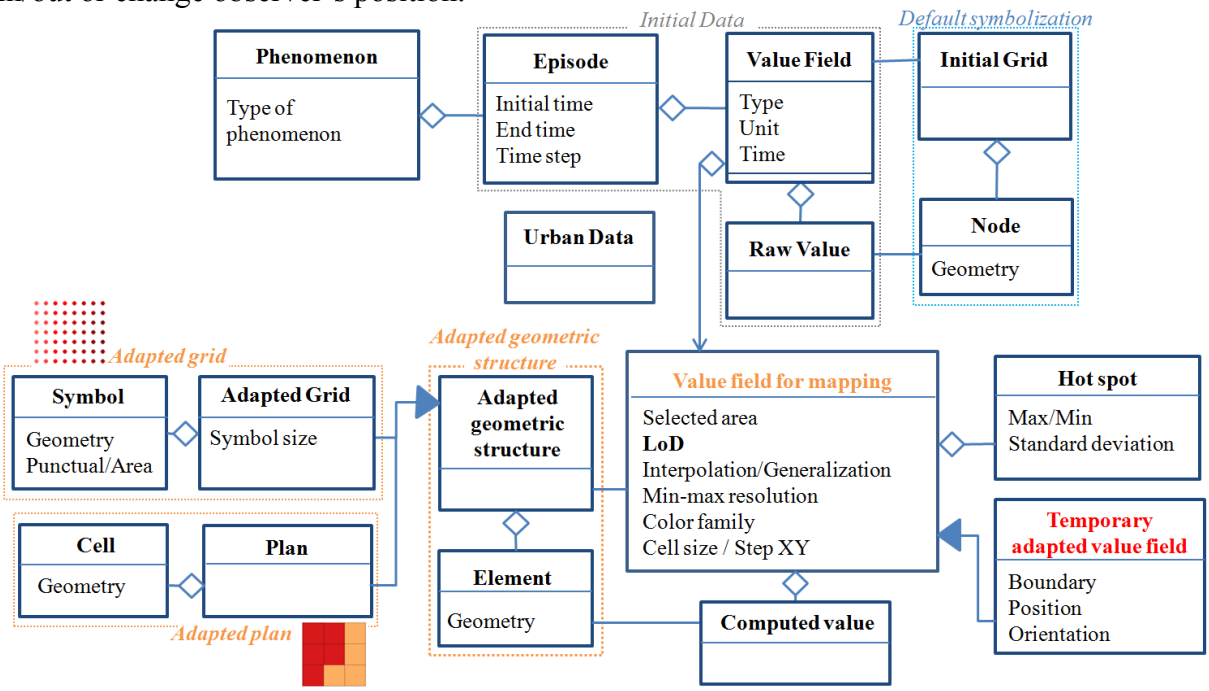


Figure 1: Model to facilitate the representation of data fields

3. Process to optimize phenomena mapping

We are in the process of writing a plug-in in QGIS to allow the automatic optimization of phenomena mapping. Once a user chooses an area, the plug-in automatically optimizes the corresponding information in the database and reloads the optimized data. We describe the optimization process hereafter (Figure 2).

3.1. Preliminary optimization

This preliminary process consists in preparing LoDs by generalization or interpolation (Pham *et al.*, 2015) that will be used for the interactive optimization. It is based on two steps: data analysis and creation of appropriate LoDs.

Step 1: Analyze data

From user's requirements, scale of phenomenon and its context, we define the necessary quantity of LoDs for the phenomenon's interpretation, maximum resolution and minimum resolution for each LoD.

We also detect the *hot spot* by looking either for the extreme values on the area or for most variables values (important standard deviation). Extreme values are important to make an alert when values exceed a safety threshold.

Step 2: Create LoDs

To represent a phenomenon at different LoDs, we create new data tables for each LoD:

- Grids to visualize the phenomenon with its geographic context: at least one grid for punctual symbols (low LoD), and one grid for area symbols (high LoD). Area symbols are very efficient for large zoom (Ruas *et al*. 2015)

- One plan to visualize the phenomenon continuously, which gives an overview of the phenomenon. Plan is perfect to represent our LoD1.

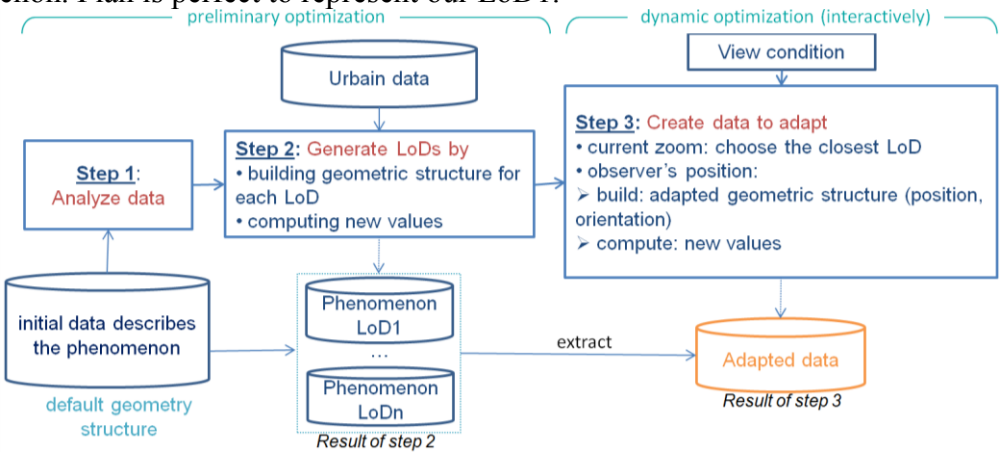


Figure 2: Principe of optimization

3.2. Dynamic optimization

To optimize the visualization of a phenomenon, we propose the adjustment of grid parameters (density, position, orientation) in accordance with each view condition defined by zoom, environment or observer's position.

Step 3: Adapt the grid for view condition

- Adapt grid density for current zoom

The grid density depends on the distance between two nodes (grid step). When we zoom in, we wish to view the phenomenon with more details. So we can create a data table for each possible LoD. But if we store too many tables in the databases, the step 2 becomes time consuming (cf 3.1.). On the other hand, building one or two LoDs is fast but the visualization may be rough. We propose to make the grid density dynamic in between stored LoDs. Thus anytime we change the zoom, the system automatically chooses the closest LoD among the tables saved in the database (step 2) and computes the new grid from the chosen one (figure 2). Current research aims to propose efficient methods for this step.

- Adapt grid orientation for environment

The grid orientation is defined by the angle between the axis of the grid and the abscissa. Optimizing consists here to find the orientation that minimizes the obscure rate of the grid in an area, or to maximize the number of non-overlapped nodes.

When an area is selected by a user, we count the number of intersections between the phenomenon data and the background (here the buildings). The obscure rate is measured by the ratio between the quantity of nodes hidden by buildings and the total ones: $R_{ob} = \frac{N_{hidden}}{N_{total}}$

If the rate of obscure is less than 10%, we validate the orientation. Otherwise, we vary the orientation angle from 0 to 45°, the best position is where R_{ob} is the smallest.

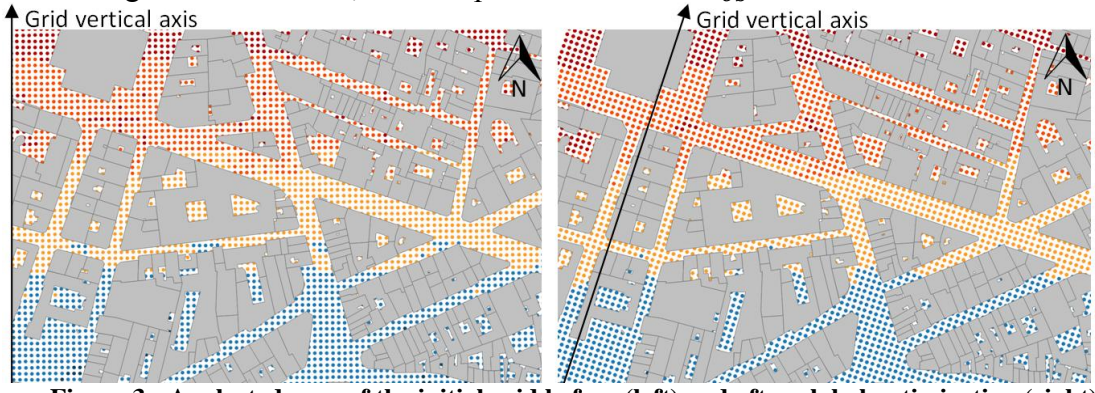


Figure 3: A selected area of the initial grid before (left) and after global optimization (right)

- Adapt grid for observer's position

We propose to create a visual field to simulate the point of view of a human that walks in the geographical space. The grid readjusts according to the observer's position.

Figure 4: Visualization with point of view of an observer

If user focuses on one street, we orient the grid according to this street in order to maximize the number of visibles nodes.

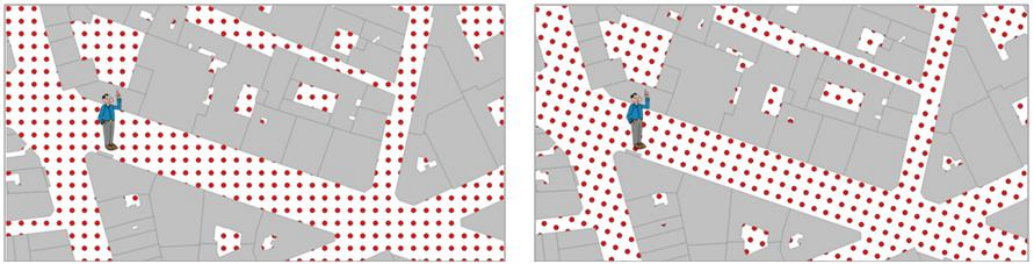


Figure 5: Initial grid before (left) and after optimization according to the street orientation where observer is located (right)

Conclusion

In this paper we propose methods for optimizing visual perception by visualizing a phenomenon at different LoDs and adapting the representation to view condition. Our current work is to complete optimization methods including symbolization. Future research will focus on evaluation.

References

Andrienko N, Andrienko G, Gatalsky P, 2003, Exploratory spatio-temporal visualization: an analytical view. *Journal of Visual Languages and Computing*, 14(6):503-541

Pham H, Ruas A, Libourel T, 2015, Representing Urban Phenomena in Their Context and at Different LoD: from Raw Data to Appropriate, *3rd Eurographics Workshop on Urban Data Modelling and Visualisation*, Delf, Neitherland.

Ruas A, 2015, From a phenomenon to its perception: models and methods to represent and explore phenomena on GIS, *Modern Trends in Cartography Springer LNGC*, ISBN 978-3-319-07926-4, 259-268

Ruas A, Pham H, 2015, Symbolization and Generalization to Map Water Pipe Data Flow and Water Quality at Different Scales, *The Cartographic Journal*, 52(2):149-158

Thomas J J and Cook K A(eds), 2005, *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, IEEE CS Press.

A spatial mixture model to account for risk discontinuities: Analyzing attempted suicide in Waterloo Region, Ontario

M. Quick[1], J. Law[1,2]

[1]School of Planning, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1
Email: {mquick; jane.law}@uwaterloo.ca

[2]School of Public Health and Health Systems, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1

Abstract

Bayesian spatial analysis of small-area data often models spatially structured random effects via the intrinsic conditional autoregressive prior distribution (ICAR). For outcomes exhibiting abrupt variations in risk between adjacent areas, spatial smoothing imposed by ICAR may challenge the identification of high risk areas. This research explores a mixture model that weights spatially structured and unstructured random effects, increasing sensitivity to risk discontinuities. The outcome analyzed is attempted suicide in the Region of Waterloo, Ontario, Canada. Compared to a conventional non-mixture Bayesian spatial model, diagnostics suggest that the mixture model has superior fit. Maps of spatially structured random effects from mixture and non-mixture models highlight specific small-areas where the mixture model captures greater levels of residual spatial structure.

1. Introduction

Spatial analyses of areal health data that apply Bayesian models often include unstructured and spatially structured random effects parameters to address methodological challenges including spatial autocorrelation and overdispersion (Best *et al.* 2005; Latouche *et al.* 2007; Haining *et al.* 2009). Spatially structured random effects capture residual spatial autocorrelation and are commonly modeled via the intrinsic conditional autoregressive distribution (ICAR) (Besag *et al.* 1991). Interpretation of spatially structured random effects as a surrogate for unmeasured geographically-specific characteristics or processes is similar to the spatial error model in frequentist spatial regression, however spatially structured random effects are full probability distributions estimated for each small-area (Anselin 2002; Law and Chan 2011).

Using ICAR, area-specific spatially structured random effects are often smoothed to the mean value of spatially structured random effects in adjacent small-areas (Richardson *et al.* 2004; Congdon 2007). Past research has suggested that when observed outcomes exhibit abrupt variations between adjacent small-areas, for example at the rural-urban boundary, smoothed estimates challenge the identification of high risk areas (Lawson and Clark, 2002; Congdon 2007). This research applies a Bayesian spatial model with continuous mixture parameter that weights the relative contribution of spatially structured and unstructured random effects for each small-area, as applied to attempted suicide in the Region of Waterloo, Ontario, Canada.

2. Study region and data

In 2006, the Regional Municipality of Waterloo was composed of 720 census dissemination areas (DAs) and a population of 475,351. Census data from 2006 data was analyzed due to limited data availability for rural DAs in the 2011 Canadian Census (Bell and Wei 2014). Waterloo Region is composed of three urban municipalities: Waterloo, Kitchener, and Cambridge, as well as four rural townships (Figure 1). Attempted suicide counts were obtained

from Waterloo Regional Police Services for four years (2011 to 2014) and expected counts were calculated based on residential population (Table 1). Abrupt variations occur at the urban-rural boundary, where DAs with higher observed and expected counts are located in urban municipalities (Figure 1). This is consistent with past research exploring both fatal and attempted suicides at the small-area scale (Hempstead 2006; Cheung *et al.* 2012).

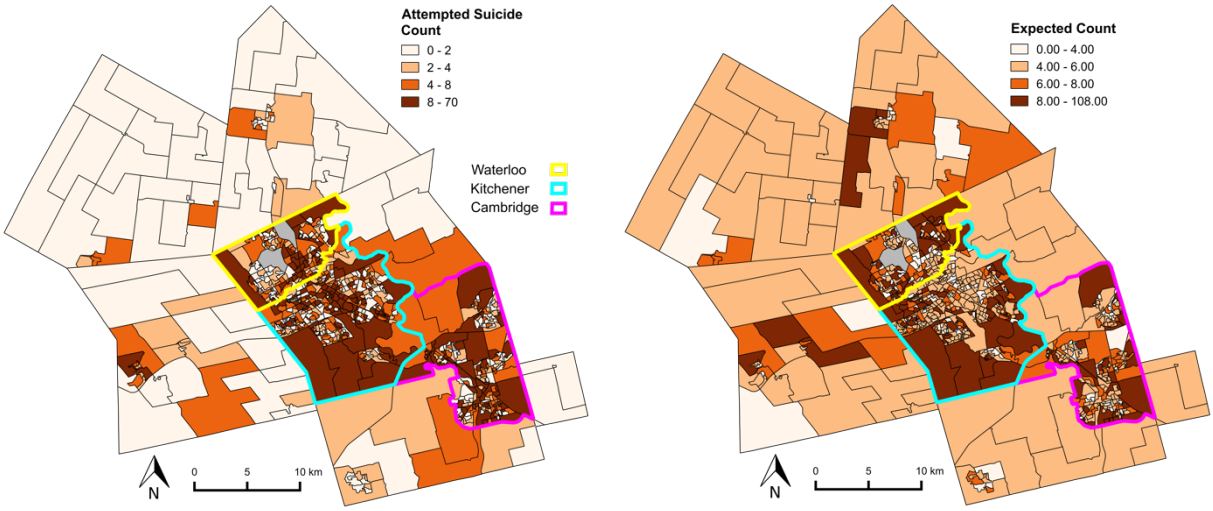


Figure 1. Observed (left) and expected count (right) of attempted suicide. Areas in grey were excluded due to incomplete data.

Past spatial analyses have found that small-area suicide risk is associated with three dimensions of the physical and social environment: rurality, social deprivation, and social fragmentation (Hempstead 2006; Rezaeian *et al.* 2007; Congdon 2011). Operationalizing these dimensions were dwelling density, percent of low-income families, and percent of one-person households, respectively (Table 1).

Table 1. Descriptive statistics of attempted suicide count and covariates.

	Mean	Std. Dev.	Min	Max
Attempted suicide count	6.16	8.22	0	70
Expected suicide count	6.16	5.37	2.24	107.03
Dwelling density (dwellings per km^2)	1,147.32	965.64	2.98	8,814.14
Low-income families (%)	5.27	6.73	0.00	37.50
One-person households (%)	21.39	13.45	0.00	68.75

3. Spatial mixture model

Attempted suicide count (O_i) for DAs i (1, …, 720) are Poisson distributed (Model 1). The Poisson distribution is often used for rare count data at the small-area scale (Militino *et al.* 2001; Richardson *et al.* 2004).

$$O_i \sim Poisson(\mu_i) \quad (1)$$

Model 2 shows the common Bayesian spatial model with expected count (E_i), overall risk (α), unstructured random effects (u_i), spatial random effects (s_i) and three covariates ($\beta_1 x_{1i}$: dwelling density, $\beta_2 x_{2i}$: low-income families, $\beta_3 x_{3i}$: one-person households) (Besag *et al.* 1991; Latouche *et al.* 1997). Covariates were standardized for model convergence. The prior for α is an improper flat distribution. Priors for β_1, β_2, and β_3 are non-informative *Normal*(0, 10000), for u_i is *Normal*(0, τ_u), and for s_i is the ICAR distribution with variance τ_s (Besag *et al.* 1991).

Spatial adjacency was defined to include all adjacent areas. Non-informative *Inverse-Gamma*(0.5, 0.0005) distributions were assigned for variance parameters τ_u and τ_s (Kelsall and Wakefield 1999).

$$\log(\mu_i) = \log(E_i) + \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i + s_i \quad (2)$$

Model 3 extends Model 2 with a mixture scheme, where u_i and s_i are weighted according to mixing parameter π_i. This model has been applied in past research (Congdon, 2007). The prior distribution for π_i is *Beta*(1,1), which constrains $1 < \pi_i > 0$. For groups of adjacent areas with dissimilar risk, π_i may take a value close to one, emphasizing unstructured random effects, whereas for groups of adjacent areas with similar risk, π_i may take a small value and emphasize spatial smoothing. Note that a variety of mixture schemes are possible in this modeling framework, including discrete mixtures that involve model switching based on, for example, observed outcome counts (i.e., zero-inflated models (Neelon *et al.* 2013)), or continuous mixtures of Gaussian and non-Gaussian spatially structured random effects (Lawson and Clark 2002).

$$\log(\mu_i) = \log(E_i) + \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + (\pi_i) u_i + (1-\pi_i) s_i \quad (3)$$

Convergence for Models 2 and 3 occurred after 50,000 iterations and posterior inferences were drawn from an additional 25,000 iterations. Fit was assessed via the Deviance Information Criterion (DIC), which is a penalized goodness of fit measure that penalizes model complexity (Spiegelhalter *et al.* 2002). Smaller DIC values indicate superior model fit.

4. Results

Model 3 demonstrates better model fit (DIC = 3,247.54) than Model 2 (DIC = 3,404.30). despite the additional mixture parameter. In Model 3, all covariates were associated with attempted suicide; dwelling density was found to be negatively associated (β_1 = -0.24 (95% Credible Interval: -0.32, -0.16)), while the percent of low-income families (β_2 = 0.12 (0.05, 0.19)) and the percent of one-person households were positively associated with attempted suicide risk (β_3 = 0.43 (0.34, 0.52)).

Spatially structured random effect estimates from Model 2 (s_i) and Model 3 ($(1-\pi_i)s_i$) are shown in Figure 2. In general, estimates from both Models were similar through most of the study region. Larger values from Model 3 than Model 2, observed in south Kitchener, south Cambridge, and in rural DAs in the west, can be attributed to a small π_is and suggest stronger residual spatial autocorrelation among adjacent areas. Smaller values, on the other hand, such as those in southeast Cambridge along the rural-urban boundary, can be attributed to large π_is and relatively less residual spatial structure.

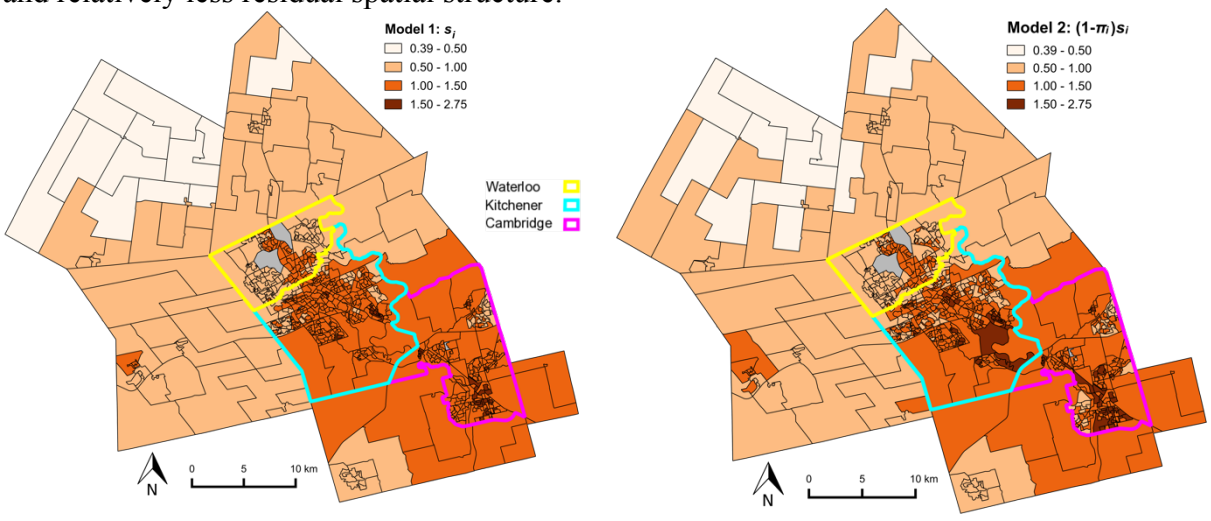


Figure 2. Spatially structured random effects from Model 2 (s_i) (right) and from Model 3 ($(1-\pi_i)s_i$) (left).

This research has applied a Bayesian spatial mixture model to analyse attempted suicide at the small-area scale in the Region of Waterloo, Ontario, Canada. The mixture scheme is continuous, which allows for the relative weight of spatially structured random effects and unstructured random effects to be adaptive to the abrupt variations in risk between adjacent small-areas. Future research should explore the geographic patterns of π_i, further investigate model fit by evaluating area-specific deviance residuals and associated posterior probability of residuals exceeding threshold values, and consider the role of spatially-varying coefficients in accommodating discontinuous risk (Richardson et al., 2004).

Acknowledgements

The authors acknowledge funding from the Social Sciences and Humanities Research Council of Canada Grant 767-2013-1540. The authors thank the Waterloo Regional Police Service for providing data.

References

Anselin L, 2002, Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27:247-267.

Bell S and Wei T, 2014, Mapping the spatial pattern of the uncertain data: A comparison of global non-response rate (GNR) between metropolitan and non-metropolitan area in Ontario. Spatial Knowledge and Information - Canada.

Besag J, York J, and Mollie A, 1991, Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1-59.

Best N, Richardson S, and Thomson A, 2005, A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14: 35-59.

Cheung YTD, Spittal MJ, Pirkis J, and Yip PSF, 2012, Spatial analysis of suicide mortality in Australia: Investigation of metropolitan rural-remote differentials of suicide risk across states/territories. *Social Science and Medicine*, 75:1460-1468.

Congdon P, 2007, Mixtures of spatial and unstructured effects for spatially discontinuous health outcomes. *Computational Statistics and Data Analysis*, 51:3197-3212.

Congdon P, 2011, The spatial pattern of suicide in the US in relation to deprivation, fragmentation and rurality. *Urban Studies*, 48(10):2010-2122.

Haining R, Law J, and Griffith D, 2009, Modelling small area counts in the presence of overdispersion and spatial autocorrelation. *Computational Statistics and Data Analysis* 53:2923-2937.

Hempstead K, 2006, The geography of self-injury: Spatial patterns in attempted and completed suicide. *Social Science and Medicine*, 62:3186-3196.

Kelsall JE and Wakefield JC, 1999, Discussion of 'Bayesian models for spatially correlated disease and exposure data' by Best NG, Arnold RA, Thomas A, Waller L, and Conlon EM. In Bernardo JM, Berger JO, Dawid P, and Smith AFM *Bayesian Statistics 6*:131-156. Oxford University Press, Oxford, United Kingdom.

Latouche A, Guihenneuc-Jouyaux C, Girard C, and Hemon D, 2007, Robustness of the BYM model in absence of spatial variation in the residuals. *International Journal of Health Geographics*, 6: 39.

Law J and Chan PW, 2011, Monitoring residual spatial patterns using Bayesian hierarchical spatial modelling for exploring unknown risk factors. *Transactions in GIS*, 15(4):521-540.

Lawson AB and Clark A, 2002, Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine,* 21:359-370.

Militino AF, Ugarte MD, and Dean CB, 2001, The use of mixture models for identifying high risks in disease mapping. *Statistics in Medicine* 20:2035-2049.

Neelon B, Ghosh P, and Loebs PF, 2013, A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society*, 176(2):389-413.

Rezaeian M, Dunn G, St Leger S, and Appleby L, 2007, Do hot spots of deprivation predict the rates of suicide within London boroughs? *Health and Place*, 13:886-893.

Richardson S, Thomson A, Best N, and Elliott P, 2004, Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112:1016-1025.

Spiegelhalter D, Best NG, Carlin BP, and van der Linde A, 2002, Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, 64(4):583-639.

Comparing Geospatial Ontologies with Indigenous Conceptualizations of Time

Geneviève Reid [1], Renee Sieber[1]
[1]McGill University, Department of Geography, Burnside Hall Building, Room 705, 805 Sherbrook Street West, Montreal, Quebec H3A 0B9
Email: genevieve.reid@mail.mcgill.ca
renee.sieber@mcgill.ca

Abstract

The geographic domain has widely been studied in ontology research. However, integrating the conceptualization of time and temporal referencing of geographic concepts in data models is a complex task that has by no means been "solved". Existing geospatial ontologies have adopted a space-time model that, for example, distinguishes *endurant* entities (lasting through time, e.g., fixed natural features) from *perdurant* entities (e.g., processes or events). Such a model might exclude indigenous conceptualizations of time that are far more sophisticated. We find that conventional ontologies make assumptions about time that fail to take into consideration indigenous notions including: 1. Time is not linear; 2. Nothing is completely fixed in time; 3. Time has agency; and 4. Time is not temporal but social.

1. Time and Geospatial Ontologies

In GIScience, indigenous conceptualizations of space and time have been depicted as being in direct opposition to those used to design geospatial technologies (Rundstrom 1995; Veland *et al.* 2014). Indeed, geospatial technologies emphasize a more static view of the world that is often inconsistent with indigenous perspectives on space and time. Compared with geographic features, notions of time have received less attention both in geospatial ontologies and in indigenous ontologies research. Time and temporal referencing of geographic concepts are nonetheless challenging to geospatial ontologies applied in indigenous contexts. Including indigenous conceptualizations in geospatial ontologies and in the Geospatial Semantic Web is crucial. Wellen and Sieber (2013) argue that developing an inclusive semantic interoperability is not only possible, but also critical to ensure future accessibility of geospatial technologies for indigenous communities, and minimize loss and misinterpretation of information when geospatial ontologies are used to record indigenous knowledge.

1.1 Endurant/Perdurant Model

An important contribution to the conceptualization of time in geospatial ontology development, was the distinction between *endurant* objects that endure through time and *perdurant* objects that happen in a certain time (e.g., processes or events) (Agarwal 2005). Grenon and Smith (2004) propose a spatio-temporal ontology of change and processes called SNAP/SPAN based on the duality *endurant/perdurant*.

The SNAP/SPAN model distinguishes *endurant* entities, which have spatial properties, from *perdurant* entities, which have temporal properties. Temporal intervals and instants describe *perdurant* entities through linear time. Even though the SNAP/SPAN distinction is widely adopted in geospatial ontologies (Agarwal 2005), philosophical assumptions about time behind this model can fundamentally differ from indigenous conceptualizations.

2. Indigenous Concepts of Time

2.1 Time is not linear

Indigenous conceptions of time can be more complex than a linear passage of time from the past, to the present, and towards the future. Time could be viewed as a spiral, a branch, a triangle, or a cycle. Figure 1 shows these different representations that time could take.

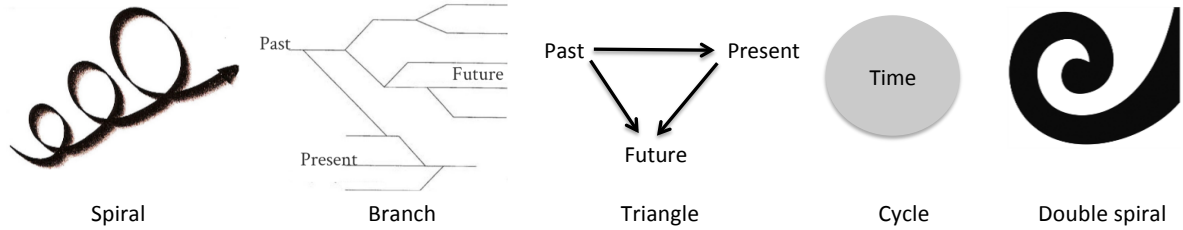


Figure 1: What is time?

In many Native American cultures, the conceptualization of cyclical time predominates (Fixico 2003). Time, as a cycle and as a circle, does not 'go' anywhere. Rather than following a direction, time is conceived as circular processes, including the diurnal, solar, lunar, and seasonal cycles. Events and activities are understood as part of these daily, seasonal, and annual cycles. Our research sees that Eastern Cree hunting practices are tied to the notion of cycles: cycles of returning animals, cycles of resting the land to restore animals' habitat, and cycles of seasons affecting the animals' behaviors and movements across the land (Berkes 2012; Preston 2002). Indigenous conceptualizations of cyclic and seasonal time are complex; many cycles are interwoven together (e.g., life cycles of plants or animals indicates life stages in other species) and linked to language and spiritual notions (e.g., cause and effects of changes) (Lantz and Turner 2003).

In the Māori tradition in New Zealand, the concept of time is represented in the Koru symbol of the double spiral where "each circumambulation of the spiral incorporates the past into both the present and the future and, in doing so, reconstitutes both" (Murton 2011: 82). In the Lake Titicaca area of South America, Aymara language and culture has a unique conception of time. Contrary to most conceptions, in Aymara, the past – which is known– is conceived in front of people, where it is visible; whereas the unknown future is in the back (Núñez and Sweetser 2006).

A triangle conceptualization of time emphasizes the direct relationship between past and future. The importance of that connection is often expressed by the Eastern Cree of Wemindji in Northern Quebec. When talking about future aspirations, people often directly make a connection to the past without any references to the present, as voiced by a participant: " [In the future] I want Wemindji to look like it was in the past" (Elder woman, Focus group, Wemindji 2013).

A structure with multiple branches emerging from the past to the present and branching off again towards the future represents the multiplicities of stories in the past and multiple alternate scenarios for the future. This branching time structure is an unsolved issue in computerized data models (Ott and Swiaczny 2001). To date, geospatial ontologies are ill equipped to deal with the branching time model or with other non-linear conceptualization of time such as the cycle, spiral or triangle (multidirectional) time structures.

2.2 Nothing is "endurant" or completely fixed in time

In indigenous contexts, temporally fixing elements of landscape as endurants might be problematic. Scholars have found that Australian indigenous storytelling processes about places proved the categorization of features as endurant through time to be inadequate

(Veland *et al.* 2014). For example, placenames–rather than being fixed–represent ephemeral expressions emerging through a narrative process (ibid.).

Creation stories are part of many indigenous cultures (McGregor 2004). These stories about how everything came to be on Earth provide an understanding of the existence of geographic features. For example, the Eastern Cree have a story of the Wolverine that was sprayed in the eyes by the Big Skunk, and walked all the way to the coast to wash up. This creation story explains how the water in the bay became salty and rivers and lakes inland remained fresh water (Preston 2002: 159-163). Other legend stories from our ongoing research in Wemindji explain how natural features, such as mountains, hills, lakes, and rivers, were formed (stories audio recorded in the community in 1984 by Luke Shashaweskum). For example, part of a longer story explains that a hill was formed and got its shape when a Giant dropped his cooking pail while being pursued and killed by a Shaman (Ronnie Georgekish, Tallyman VC22, Interview, April 23, 2016).

2.3 Time has agency

Anthropologists have widely studied the notion of agency and natural features (e.g., Cruikshank 2010; Hallowell 2002). They describe how indigenous ontologies often conceptualize the land itself and the elements constituting the land (geographic entities such as mountains, rivers, islands, trees; natural phenomena such as wind, thunder; sun, moon; animals and other spirits present on the land) as living beings, filled with spiritual powers and can all be considered as 'persons'. Even though the notion of agency for time has been less explored, it represents nonetheless an important challenge to geospatial ontologies.

The Runa communities of Ecuador hold a notion of a 'living future' (Kohn 2013). Events, activities and practices of everyday life are interlinked and influenced by a future that manifests itself through relationships among humans, animals, nature, and more-than-human beings (Kohn 2013). For the Eastern Cree, the notion of a 'living past' is often expressed. For example, a young Cree woman activist explained in a TV interview that, for Cree people, each step one takes forward is supported by a thousand ancestors (Maïtée Labrecque-Saganash in Radio Canada 2016). People in Wemindji often mention how spirits of ancestors are part of the land and guide the hunters. Furthermore, Wemindji's Wellness & Culture Department staff often refers to the Cree Nation of Wemindji's slogan: "A Community Where Tradition Lives On" (www.wemindji.ca). The notions of a 'living past' and a 'living future' emphasize the roles of time in actually influencing and affecting events, activities, behaviors, and relationships for humans, animals, nature, and more-than-human beings.

2.4 Time is not temporal but social

Space-time data models conceptualize *perdurant* entities (processes) as they unfold through a temporal interval (Grenon and Smith 2004). These 'time-based' time intervals can be problematic when the concept of time is not perceived as independent from events and objects.

Sinha *et al.* (2011) show that in Amondawa culture and language the concept 'time' as an abstract domain independent of the events that occur 'in time' does not exist. For Amondawa people, time is not based on countable units but based upon the interplay between ecological facts in the natural environment and social structures (ibid.). The social structure of time is based on the rhythms of working activities and the stages of life. Instead of indicating the passage of time with nominal age of people, Amondawa people change their proper names to indicate the transition in stage of life, the kinship and the role in the family or in the community. Amondawa time intervals are event-based and social, rather than 'time-based'.

3. Conclusion

Excluding indigenous conceptualizations from geospatial ontologies and from the Geospatial Semantic Web puts indigenous communities at greater risk than they already are of losing their knowledge or having it stripped of significance (Wellen and Sieber 2013). Rather than being exclusionary, developments towards semantic interoperability can be (and ought to be) inclusive (ibid.). However, conventional geospatial ontologies fail to take into consideration indigenous conceptualizations of time. Further research should allow the integration of indigenous notions such as 1. Time is not linear; 2. Nothing is completely fixed in time; 3. Time has agency; and 4. Time is not temporal but social.

Acknowledgements

We owe our gratitude to the Cree Nation of Wemindji and to Wemindji community members.

References

Agarwal P, 2005, Ontological considerations in GIScience. *International Journal of Geographical Information Science*, *19*(5): 501–536.

Berkes F, 2012, *Sacred Ecology* (Third ed.). Routledge, New York; London.

Cruikshank J, 2010, *Do Glaciers Listen?: Local Knowledge, Colonial Encounters, and Social Imagination*. UBC Press, Vancouver; Toronto.

Fixico D, 2013, *The American Indian mind in a linear world: American Indian studies and traditional knowledge*. Routledge, New York.

Grenon P, and Smith B, 2004, SNAP and SPAN: Towards Dynamic Spatial Ontology. *Spatial Cognition & Computation*, *4*(1): 69–104.

Hallowell A. I, 2002, Ojibwa ontology, behavior, and world view. *Readings in Indigenous Religions*, *22:*17-42.

Kohn E, 2013, *How Forests Think: Toward an Anthropology Beyond the Human*. Univ of California Press, Berkeley; Los Angeles; London.

Lantz T. C, and Turner, N. J, 2003, Traditional phenological knowledge of Aboriginal peoples in British Columbia. *Journal of Ethnobiology, 23*(2): 263–286.

McGregor D, 2004, Traditional Ecological Knowledge and sustainable development: Towards Coexistence. In M. Blaser, H. A. Feit, & G. McRae (Eds.), *In the Way of Development: Indigenous People, Life Projects, and Globalization*, Zed Books; International Development Research Centre, London; New York; Ottawa, 92–110.

Murton B, 2011, Embedded in place "Mirror knowledge" and "simultaneous landscapes.". In D. M., Andrew G. Turk, Niclas Burenhult and David Stea Mark (Ed.), *Landscape in Language: Transdisciplinary perspectives,* John Benjamins Publishing Company, Amsterdam; Philadelphia, 73–100.

Núñez R. E, and Sweetser E, 2006, With the Future Behind Them: Convergent Evidence From Aymara Language and Gesture in the Crosslinguistic Comparison of Spatial Construals of Time. *Cognitive Science*, *30*(3): 401–450.

Ott T, and Swiaczny F, 2001, *Time-integrative geographic information systems: Management and Analysis of Spatio-Temporal Data*. Springer, Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo.

Preston R. J, 2002, *Cree narrative: expressing the personal meanings of events* (2nd ed.). McGill-Queen's Press, Montreal; Kingston; London: Ithaca.

Radio Canada, 2016, Vivre de fierté et d'espoir. *Tout le monde en parle*, http://ici.radio-canada.ca/tele/tout-le-monde-en-parle/2015-2016/segments/entrevue/6448/samian-maitee-labrecque-saganash

Rundstrom R. A, 1995, GIS, Indigenous Peoples, and Epistemological Diversity. *Cartography and Geographic Information Science*, *22*(1): 45–57.

Sinha C, Sinha V. D. S, Zinken J, and Sampaio W, 2011, When time is not space: the social and linguistic construction of time intervals and temporal event relations in an Amazonian culture. *Language and Cognition*, *3*(1): 137–169.

Veland S, Lynch A, Bischoff-Mattson Z, Joachim L. E. E, and Johnson N, 2014, All Strings Attached: Negotiating Relationships of Geographic Information Science. *Geographical Research*, *52*(3): 296–308.

Wellen C. C, and Sieber, R. E, 2013, Toward an inclusive semantic interoperability: the case of Cree hydrographic features. *International Journal of Geographical Information Science*, *27*(1): 168–191.

Deriving Hospital Catchment Areas from Mobile Phone Data

Bernd Resch[1,2], Azmat Arif[1], Gautier Krings[3], Guillaume Vankeerberghen[3], Marc Buekenhout[4]

[1]University of Salzburg, Department of Geoinformatics – Z_GIS, Schillerstrasse 30, 5020 Salzburg, Austria
Email: bernd.resch@sbg.ac.at, azmat.arif@stud.sbg.ac.at

[2]Harvard University, Center for Geographical Analysis, 1737 Cambridge Street, Cambridge, MA 02138, USA

[3]Real Impact Analytics, 5 Place du Champ de Mars, 1050 Brussels, Belgium
Email: {gautier.krings; guillaume.vankeerberghen}@realimpactanalytics.com

[4]Email: mbuekenhout@msn.com

Abstract

Delineating catchment areas of medical facilities is essential for estimating the quality of a health-care system and to maximise the efficiency of health service provision. One critical shortcoming of previous approaches are manifested in their comprehensive assumptions about a hospital's patients by using census data or gravity models. In contrast, our approach uses anonymised mobile and landline phone data to derive hospital catchment areas. Our goal is not to assess the quality of the health care system, but to identify the geographic areas, in which people actually use a hospital. Thus, our results reveal new insights into the catchment areas of hospitals by minimising assumptions about demographic factors.

1. Introduction and Related Work

Adequate provision of health services is a central priority of health professionals and policy makers worldwide. More, the efficiency of health care systems, i.e., the provision of best possible service with minimum resources, is critical to public providers (Fransen et al. 2015). These requirements have led to a number of studies to analyse catchment areas and service quality of medical facilities. Even though geospatial analysis methods have existed for decades, there is still a general lack of studies that have mapped and examined health service catchments in practice, not only from a theoretical viewpoint (Schuurman et al. 2006).

Previous approaches for delineating medical catchment areas comprise statistical population-to-provider ratios, gravitational models, travel cost estimation, analysis of the physical distance between hospitals, census-based patient-origin analysis, commuter-based approaches of modelling spatial accessibility (Fransen et al. 2015; Wang and Wheeler 2015), or the two-step floating catchment area method based on the physician-to-population ratio (Luo and Wang 2003). The major drawback of these approaches is that they make far-reaching assumptions about a hospital's patients by applying census data, travel times or gravity models. Moreover, they do not take heterogeneous activity and mobility patterns into account or only derive them from static census data.

Thus, the approach proposed in this paper uses anonymised mobile and landline phone data to delineate hospital catchment areas. Like this, we aim to identify the geographic areas, in which people use a hospital instead of assessing the quality of the health care system per se. Therefore, we analyse calls to and from hospitals in Trinidad and Tobago. This goes beyond just using patient records in that we are able to draw conclusions from a wider range of communication with a hospital (enquiries, arrangement of appointments, follow-up care, visitors, etc.), beyond patients' hospital stays.

2. Data Sources

For retrieving the **hospital polygons**, we manually digitised 117 hospitals in Trinidad and Tobago using Open Street Map (OSM).

The **cell phone network data**, which was provided by a national operator with a market share of more than 50%, contain locations of the cells, including the antennas' mounting angles. The data also contain the properties of each **call** (originating cell, destination cell, duration, etc.). Overall, we used more than 4,000 antennas and about 2.5 billion calls over a period of six months. The call data are fully anonymised, i.e., no conclusions can be drawn to individual subscribers.

In addition to the operator's data, we used network data from **OpenCellID** (http://opencellid.org), a collaborative open-source project collecting data about mobile cells around the world. OpenCellID enrich our provider's data by additional cell locations and the estimated range of each antenna.

3. Methodology and Results

3.1 Determining Service Areas for Mobile Cells

To determine the service areas for each antenna, we use the following parameters: location and range (in meters) from OpenCellID, direction (mounting angle) from the cell phone network operator, and azimuth (beam width).

3.2 Identifying Calls to Hospital Landline Numbers

Next, we identify the calls that were made from and to hospital landline numbers to and from the cell phone network through a simple attribute join on the hospitals' phone numbers. The geometric structure of the resulting graph resembles a star-like pattern with a hospital in its centre, as expected. Through mapping the calls, we can distinguish between the different levels of spatial influence of the hospitals.

3.3 Identifying Calls to and from Cell Phones Located within a Hospital

Then, we determine calls that are made from and to cell phones which are potentially located within a hospital. Therefore, we model each antenna's radiation pattern as a 3-dimensional distribution curve, which reflects the fact that the probability of a cell phone actually being connected to an antenna decreases with distance to the antenna. We approximate the model from Buvaneswari et al. (2007) by using ellipses, as shown in Figure 1.

We developed the following six-step algorithm. 1.) for each cell, we calculate the centre of the ellipse, which represents the **antenna's coverage** given its location and range. 2.) for each cell, we create **ellipse segments** representing a cell's decreasing coverage over distance, according to the coverage model by Buvaneswari et al. (2007). 3.) we orient the ellipses using the antenna's azimuth value, and intersect them with the opening angle of the cell (90°). 4.) we assign a **proportional percentage of the cell's calls to each segment**, which represents the probability that a call has been made in a segment. Therefore, we use the well-known path loss formula $L=10*n*log(d)$, which characterises the decrease of the receivable power radiated by an antenna with increasing distance (d). In the formula, n is the free space component, which is five in our setting – a mixture between urban, indoor and rural places. 5.) we intersect the ellipse segments with the **hospital polygons**. 6.) we calculate the **number of calls in a cell**, which are made to and from cell phones located within the hospital using a spatial join and subsequent summation.

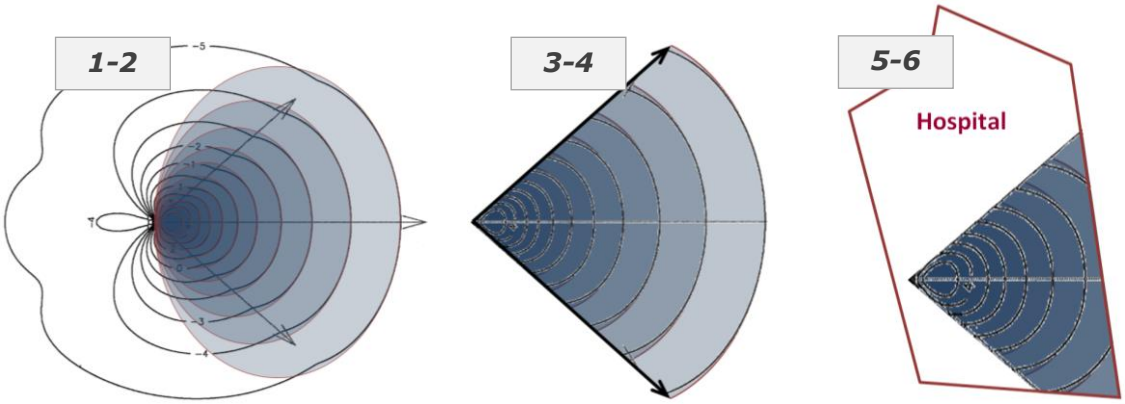


Figure 1: Identifying Calls from and to Cell Phones in a Hospital.

3.4 Computing the Catchment Areas for each Hospital

In the next step, we determine the catchment area of each hospital, where a hospital is identified using landline calls and cell phone calls (s. sub-sections 3.2 and 3.3). First, we compute the probability that a cell phone call has been made from or to a hospital to or from a mobile phone in a remote cell. This is done by identifying distinct origin-destination pairs for each call – a hospital and a remote cell. Then, we use a regular grid (cell size 1x1 km) to compute the number of calls made to and from mobile phones located within each grid cell, analogously to step 6 described above. Finally, we compute the catchment area for each hospital by summing up for each grid cell the number of calls to and from a hospital. Figure 2 shows an exemplary catchment area for one hospital.

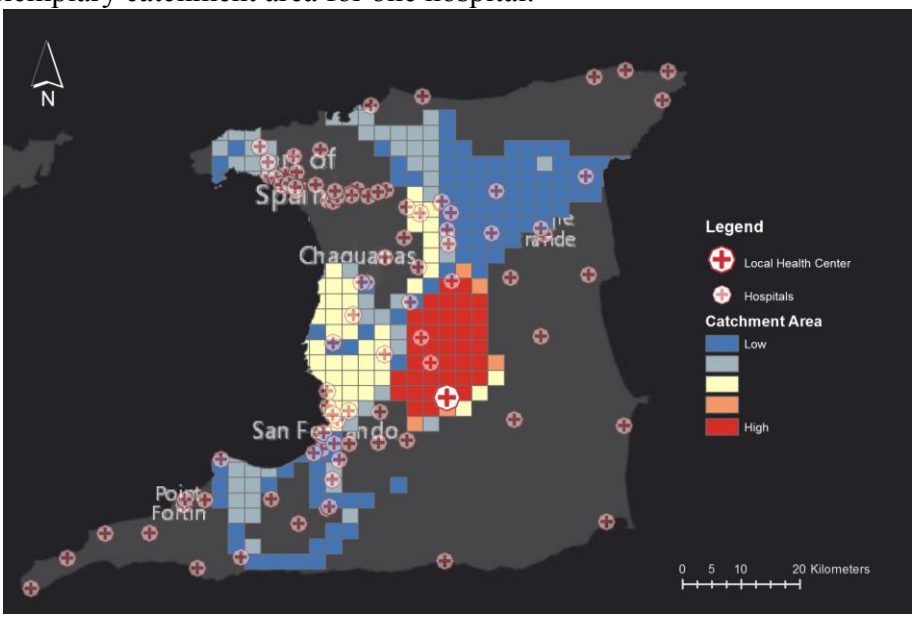


Figure 2: Catchment Area for a Hospital.

4. Discussion and Conclusion

This paper presents an approach to derive the catchment areas of hospitals from anonymised mobile and landline phone data. In contrast to previous approaches, we identify the geographic areas, in which people actually use a hospital. Therefore, we analyse calls to and from hospitals. Our results reveal new insights into the catchment areas of hospitals by minimising assumptions about demographic factors. Even though our case study uses data from Trinidad and Tobago, our approach is transferable to other regions worldwide.

In performing our research, we identified a number of limitations. First, not all calls that are made or received within a hospital are necessarily related to the hospital's "business", which induces an unknown bias. Moreover, we have not accounted for the mobile cells' spatial density in our analysis. Thus, our current method does not consider that the probability of a cell phone being connected to a particular antenna decreases if many cells' coverage areas overlap. Furthermore, the use of grid cells in the calculation of the catchment areas may lead to a modifiable areal unit problem (MAUP). This could potentially be mitigated by intersecting all ellipse segments, which was too computationally intensive given our extensive dataset. Finally, evaluation and validation of our results are difficult as no data for ground-truthing or comparable studies exist.

References

Buvaneswari A, James DA, Liu C, Graybeal JM, Lambert D and MacDonald WM, 2007, A Statistical View of the Transient Signals That Support a Wireless Call. *Technometrics*, *49*(3):305–317.

Fransen K, Neutens T, De Maeyer P and Deruyter G, 2015, A Commuter-based Two-step Floating Catchment Area Method for Measuring Spatial Accessibility of Daycare Centers. *Health and Place*, *32*:65–73.

Luo W and Wang F, 2003, Measures of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region. *Environment and Planning B: Planning and Design*, *30*(6):865–884.

Schuurman N, Fiedler RS, Grzybowski SCW and Grund D, 2006, Defining Rational Hospital Catchments for Non-urban Areas Based on Travel-time. *International Journal of Health Geographics*, *5*(1):43.

Wang A and Wheeler DC, 2015, Catchment Area Analysis Using Bayesian Regression Modeling. *Cancer Informatics*, *14*(Suppl 2):71–79.

Space-Time Topological Graphs

C. Robertson[1]

[1]Wilfrid Laurier University, 75 University Ave, Waterloo, ON
Email: crobertson@wlu.ca

Abstract

The integration of space and time in geospatial analyses is an active research area. While methods for characterizing sequences of spatial point objects over time have evolved greatly over recent years, the treatment of space-time change in complex geographic objects such as linear features and polygons has seen much less attention. In this paper we consider the directed graph as a representation for temporal change in spatial objects, where edges are formed by spatial relationships. With this representation, the large foundation of graph-theoretic metrics and algorithms can be used for the analysis of space-time change in geographic objects. We present preliminary findings using several exemplar datasets: range expansion of polygons representing animal home ranges, the spatial path of Hurricane Katrina, and a forest insect outbreak. In particular, we investigate the dynamics of space-time change in these datasets using well-known metrics for characterizing graph structures.

1. Introduction

Understanding space-time change continues to be a challenging and increasingly required task for geospatial analysis, as more spatial datasets are evolving into long-term records of spatial and temporal change (e.g, satellite archives, long-term radio collar studies). However, models of space-time change remain to be fulll developed. Stell et al. (2011) introduced a bigraph representation of space-time change, whereby objects were represented as nodes at discrete time points, and linked through relations. As well, Del Mondo et al. 2012 further developed this model with an explicit spatiotemporal graph representation, distinguishing between filial relations (based on explicit identity) and purely topological relations (derived from spatial relationships). We build on these ideas from the perspective of space-time analysis in order to analyze spatial changes in polygon objects over time. We use several space-time polygon datasets in order to explore a how space-time topological graphs can be used as a representation for spatial-temporal analysis generally.

1.1 Polygon models of space-time change

Sadahiro and Umemura (2001) introduced a framework for the analysis of spatio-temporal polygon distributions based on events derived from spatial overlap relations in neighbouring time-steps. The scope of events was extended in Robertson et al. (2007) to include proximity-based spatial relations, describing various types of movement occurring over the change interval, known as spatial-temporal analysis of moving polygons (STAMP). The STAMP framework is temporally discrete, such that an appropriate temporal grain is required a priori in order to represent change. Metrics associated with space-time change include counts of event types, area change, directional change, all of which can be accumulated along paths of the graph representing continuous spatial relationships over time. In Sadahiro (2001) the graph representation of polygon change events was introduced. In this paper we build on this framework order to integrate existing graph metrics into the spatial-temporal analysis of polygon distributions.

2. Methods

2.1 Space-time topological graphs

We define the space-time topological graph g by the set of edges E and vertices V representing polygons where each edge e_{ij} represents a spatial relationship between polygons from neighbouring time periods i and j (Figure 1). Since edges are determined by time, g is both directed and acyclic, while edges can be either binary or weighted. Each change interval generates a bipartite graph, yielding a k-partite graph for the full set of time periods. We could conceive of g as a set of dynamic spatial weights matrices. The properties of g in this context have not been explored in great detail since its inception in Sadahiro (2001), which examined changes in polygon area and event types. We therefore examine some properties of g derived from different datasets with the aim of exploring g as a general representation for spatial-temporal analysis. Weights were derived as the ratio of T1∩T2 area to the T2 area in each polygon event-grouping. In general, weights can be tailored to the application and spatial representation being used.

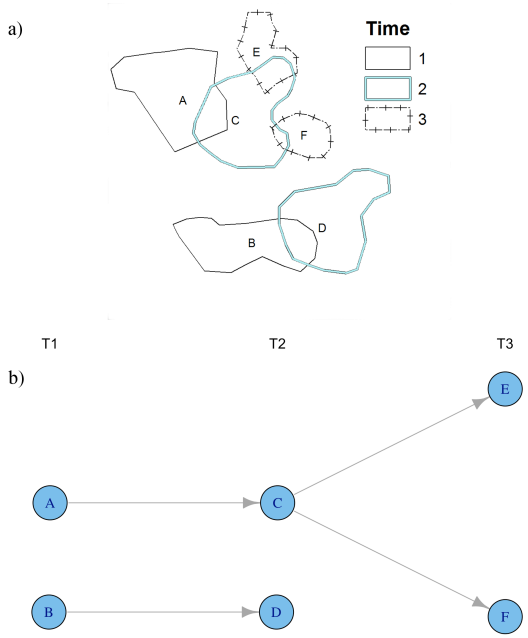


Figure 1. a) Space-time polygon distribution and b) its representation as a space-time topological graph.

2.2 Datasets

Three polygon datasets were used to investigate g based on the STAMP event framework (Robertson et al. 2007) as outlined in Table 1. The forest insect (mountain pine beetle – MPB) outbreak data were obtained for the Morice Timber Supply Area in west-central British Columbia. These data represent hotspots derived from thresholded kernel density estimates of points representing clusters of attacked trees, during the initial years of the mountain pine beetle outbreak that occurred in British Columbia. The caribou range polygons represent winter seasonal home ranges derived from radio-collared caribou in the Bathurst herd. Finally, the Hurricane Katrina dataset was obtained from US NOAA Hurricane Research Division's H*Wind product and polygons represent 39 miles per hour isotachs (contours derived from wind speed fields) at 3 hour intervals. Each of these represents significant environmental change events at least in part due to climate change, and phenomena for which better spatial forecasting models are needed.

Table 1. Space-time polygon exemplar datasets of environmental change.

Dataset	Temporal Scale	N	Location
MPB	Annual (1995 - 2003)	711	British Columbia, Canada
Caribou	Annual (1996 - 2014)	36	Northwest Territories, Canada
Katrina	3 hours (Aug 25 21:00 - Aug 29 21:00)	33	Southeast USA

2.3 Analysis

Analysis of graph data typically describes properties of the graph as a whole, its edges, or its nodes. Depending on the application, different properties may be of interest. For example, measures of centrality are often used to identify important nodes in social networks (i.e., super nodes) that are connected to a disproportionate number of nodes in the network. We used three well-known metrics for exploring node centrality, connectivity, and community/clustering of *g* in our space-time polygon datasets (Table 2).

Table 2. Graph metrics used to explore space-time topological graphs.

Metric	Level	Property	Description
Betweenness centrality	Node	Node importance	Number of shortest paths from all nodes to all others that pass through that node
Degree Assortativity	Graph	Topology	Similarity in number of edges among connected nodes
WalkTrap	Node	Community / cluster detection	Modularity maximizing algorithm based on random walks from each node

3. Results

The caribou data had a graph size of 48 edges and 32 nodes, and a density of 0.048, while the MPB data exhibited 613 edges, 557 nodes, and density of 0.002. The Katrina graph was much less complex, as splitting and merging events did not occur, yielding a graph with 45 edges, 33 nodes, and density of 0.043. These graphs along with their clusters detected from the WalkTrap algorithm are presented in Figure 2. The results obtained from the community detection methods for all three datasets provided interesting partitions for identifying sub-periods and polygon groups that align with our understanding of these spatio-temporal processes. Graph metrics showed expected patterns with the MPB and Katrina having more spatial-temporal centrality, and higher similarity in node degree (representing topological relations between polygons).

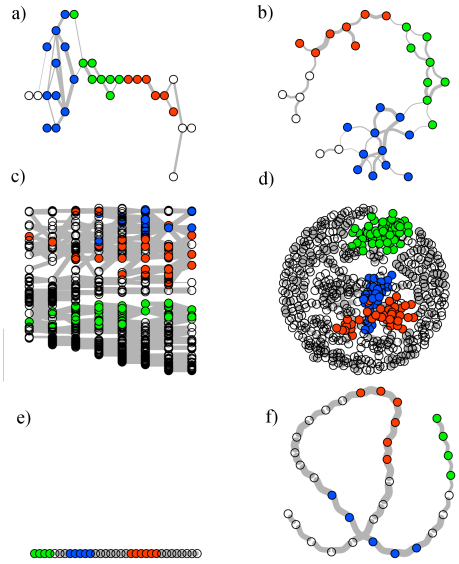


Figure 2. Graphs and cluster nodes (coloured) for a-b) caribou home ranges, c-d) forest insect outbreaks, and e-f) Hurricane Katrina. Left plots have temporal ordering maintained, right use the Fruchterman & Reingold (1991) graph layout algorithm.

Table 3. Graph metrics used to explore space-time topological graphs.

Metric	Caribou	MPB	Katrina
Betweenness centrality	0.08188	0.00004	0.16667
Degree Assortativity	0.03534	-0.16276	1.0

4. Conclusions

While interpreting the patterns of these analyses is beyond the scope of this short paper, the results highlight the potential utility in using mathematical graphs as general models for spatial-temporal analysis and relating spatial processes to their space-time topology. In particular, the community detection of clusters of nodes relating to important events in the space-time process may help to characterize signatures of environmental change events as they arise. Further research into which metrics are ideal in different contexts and the ways spatial measures can be integrated as edge weights remain areas to explore.

References

DelMondo, G, Rodriguez, MA, Claramunt, C, Bravo, L, and Thibaud, R, 2013, Modeling consistency of spatio-temporal graphs. *Data & Knowledge Engineering*, 84: 59-80.

Fruchterman, TM and Reingold, EM, 1991, Graph drawing by force-directed placement. *Software: Practice and experience*, 21:1129-1164.

Robertson C, Nelson TA, Boots B and Wulder MA, 2007, STAMP: Spatial-temporal analysis of moving polygons. *Journal of Geographical Systems*, 9:207–227.

Sadahiro Y and Umemura M, 2001, A computational approach for the analysis of changes in polygon distributions. *Journal of Geographical Systems,* 3, 137-154.

Sadahiro Y, 2001, Exploratory Method for Analyzing Changes in Polygon Distributions. *Environment and Planning B: Planning and Design*, 28: 595–609.

Stell, J, DelMondo, G, Thibaud, R, and Claramunt, C, 2011, Spatio-temporal evolution as bigraph dynamics. *International Conference on Spatial Information Theory*. Springer Berlin Heidelberg.

A Multidirectional Optimal Ecotope-Based Algorithm to Delineate a Commuter Shed

D. Schleith, M. J. Widener

University of Cincinnati, 401 Braunstein Hall,
Cincinnati, OH 45221-0131
Email: schleidk@mail.uc.edu

University of Toronto, Sidney Smith Hall, 100 St. George St, Room 5037,
Toronto, ON M5S 3G3
Email: michael.widener@utoronto.ca

1. Introduction

In commuting research the geographic area under investigation is of crucial importance. When examining commutes occurring in a region of interest, the selection and use of different city, county, or metropolitan region boundaries will have a large impact on analyses of travel times and distances, whether a transit network provides adequate access to jobs, levels of congestion, and so on. This is closely linked to the spatial form of cities (especially in the North American context) where a relatively dense city is surrounded by suburbs with progressively lower densities. Determining what actually constitutes a commuting region (or "commuter shed") is typically a matter of using administrative boundaries prescribed by the U.S. Census Bureau. In general though, the metropolitan region is often used because it represents a big enough area to capture most of the economic activity occurring inside. The issue with metropolitan boundaries, however, is summarized in Morrill et al. (1999), "… metropolitan areas are widely recognized as far from consistent in meaning or adequate in definition." The problem is largely attributed to the use of counties as building blocks. Counties that are selected to comprise a metropolitan region are those neighboring the county or counties containing the largest principal city. The neighboring counties are included if they are socially and economically connected to the principal county, as measured by the number of commuters coming into the central county (Office of Management and Budget, 2010). Counties have a large spatial extent, and oftentimes include vast rural spaces with little relationship to the urbanized area of interest to many researchers. A method for providing a more precise measure is warranted.

In many ways a commuter shed is like a cluster of commuting activity, where there are significant links between residents moving between relevant, contiguous zones. Here, we use a cluster detection method to delineate the commuter sheds of the counties that make up the Miami, FL metro region. We take census tracts as the building blocks to provide a more precise representation of the commuter shed, and test the spatial interaction of these tracts using the percentage of commutes into the various zones and an advanced spatial clustering statistic.

2. Relevant Literature

As previously mentioned, commuter sheds are the de facto analysis areas of most commuting research, the results of which are sensitive to the definition of the study area. Researchers are therefor interested in these definitions, as they attempt to accurately describe settlement patterns across the country and provide a reasonable assessment of how people move within an urban region. The current method used by the US Census examines the "percent of commutes" in a county to the nearest county containing a central city. So, an outlying county is included if it has at least 15% of its commuters working in that central county (or counties). But an outlying county can only be assigned to one central county and the determination is made based on commutes from the possible central counties added to commutes to the central county.

Generally, this effort is to determine the location of suburbs of various urban areas and which rural places are not exurban suburbs but places unto themselves.

3. Methods

This research expands on Morrill et al.'s work by using a novel cluster detection method to delineate the commuter shed of Miami, Florida. The method is called A Multidirectional Optimal Ecotope-Based Algorithm or AMOEBA, developed by Aldstadt and Getis (2006). AMOEBA uses the Gi* statistic to determine spatial interactions (Getis & Ord, 1992). The method calculates Gi* for all of the zones in a region iteratively determining whether to include a zone in the cluster based on the mean Gi* score. The variable used here to determine the presence of a cluster is the percentage of the commutes that terminate within an area of interest. To test the sensitivity of defining commuter sheds in a multi-core region using a clustering analysis, as the identification of clusters will be subject to the number of total tracts included, we use study areas that include tracts within 60-, 90-, 150- and 180-mile buffers of Miami-Dade, Broward, and Palm Beach counties, as well as the Miami Urbanized Area. Finally, as the high-resolution commuting data are stratified by a number of interesting categories (described below), we additionally run the AMOEBA method on these different sets of workers to see how commuter sheds vary.

4. Data

The data used here are origin-destination pairs from the Census' LEHD dataset, referred to also as the LEHD Origin-Destination Employment Statistics (LODES). These data are available annually for most of the U.S. between 2002 – 2014. Several commuting studies have made use of the LEHD dataset like Horner et al. (2015), demonstrating the utility of high-resolution commuter data.

The OD data files contain counts of the number of workers making trips between census blocks, however, in order to execute the computation of the clustering statistic in a reasonable amount of time, these counts are aggregated up to the census tract level. These OD files include counts stratified by three income categories, three age categories, and three industry classification categories. The LODES data also distinguishes between primary and secondary employment, so only primary employment is considered in this study. As administrative boundaries often change over time, the LEHD data program has resolved the 2002-2009 data to the corresponding 2010 blocks.

5. Results

As mentioned previously, the U.S. Census Bureau uses commute trips into the county or counties that contain the urban area. Figure 1 shows the census tracts within 90 miles of Miami-Dade County and the percentage of commutes that terminate within Miami-Dade County.

Figure 2 shows the results of the AMOEBA method applied to the city of Miami, Miami-Dade county, the Miami urbanized area, and finally the entire three county metro region, each with a 150-mile buffer. With the clustering results of all counties displayed the overall cluster appears to match well with the Census Bureau's delineation (all three counties). What is interesting here is that, with the exception of a little overlap at the boundary of each county, each cluster is relatively relegated to the county it comes from. Each of these counties is home to a relatively large city and the clustering suggests that the Miami metro region might more accurately be described as three metropolitan regions.

All together these clusters present a novel way to depict an area's commuter shed by improving on the current U.S. Census Bureau definition. Using a cluster detection method with high-resolution data, as we do here, removes many of the issues that arise when defining

unique commuter sheds related to particular city. Table 1 shows the total number of tracts that constitute the high cluster, low cluster, and the outside cluster.

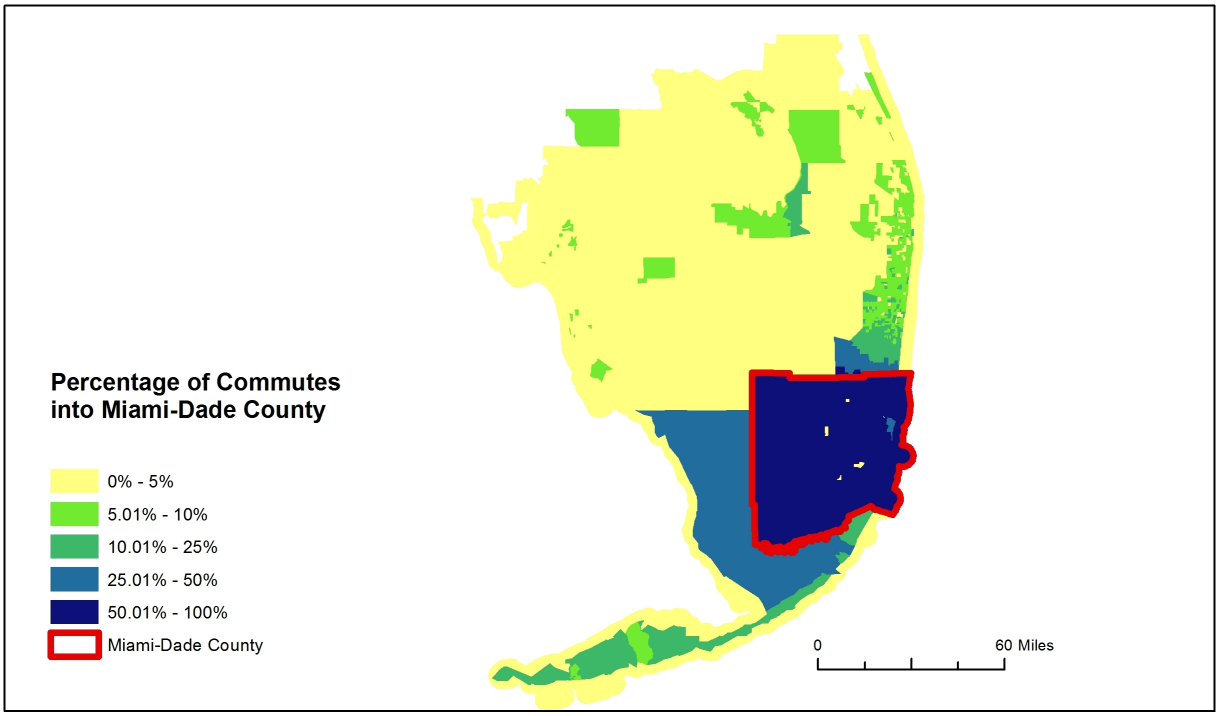


Figure 1. Commutes into Miami-Dade County in 2013.

Table 1. Number of tracts in each AMOEBA cluster.

	60 miles			90 miles			120 miles			150 miles		
	high	low	outside	high	low	outside	high	low	outside	high	low	outside
Cities												
Miami	497	630	26	497	751	27	503	993	21	506	1178	18
Fort Lauderdale	277	954	7	279	1054	6	282	1287	13	290	1295	10
West Palm Beach	190	872	18	193	1198	16	197	1536	12	199	2038	11
Counties												
Miami-Dade	516	736	12	519	1037	10	519	1176	11	519	1384	13
Broward	360	994	4	361	1276	5	361	1429	5	366	1950	8
Palm Beach	331	1127	7	330	1450	8	330	2118	8	330	2914	8
Metro Region	1206	282	5	1207	597	6	1207	1250	10	1207	2035	11
Urbanized Area	1195	168	5	1199	310	7	1202	636	7	1205	1207	12

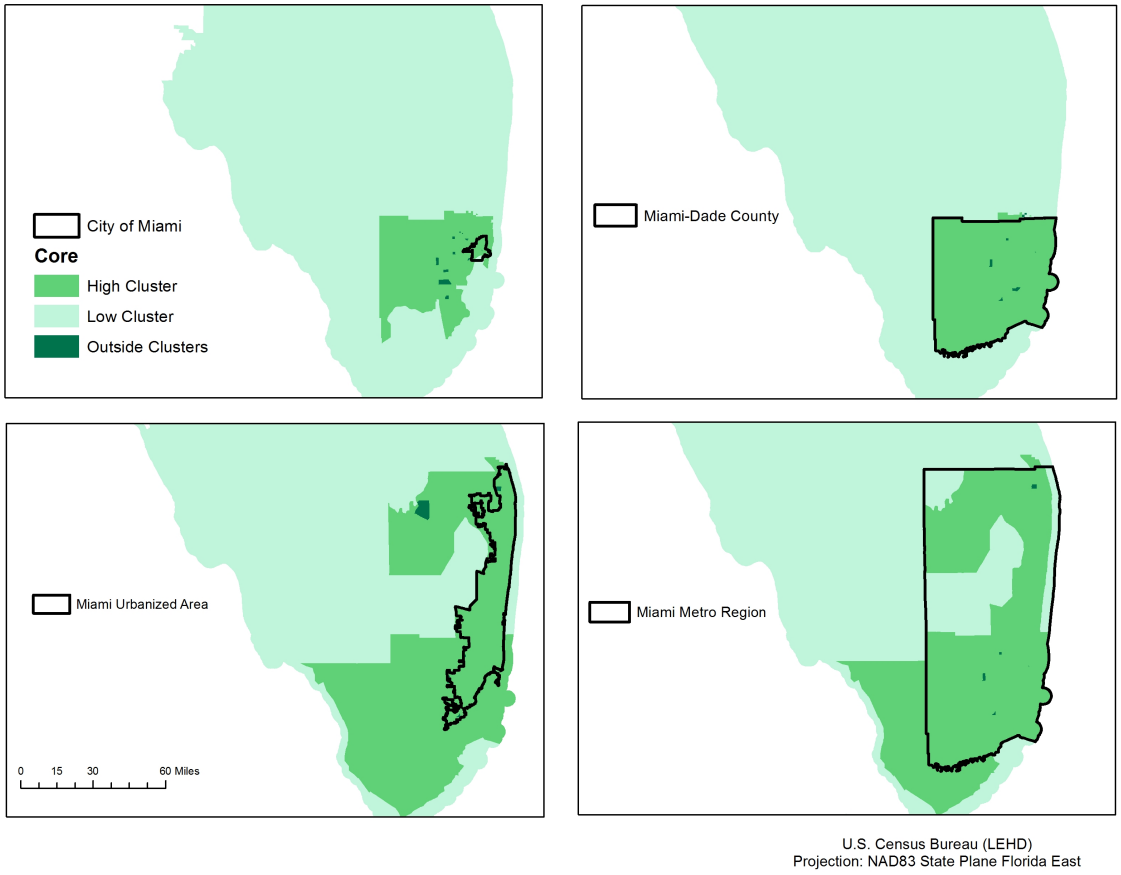


Figure 2. AMOEBA results for Miami, Miami-Dade County, the Miami urbanized area, and the metro region.

References

Aldstadt, Jared, & Getis, Arthur. (2006). Using AMOEBA to Create a Spatial Weights Matrix and Identify Spatial Clusters. *Geographical Analysis, 38*(4), 327-343. doi: 10.1111/j.1538-4632.2006.00689.x

Getis, Arthur, & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis, 24*(3), 189-206. doi: 10.1111/j.1538-4632.1992.tb00261.x

Horner, Mark W., Schleith, Daniel K., & Widener, Michael J. (2015). An Analysis of the Commuting and Jobs–Housing Patterns of Older Adult Workers. *The Professional Geographer, 67*(4), 575-585. doi: 10.1080/00330124.2015.1054018

Morrill, R., Cromartie, J., & Hart, G. (1999). Metropolitan, urban, and rural commuting areas: Toward a better depiction of the United States settlement system. *Urban Geography, 20*(8), 727-748.

Privacy Considerations for Duplicate Points in Masked Geodata

D. E. Seidl[1]

[1]Department of Geography, San Diego State University – UC Santa Barbara, 5500 Campanile Drive, San Diego, CA 92182
Email: dseidl@mail.sdsu.edu

Abstract

Reversal of a geomasking procedure can stem from the decryption of just a few points in a data set. This study explores the risks to privacy from the existence of duplicate data coordinates, and how such points are treated differently according to masking technique. Analysis of duplicates is conducted on a sample of urban foreclosure data, though the presence of duplicates should be considered on a case-by-case basis before releasing a masked data set. A nearest neighbour distance calculation for multi-unit parcels is recommended for weighting displacement distances in masking procedures on geodata with duplicate coordinates.

1. Introduction

Geomasking techniques, which alter point distributions to protect privacy, are seeing increased use in public-facing applications. The citizen science site, iNaturalist, for instance, allows users to upload species observations with coordinates randomized within a 0.2 by 0.2 degree area (http://www.inaturalist.org/pages/help) (retrieved May 10, 2016). With wider use of masked geographic data, there is greater potential for users to decrypt masking techniques and ascertain original locations. The presence of duplicate sets of points at the same coordinates can pose differential risks to privacy when those points are masked.

Early on in geomasking studies, researchers warned that releasing multiple versions of masked data could result in the reversal of the masking procedure (Armstrong, Rushton, and Zimmerman 1999). Zimmerman and Pavlik (2008) later demonstrated that the release of multiple masked data sets increases reverse engineering probability, since randomized points converge around original locations. Others have noted that if an adversary is able to determine the distance threshold used in masking point data, it might be possible to re-identify original locations (Zhang *et al.* 2015). A less-explored possibility is that multiple releases of points within a single data set—i.e. duplicate points—could reveal additional information about housing type, which could subsequently be used to uncover household identities. This is particularly true if that housing type is rare, or an anomaly in the study area.

This study explores the privacy risks associated with the presence of duplicate points in a geographic data set and summarizes how the treatment of duplicate points varies by masking technique. For example, data representing different households may be matched to the same latitude and longitude if the data subjects live in the same apartment building or residential parcel. Foreclosure data are an example of how sensitive data about disparate households may be tied to the same coordinates. Multiple foreclosures in the same building will generally geocode to the same location. The treatment of duplicate points by a masking technique can adversely impact cluster detection, or increase the risk of household re-identification. Figure 1 demonstrates how maintaining a set of duplicates together when masked impacts the privacy risk when a map user is informed of auxiliary housing geodata. Duplicate points in the masked data would suggest a common origin in the same housing parcel, and if there is only a single multi-unit residential parcel nearby, an adversary may make an educated guess

at the original parcel. The adversary also gains new information regarding approximate distances used in the masking procedure.

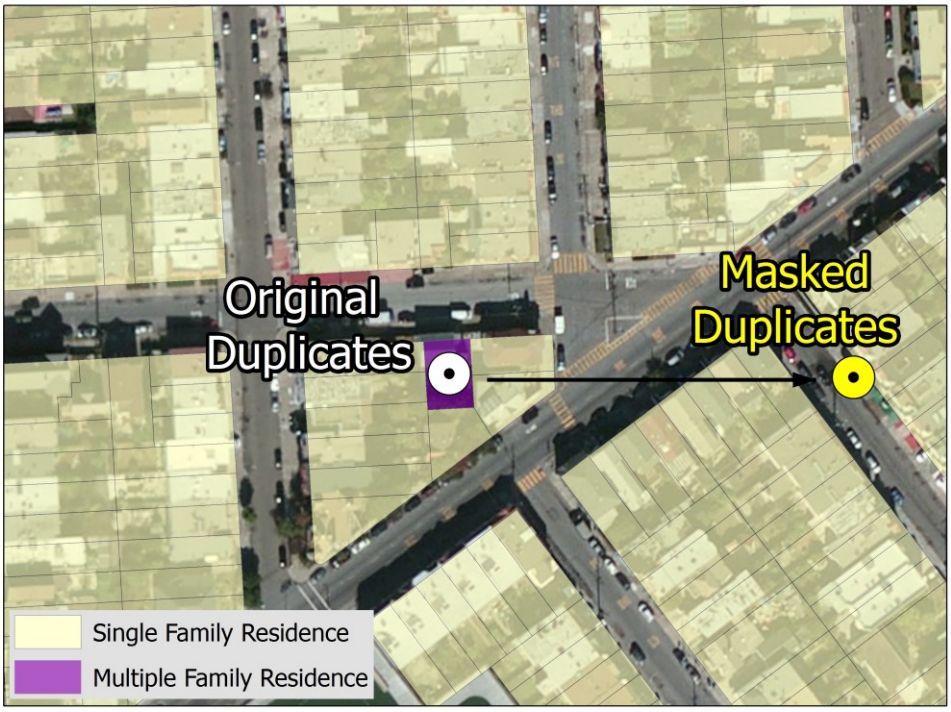


Figure 1. Example of duplicate masked points from unique multi-family residence

1.1 Duplicate Point Treatment by Masking Technique

The displacement of duplicate points varies according to masking technique. Either each point is treated separately and re-located in a somewhat randomized manner, or each set of duplicate points is displaced together and assigned new identical coordinates in masked form. The treatment of duplicate points by known masking techniques is summarized in Table 1. Random perturbation, which displaces points a random distance and direction within a distance threshold, results in differential displacement of duplicates, as do its variants of weighted, donut, and Gaussian perturbation, which also rely on randomization (Zandbergen 2014). Location swapping, in which displacement options are limited to the same land use type as the original point, also re-locates each duplicate point separately (Zhang *et al.* 2015). Affine transformations, with translate, rotate, or change the scale of point distributions (Armstrong, Rushton, and Zimmerman 1999), displace duplicates together. Grid masking (Leitner and Curtis 2006), masking based on the Military Grid Reference System (MGRS) (Clarke 2016), and Voronoi masking (Seidl *et al.* 2015) also re-locate duplicate points to new identical coordinates. Grid masking, which snaps points to regularly-spaced grid cells, may pose less of a risk of identifying multi-unit parcels, since single-family residence points also tend to be aggregated together under this technique.

Table 1. Duplicate point displacement by technique.

Masking Technique	Duplicate Displacement Type
Random Perturbation	Separate
Weighted Random Perturbation	Separate
Donut Masking	Separate
Gaussian Perturbation	Separate
Location Swapping	Separate

Affine Transformations	Together
Grid Masking	Together
MGRS Masking	Together
Voronoi Masking	Together

2. Methods

2.1 Data

Property foreclosures are an example of geodata which flag sensitive personal financial status, but also remain generally available to the public. Multiple foreclosures may occur in the same residential parcel, creating instances of duplicate points. Real estate websites, such as Zillow and RealtyTrac, post foreclosure and pre-foreclosure geodata in online map applications. Pre-foreclosure properties are not for sale, but instead indicate where the owner has defaulted on loans. This publicly available information constitutes flagging of financial status at a high granularity. In February 2016, there were 326 foreclosure listings in the city boundary of San Francisco. Of the listed properties, 6 locations had at least 1 duplicate address, and there were 22 overall duplicates. While this study focuses on foreclosure data, the considerations are relevant to any datasets that might include duplicate locations. Examples are survey data with participants from multi-unit households, crime data, and social media data.

2.2 Household Re-identification

The probability of household re-identification is dependent on contextual geographic data accessible alongside the masked coordinates. In San Francisco, geographic boundaries of land use parcels and the quantity of residential units in each are freely available from city data repositories. This means that if duplicate points are kept together when masked, and if multi-unit parcels are rare in the study area, the map user has a strong likelihood of uncovering the original location of the duplicate points. Privacy in geomasked data sets is often measured as the number of potential residences closer to the masked location than the original location, a concept referred to as spatial *k*-anonymity (Zandbergen 2014). For duplicate points kept together by affine transformations and grid, MGRS, and Voronoi masking, a more accurate risk of original parcel identification is presented by the number of multi-residence parcels neighbouring the masked location. Weighted distance thresholds in masking procedures work proactively from the other side by addressing distance to neighbours before masking is performed. One way of accomplishing this is by calculating the distance to *k* nearest neighbours (Seidl *et al.* 2015).

There are 40,773 parcels in San Francisco with more than 1 residential unit, or approximately 30% of all residential parcels in the study area. For the 6 locations where foreclosures are duplicated, the distance to *k* nearest neighbours of multi-unit housing from 1 to 10 is calculated. This process provides a means of assessing the likelihood of having adequate multi-unit neighbouring parcels to maintain anonymity if duplicates remain at matching coordinates in the masked data. San Francisco is a densely populated urban area, so in more rural regions, a multi-family household may be a stronger unique identifier, resulting in a greater distance to the nearest multi-family household.

3. Results

For each of the 6 lettered duplicate foreclosure points, the distance to *k* nearest neighbours is shown in Table 2. Multi-unit parcels are relatively close to one another in these urban data; on average it was 12.5 meters to the nearest other parcel with multiple residential units,

though this varied between points. Duplicate points are likely to be more detrimental to confidentiality in rural areas where multi-unit parcels are rare. This method is adequate if each duplicated household only has one other corresponding coordinate, but may be extended to consider the *k* nearest neighbours with at least as many units as the number of duplicated points at a location. Otherwise, a location with 15 duplicated points may be identified as unique among smaller multi-unit parcels.

Table 2. Distances to *k* nearest neighbours for multi-unit parcels (meters).

	Point						
k	A	B	C	D	E	F	Average
1	32.0	2.6	11.7	5.5	4.0	18.8	**12.5**
2	32.1	5.4	34.1	32.9	7.2	28.1	**23.3**
3	46.7	5.8	42.0	34.4	8.2	30.0	**27.9**
4	46.7	8.2	48.0	36.3	12.1	30.6	**30.3**
5	48.8	11.1	49.0	42.0	12.6	33.0	**32.7**
6	49.6	11.7	53.0	43.3	13.0	40.8	**35.2**
7	52.0	12.4	63.3	44.0	13.3	43.1	**38.0**
8	52.7	13.3	63.4	45.8	13.9	43.3	**38.7**
9	54.2	14.0	64.6	50.1	14.4	44.0	**40.2**
10	54.7	14.8	67.2	54.5	15.1	44.3	**41.8**

4. Conclusion

This study introduces the potential privacy risks of keeping duplicate points in masked data sets. Affine transformations, MGRS masking, and Voronoi masking, which keep duplicate points together when masked, increase the probability of identifying original corresponding parcels. Grid masking also keeps duplicate points together, but is more likely to group them with other point locations, as long as the distance threshold is sufficiently large. Calculating the distance to neighbouring multi-unit parcels can help to weight masking to protect anonymity with duplicate points.

References

Armstrong MP, Rushton P and Zimmerman DL, 1999, Geographically masking health data to preserve confidentiality. *Statistics in Medicine* 18(5): 497–525.

Clarke, KC, 2016, A Multiscale masking method for point geographic data. *International Journal of Geographical Information Science* 30(3): 300-315.

Leitner M and Curtis A, 2006, A First step towards a framework for presenting the location of confidential point data on maps—results of an empirical perceptual study. *International Journal of Geographical Information Science* 20(7): 813–822.

Seidl D, Paulus G, Jankowski P and Regenfelder M, 2015, Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography* 63: 253-263.

Zandbergen, PA. 2014. Ensuring confidentiality of geocoded health data: assessing geographic masking strategies for individual-level data. *Advances in Medicine* 1–14.

Zhang S, Freundschuh SM, Lenzer K and Zandbergen PA, 2015, The Location swapping method for geomasking. *Cartography and Geographic Information Science* (Online): 1–13.

Zimmerman, DL and Pavlik C, 2008, Quantifying the effects of mask metadata disclosure and multiple releases on the confidentiality of geographically masked health data. *Geographical Analysis* 40(1): 52–76.

Toward A Location Based Service for Assessing and Recommending Landscape Views

S. Soleimani[1], M. R. Malek[1], G. Sinha[2]

[1] Geodesy & Geomatics Department, K.N.Toosi University of Technology, Tehran, Iran
Email: {ssoleimani; mrmalek}@kntu.ac.ir

[2]Department of Geography, Ohio University, Athens, Ohio, 45701, USA
Email: sinhag@ohio.edu

Abstract

The rationale and initial design considerations are presented for a location based service (LBS) that will offer people personalized recommendations for aesthetically desirable places. This will be based on people's personal characteristics and environmental factors known to affect landscape perception. A prototype user-interface has been designed for eliciting users' perceptions of landscape elements and colors they perceive from photographs of landscape views. In the long-term, a VGI repository database will be maintained to make personalized location specific recommendations about suitable landscape views for user-specified locations. A prototype fuzzy logic recommendation system comprising 72 rules has also been designed to model how personal factors (gender, age, health status) and perceived colors likely determine the aesthetic suitability of landscape views. An initial trial revealed some limitations and stressed the need for further testing of the feasibility of fuzzy logic as the reasoning framework for this LBS for recommending suitable landscape views.

1. Introduction

As urbanity imposes psychological demands that become excessive, people seek stress-reducing (especially natural) environments where they can recreate and relax. The investigation of the ecological and visual aspects of physical landscapes is critical for understanding and addressing many of the psychosomatic disorders of urban dwellers. In this paper, we focus on visual perception of people's everyday environments. Human-beings engage in a wide variety of important decisions in their everyday lives, based on visual properties of the geographic environment and their geographical knowledge. Examples include coping with environmental visual pollution, finding one's way in spatially extended areas, and finding a landscape view appropriate to one's emotional status. Color, shape, composition, and configuration of objects are important components of visual stimuli that should be studied to understand our perceptions of and reactions to physical landscapes in urban environments.

In this short paper, we present the rationale and initial design considerations for a prototype Location Based Service (LBS) for landscape view assessment. This LBS will ultimately offer people personalized recommendations for places (near their chosen location(s)) that are most likely to appeal to them aesthetically. The LBS is designed to make such personalized recommendations based on available knowledge of how people's personal characteristics affects landscape perception. A prototype user-interface has been designed for eliciting users' perceptions about the types of landscape elements and the colors they perceive in images showing landscape views. An important goal of this project is to maintain a large database of people's perception of landscape views, based on volunteered geographic information (VGI) generated from people tagging images uploaded by them and/or others.

Another important component of this LBS is a fuzzy logic decision system which implements rules reflecting knowledge from research on how personal characteristics (gender, age, health status) and environmental variables together affect the aesthetic and emotional appeal of landscape views. Currently, the dominant color of landscape views is the only environmental variable being used in the fuzzy logic rule-base for this exploratory stage of the project. Color perception and how it varies with age, gender and health condition is heavily researched in multiple fields, which gave us confidence in developing our first set of rules based on color perception alone. Over time, the rule base will be refined and greatly extended to reason with multiple environmental factors that affect people's perception and emotional response to landscape views. Despite the additional effort needed, the fuzzy logic system offers scope for more nuanced reasoning, compared to conventional Boolean logic reasoning.

2. Overview of LBS System Design

The theoretical concepts and system design decisions for the LBS prototype (currently implemented in a Visual Basic programming environment) is summarized in this section. This is a refinement and development from earlier work completed for the first author's Master's thesis project (Soleimani 2015).

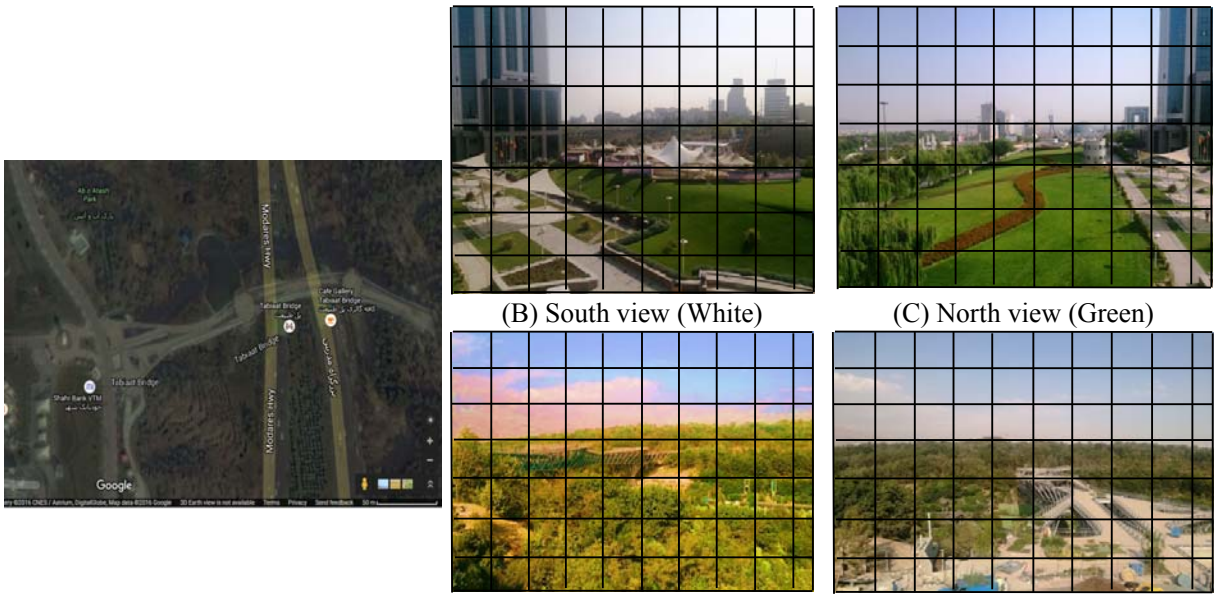
The user-interface for eliciting VGI allows users to both upload and interactively create descriptive tags images of landscape views. Multiple images for different directional views of a location are encouraged to be input (see Figure 1). The image tagging process is currently set up to avoid free-form descriptions; instead, users choose from a pre-defined list of landscape elements and colors to describe the salient elements and their characteristic colors (Le Yaouanc *et al.* 2010; O'Connor and Kroefges 2008). They also estimate distances to landscape elements from the photograph's vantage point (but this information is not being used by the system currently). A grid of cells (see Figure 1) is also displayed to help users count the number of cells each specified landscape element is covered by in the image—this allows the system to calculate the proportion of all identified colors in the image and assign a dominant color for every image. Additionally, based on research on the impact of colors on people's emotions, the LBS system categorizes colors as follows: *very warm* (yellow and orange), *warm* (brown and red), *cold* (green, blue, and magenta) and *neutral* (grey, black, and white) (Chappel et al. 2014).

The fuzzy logic decision system was implemented in MATLAB software and comprises 72 independent if-then combinatory rules based on all possible combinations of 4 input linguistic variables (gender, age, health condition, and dominant color category)—the desired consequent for each rule maybe one of four suitability values (good, fair, acceptable, bad) (see Table 1 for examples). The specific methods for deriving fuzzy sets and membership functions for the linguistic variables are based on our interpretation of research findings from color perception studies (Kaya and Epps 2004; Chappel et al. 2014; Sokolova and Fernández-Caballero 2015).

3. Experimental Validation

Experimental testing and VGI collection is yet to begin because system design issues are still being resolved. The authors have been using only themselves as test subjects, but an informal validation trial was conducted with a 35-year female cancer patient known to the first author. She chose to upload four images of the Fire-and-Water-Park of Tehran (Figure 1). It took her the longest to process her first image, but it got progressively faster. Her color tags for landscape elements were surprising and revealed that her judgment was based not just on

colors apparent in the uploaded images, but also (and probably more so) by her memory of prior *in situ* observations in the park. Based on her personal profile (female, middle-aged, cancer afflicted), Table 1 shows the four possible rules that are available for this user in the current system. Based on the dominant colors for views in Figure 1, cold color dominant views are found towards the north and east. Hence, given the rules, the system should yield relatively higher suitability membership values (good, fair) for the north and east view photos and lower membership values for the others.



(A) Location of Fire-and-Water Park, in Tehran, Iran

(B) South view (White)

(C) North view (Green)

(D) East view (Yellow)

(E) West view (Gray)

Figure 1. Fire-and-Water-Park of Tehran, Iran as visible in satellite imagery view (A) and four different directional views (B – E) and dominant color tagged by a user.

Table 1. Four possible rules for a female, middle-aged user suffering from cancer

if	Gender	Age	Disease	Dominating Color of View	then	(Desired) View Suitability
R_1	Female/ Male	Young/Middle-aged/ Elderly	Cancer/Neurological/ Vision	Cold/ Neutral/ Warm/Very warm		**Good**
R_2	Female/ Male	Young/Middle-aged/ Elderly	Cancer/Neurological/ Vision	Cold/ Neutral/ Warm/Very warm		**Fair**
R_3	Female/ Male	Young/Middle-aged/ Elderly	Cancer/Neurological/ Vision	Cold/ Neutral/ Warm/Very warm		**Acceptable**
R_4	Female/ Male	Young/Middle-aged/ Elderly	Cancer/Neurological/ Vision	Cold/ Neutral/ Warm/Very warm		**Bad**

The final step of defuzzification in any fuzzy logic system is necessary to convert a fuzzy set output to a crisp number could then allow choice of the appropriate linguistic label to describe the obtained output. For this experiment, the recommended suitability scores from

five available defuzzification methods in MATLAB confirmed that the suitability assessment for a view may depend strongly on the defuzzification method. As can be inferred from the suitability scores tabulated in Table 2, the bisector, centroid and MOM methods yield such similar scores for all views images that it is not possible to distinguish between them; the LOM method would recommend the neutral colored south and west views (suitability scores of 0.66 vs. 0.41 for north and east views); only the SOM defuzzification method would lead to the desired recommendation of the north and east cold views (suitability scores .41 compared to 0.16 for the south and west views) as being more suitable than the other two views. This clearly suggests that the fuzzy logic fuzzification and defuzzification methods need to be explored more thoroughly for this prototype LBS system.

Table 2. Defuzzification results

Method	North View	South View	West View	East View
Bisector	0.43	0.41	0.41	0.41
Centroid	0.45	0.41	0.41	0.41
Smallest of maximum (SOM)	0.41	0.16	0.16	0.41
Largest of maximum (LOM)	0.41	0.66	0.66	0.41
Mean of maximum (MOM)	0.41	0.41	0.41	0.41

4. Conclusion

The LBS prototype system design and even the underlying rationale will continue to evolve and be refined, since these rules are based on a much-simplified model and may be modified and refined based on more in-depth review of literature, recommendations from experts, and future human-subject validation trials. In the next iteration, the fuzzy rule base should be updated to decide the color category based not on the dominant color in a landscape view but the proportions of all colors tagged by users. Other factors (e.g., configuration, texture, elements) that affect landscape perception must eventually be incorporated into the fuzzy logic system. Experimental validation suggested the need for an improved user-interface, and reducing users' cognitive load while processing images; the prototype must also be migrated to a web-based development environment for collecting sufficient VGI. Finally, the VGI elicitation process can yield not just a database of landscape view descriptions, but can also be designed to capture users' assessments of the suitability of those views. This component of the VGI database will be most critical for validating and improve the rule-base that is supposed to capture how people vary in how they perceive different views of landscapes.

References

Chappell M, Davis B, Pennisi B and Sullivan M, 2014, Landscape Basics: Color Theory. *UGA Extension Bulletin 1396*.

Kaya N and Epps HH, 2004, Relationship between Color and Emotion: A Study of College Students. College Student Journal, 38(3): 396–405.

Le Yaouanc JM, Saux É and Claramunt C, 2010, A Semantic and Language-Based Representation of an Environmental Scene. *Geoinformatica,*14(3): 333–352.

O'Connor L and Kroefges PC, 2008, The Land Remembers: Landscape Terms and Place Names in Lowland Chontal of Oaxaca, Mexico. *Language Sciences*, 30(2): 291–315.

Sokolova MV and Fernández-Caballero A, 2015, A Review on the Role of Color and Light in Affective Computing. Applied Sciences, 5: 275-293.

Soleimani S, 2015, *Landscape Description in Volunteered Geographic Information (VGI) using Spatial and Temporal Relationships*. Unpublished M.Sc. Thesis, KNT University, Tehran, Iran (In Persian).

Can we use OpenStreetMap POIs for the Evaluation of Urban Accessibility?

S. Steiniger[1], M.E. Poorazizi[2], D.R. Scott[1], C. Fuentes[1], R. Crespo[3]

[1]CEDEUS & Pontificia Universidad Católica de Chile A, Avda. Vicuña Mackenna 4860, Macul - Santiago - Chile
Email: ssteiniger@uc.cl

[2]University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada
Email: mepooraz@ucalgary.ca

[3]Universidad Bernardo O'Higgins, Avenida Viel 1497 Ruta 5 Sur - Santiago, Chile
Email: ricardo.crespo@ubo.cl

Abstract

High urban accessibility, measured in terms of easy and fast to reach public and private services, such as schools, parks and shops, is considered an important indicator for quality of life. We developed a web-based platform to measure urban accessibility for transportation planning. Several cities that we aim to study are in Chile where data on urban services is scarce. We evaluated if OpenStreetMap (OSM) can be used as a source of Points of Interest (POI) to evaluate urban accessibility. Based on a field survey in Santiago de Chile, we found that completeness of the OSM POI database is geographical very diverse (7% - 73%) and therefore accessibility scores change significantly. However, scores in some areas did not change much when POIs were added to the OSM database. Given the lack of alternative data sources we recommend using OSM, but suggest evaluating POI completeness in areas with average and low accessibility scores.

1. Urban Accessibility

Urban accessibility measures how easy or difficult it is to visit places of one's day-to-day activities by measuring travel time. As daily activities are considered things such as going to work or school, visiting the library or a park, doing (grocery) shopping, going out to a restaurant, or visiting a movie theatre. Traditionally, transportation planning has focused on measuring accessibility as accessibility to work places, as work-related trips generate high travel demand. Following newer but also classic accessibility measurement theory and practice (e.g., Talen and Anselin 1998; Geurs and van Wee 2004), we developed a web-based platform that measures urban accessibility by counting activity opportunities within a certain travel time – or "travel-shed". The platform should later be used to evaluate how accessibility changes when investments into urban infrastructure or changes to public transit schedules are made.

To evaluate urban accessibility based on activity opportunities and with respect to different modes of travel (e.g. walking, biking, public transit) different data sources are needed: (a) activity opportunities in form of Point of Interest (POI) data, (b) road and pedestrian path network information, and (c) bus and subway/LRT schedules if accessibility with public transit is of interest. In this paper we aim to evaluate how completeness of the (OpenStreetMap) POI database affects calculated walkability scores.

2. Why OpenStreetMap as a Data Source

Our first platform prototype was implemented to measure accessibility in Calgary, Canada (Steiniger *et al*. 2013). For this prototype we used road network data from OpenStreetMap and Points of Interest (POI) were received from MapQuest's proprietary online database.

When we changed our case study to the metropolitan area of Santiago de Chile, an urban area formed by about 37 municipalities with more than 5 Million inhabitants, it became clear that the MapQuest POI database has a focus on North America and Europe and contains much less POI data for regions such as South America. Observing that Google's database also missed a significant number of POI for some areas of Santiago de Chile we chose OpenStreetMap (OSM) as our POI data source. OSMs open database licence and global database coverage helped in making this choice too. Subsequently we used OSM's Overpass API (Olbricht 2015) in our Santiago implementation to obtain POI data.

3. Evaluating Accessibility and OSM Completeness

Given the relatively few POI objects found in the MapQuest POI database, and noting that the Google's database seems also to miss quite some information for some areas of Santiago de Chile, it was seen necessary to evaluate the completeness of OSM's POI database and evaluate the effect of completeness on accessibility score calculation. We note here that a full coverage of all POI is not likely to be achieved for any database, as some areas of Santiago are considered as too dangerous to enter for non-residents. Hence, as a further question, it is of interest what level of POI completeness needs to be reached to obtain a reliable estimate of urban accessibility. An initial hypothesis was that data completeness may be similarly low across Santiago, and therefore accessibility scores will raise by the same amount if the database is completed. For our experiments we concentrate on walkability, as a subcomponent of general accessibility. In the next two subsections we explain briefly how walkability scores are calculated and then present our approach to completeness evaluation.

3.1 Evaluating Urban Walkability

Adopting the basic method from Walkscore.com (2011) our accessibility score is calculated in three basic steps for a particular location (x,y) (see Steiniger *et al*. 2013): (i) calculation of an accessibility area, i.e. a walkshed, (ii) retrieval of activity opportunities (i.e. POI) within that area, and (iii) calculation of a score. Thereby the contribution of each POI to the total score, i.e. its weight, is based on the POI type, number of POI of the same type, and perhaps distance to the location (x,y). The method returns a "walkscore" between 0 (not walkable) and 100 (very walkable).

The POI types considered are: (1) grocery stores, (2) restaurants, (3) shopping (shopping & business), (4) cafés and bars/pubs, (5) banks (ATMs), (6) parks, (7) schools, (8) books (libraries & book stores), and (9) entertainment (cinemas, sport venues, museums). Important is that a POI may not contribute to a total score if the area/shed contains (too) many other POI of the same type. For instance, to reach the maximum score of 100 there needs to be only one ATM and one Park, but five shops/businesses. We note that the original WalkScore.com implementation ("Street Smart") used a 1-mile circle for POI evaluation while our approach utilizes the road network to obtain a walkshed. Accessibility evaluation is implemented employing OGC Web Processing Services (WPS) with GeoServer, and by using OpenTripPlanner, PostGIS and Python scripts for network, area and score analysis (see Poorazizi *et al*. 2015 for details).

3.2 Evaluation of the OSM POI Database

The evaluation of accuracy and completeness of the OpenStreetMap database has been lately a frequent topic in the GIS literature, whereby different feature types and geographical regions were analysed (e.g. Girres and Touya 2010; Arsanjani *et al*. 2015; Camboim et al. 2015; among others). However, for our particular purpose and geographical focus we performed our own completeness evaluation. Therefore we selected 11 ground survey sites in different parts of the Santiago metropolitan area. Stratified sampling was applied with the

objective to cover areas with (i) low and high accessibility scores calculated in an initial step, and (ii) areas that are equally geographically distributed.
A team of two persons surveyed each area for objects that would fall into the 9 POI categories mentioned above, while also looking for missing streets and pedestrian paths. Some of the survey sites were considered too dangerous to go there in person. Here, we used Google's StreetView tool – which is available for almost all areas of Santiago - to map POI on screen. In total about 570 unmapped POI were found in the survey, and added to the OSM database. Afterwards we calculated walkability scores (assuming a 15 min walk at 5km/h) for a point grid/lattice covering Santiago, with a spacing of dx=320m and dy=370m. Thereby we calculated scores with the original and with the updated OSM POI database, and evaluated scores only for those grid points that fell within one of the 11 survey areas.

4. Results and Recommendations

The resulting walkability scores for the survey areas, and for Santiago, aren't distributed normal/Gaussian (we observed a mixed distribution), so it is actually not recommendable to present average values. However, to obtain an idea about score variability we present average values below. Figure 1 (left) shows a walkability score map for Santiago, calculated with the original OSM POI database.

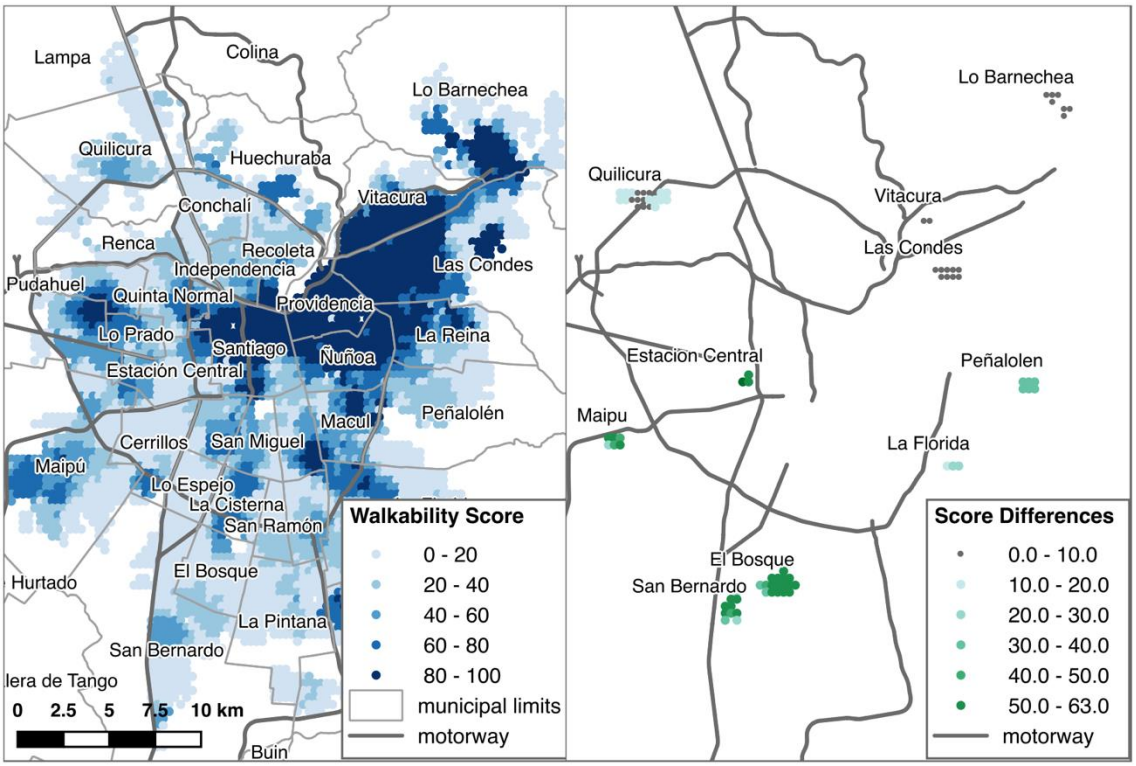


Figure 1. Walkability scores for Santiago de Chile (left) and score differences for 10 of 11 surveyed areas after OSM database updates (right).

POI completeness was in-between 7% to 73%. About half of the survey areas had a POI completeness of 20% or less. Higher percentages of completeness have been reported for other OSM feature types, such as roads (Girres and Touya 2010: 37-45% in France) and land-use (Arsanjani *et al*. 2015: 40-60 % for some urban areas in Germany). This difference can be explained by the fact that roads, paths and land cover are usually visible in satellite images, which are often used by volunteers for mapping on the desktop PC. In contrast, when POI should be mapped on the desktop PC, then the area needs to be covered by a street-image

database like Google StreetView. However, a basic set of POI may exist already due to country-wide imports of governmental datasets, containing for instance schools or hospitals. We note that in our survey we found only 2 unmapped pedestrian pathways.

Looking at the walkability scores, the score average did not change at all for two of our 11 survey areas, despite additions to the POI database. The maximum (average) score change found was 58 score points – which is significant considering the score limits [0-100]. Summarizing over all 82 sample grid points the score average rose by about 50 walkscore points (see also Figure 1 right). The two areas that did not experience any change were (a) a Central Business District (Las Condes) – in which 23 POI where added to an existing 62 POI, and (b) a suburban urban sprawl neighbourhood (Lo Barnechea) – in which 6 POI where added to 13 existing POI.

Interestingly there is no clear linear relation between the number of POI added to the database versus score increase. However, it can still be said that: If a neighbourhood was previously scoring low, then scores will raise if more POI are added. But, if a score is already high, then adding new POI to the database may not raise the score at all, due to maximal count limits for POI types during score calculation. Our conclusions are therefore as follows: (i) OpenStreetMap can be used for POI-based urban accessibility analysis, whereby database completeness has a significant influence. (ii) There is most likely no general accessibility score offset/bias for a larger city, since city districts have a high variability in POI completeness. (iii) If OpenStreetMap is used, then POI completeness needs to be evaluated for areas with low or average scores (< 70) before further conclusions can be drawn. Given the (currently few) data points, a POI database completeness of 85% or higher is needed for reliable walkability score estimation. In a next step we will investigate if there are correlations between completeness and socio-demographic factors. If correlations exist, then this would allow developing models for the estimation of completeness.

Acknowledgements

The authors acknowledge financial support by the project “AccesoBarrio” (Conicyt/Fondecyt 1150239) and the Centro de Desarrollo Urbano Sustentable (Conicyt/Fondap/15110020).

References

Arsanjani JJ, Mooney P, Zipf A. and Schauss A, 2015, Quality Assessment of the Contributed Land Use Information from OpenStreetMap Versus Authoritative Datasets. In: Arsanjani JJ et al. (Eds.), *OpenStreetMap in GIScience*, LNGC. Springer, 37–58.

Camboim SP, Bravo JVM and Sluter CR, 2015, An Investigation into the Completeness of, and the Updates to, OpenStreetMap Data in a Heterogeneous Area in Brazil. *ISPRS International Journal of Geo-Information 4*, 1366–1388.

Geurs KT and van Wee B, 2004, Accessibility evaluation of land-use and transport strategies: review and research directions. *Journal of Transport Geography*, 12:127–140.

Girres JF and Touya G, 2010, Quality Assessment of the French OpenStreetMap Dataset. *Transactions in GIS* 14, 435–459.

Olbricht RM, 2015, Data Retrieval for Small Spatial Regions in OpenStreetMap. In: Arsanjani JJ et al. (Eds.), *OpenStreetMap in GIScience*, LNGC. Springer, 101–122.

Poorazizi ME, Steiniger S and Hunter AJS, 2015, A service-oriented architecture to enable participatory planning: an e-planning platform. *International Journal of Geographical Information Science*, 29:1081–1110.

Steiniger S, Poorazizi ME and Hunter AJS, 2013, WalkYourPlace - evaluating neighbourhood accessibility at street level. In: Ellul C et al. (Eds.), *Proceedings of the 29th Urban Data Management Symposium*. ISPRS International Archives of Photogrammetry, Remote Sensing, and Spatial Information Science, London, UK.

Talen E and Anselin L, 1998, Assessing spatial equity: An evaluation of measures of accessibility to public playgrounds. *Environment and Planning A*, 30:595–614.

WalkScore, 2011, *Walk Score Methodology*. http://pubs.cedeus.cl/omeka/document/82

Community and Copyright: The Building of an OpenGIS Data Repository in Buffalo, NY

Monica Stephens[1]

[1]Department of Geography, University at Buffalo (SUNY), 105 Wilkeson Quadrangle, Buffalo, NY 14261, United States
Email: mgstephe@buffalo.edu

Abstract

This project examines the challenges and opportunities that arose in the first phase of building an open Geospatial data portal for Buffalo, NY. Community groups, academics, and non-profits articulated a desire to have an exchange of datasets and information that was accessible and stored in one place. However, the City of Buffalo, NY was unable and unwilling to share local data or provide a top-down initiative for collecting, storing, and disseminating geospatial information. Other locales like Detroit and Pittsburgh have relied on non-governmental actors to take the lead establishing data repositories with open data policies that were later adopted by the public sector. Through a Civic Engagement and Public Partnership grant with the United Way of Buffalo and Erie County, we built a prototype of an open GIS-data portal.

1. Introduction

Open data are often thought of as a foundation on which to develop new goods, services, and applications that could significantly improve quality of life and community development. Open data are freely available to everyone to use without the barriers conventionally imposed by copyrights, patents, or other forms of information control. Open data provides an opportunity for communities to improve transparency and democratic control of government, augment the efficiency and effectiveness of public services, and generate new knowledge and insight into processes of social, political, economic and environmental change.

Although open data systems have been strongly championed by city governments in places like New York City, Chicago, and San Francisco, other locales like Detroit (http://portal.datadrivendetroit.org) and Pittsburgh (http://www.wprdc.org/background) have relied on non-governmental actors to take the lead in establishing data repositories with open data policies subsequently following from the public sector. In Buffalo, New York, where local government has not adopted public policy related to open data, we partnered with community organizations to build a repository to collect and disseminate geospatial information.

1.1 Open Data in Buffalo

In Buffalo, NY local government has resisted sharing or disseminating datasets collected by the city or county. Community groups and non-profits provided motivation to share data by re-creating updated versions of community data. For example, the 'ReTree the District' project used smartphones and volunteers to crowdsource a house-by-house tree census of one neighbourhood in the city (Krolikowski 2015). The dataset they created was of higher quality and integrity than the data held by the city forester and was later adopted by the city.

Another community group, frustrated that the city budget was released only as single-page, non-machine readable PDFs (City of Buffalo 2015), scraped the budget website and posted on GitHub a version accessible for analysis, the city later submitted the budget to OpenBooks. The United Way of Buffalo and Erie County (UWBEC) recognized the need for data and technical applications available to the non-profit community, so in 2014 they

partnered with AT&T and organized a civic hackathon. This hackathon offered prize money to the applications that best addressed a number of municipal issues (housing blight, transportation, etc…) in Western New York. An initial goal in designing an open GIS repository was to disseminate this data as well as other data collected by community groups to the public.

1.2 Open Data and Community Control

Rhetoric surrounding open data implies that government and data empowered citizen can work together to usher in a new phase of information democracy. However, creating maps or geospatial analysis of local data is a process that has historically been imbued with power (Crampton 2005). Johnson (2014) criticized open data movements for marginalizing groups that are unable to contribute to the data by excluding them from resulting data sets.

Cities such as Buffalo have options to harness the power of motivated citizens through open data. Sieber and Johnson (2015) examined government open data systems and recommend using open data as a conduit for a larger "agenda of citizen inclusion and participation in decision-making." Johnson (2016) described different ways citizens can provide feedback and update governmental datasets at different levels of risk to the government. The city of Buffalo does not share geospatial data which precludes citizens from conducting independent analysis or participating in urban decision making.

2. Methods

To provide a service to the community, we decided to build a community data platform that would allow non-profits, academics, and community groups to post and access geospatial data without a human gatekeeper. Many open data systems (e.g. NYC Opendata, Detroit Open Data) rely on Socrata (https://www.socrata.com) which focuses on hosting open data, but not organizing or retrieving geospatial information. Additionally Socrata charges municipalities approximately $5,000/month for data hosting services. These limitations created the need to build a new platform that was geospatial and could build linkages within the data.

Our community data platform was implemented as a web application using C#, .Net, with MySQL for a database management system. Data is stored in a cloud storage system (cloudatcost), which also provides API access. Currently, the data can only be uploaded as .csv, .json, .xlsx (although we plan to expand this to shapefiles in the future). To add new data to the repository, any user can visit the portal and upload vector data. Figure 1 shows the current start-up screen that accepts uploaded files.

Figure 1. Data-submission page for community data portal

All aspects of the platform center around the master data repository that uploads, retrieves and presents data. The core concept for the repository is the use of a master ontology. This is a list of data attributes with a well defined name and meaning. Appropriate fields within a source data set are mapped to these master fields allowing users to establish linkages among previously disparate data sets. These linkages can subsequently be used to build interpretations and presentations of the data. Figure 2 identifies the major components and subcomponents of the system architecture.

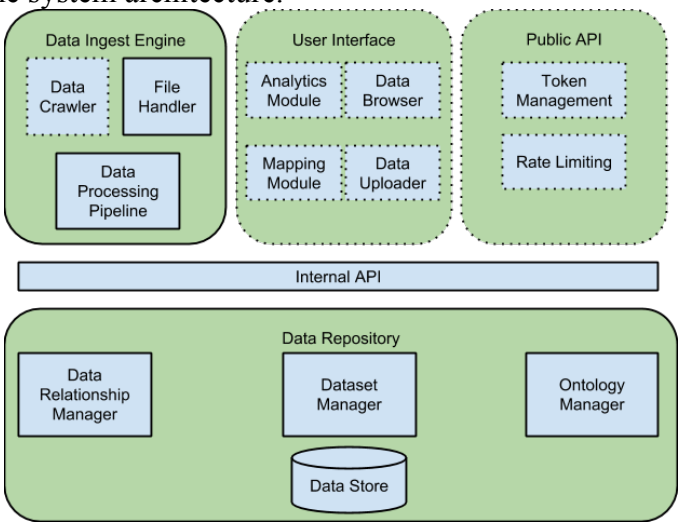


Figure 2. Major components of system architecture (dotted lines are phase 1)

In other words, if the data uploaded has an attribute similar to data that already exists in the data storage (such as a common GeoID or FIPS code from census data) then the new data is added to the existing data table. If the data is of a new type or similar data does not exist, a new data set will be created for that data in the storage. The data file is parsed based on the information provided by the uploader such as number of header rows. Phase two of this project will address the data retrieval and visualization aspects.

3. Results and Conclusion

Phase one of this project has introduced many challenges and some successes. Challenges have included training an interdisciplinary team, bridging the divide between the university and the community, and addressing cloud storage, security, copyright and server issues. Successes included building connections between the business community, non-profits, academics, and motivated community members.

Building an interdisciplinary team between Computer Science and GIS led to many disciplinary debates and conversations over ontology within systems architecture. The University at Buffalo Library was hesitant to host data without knowing the origins and was concerned about potential copyright issues if somebody uploaded copyrighted material. University at Buffalo Information Technology (UBIT) could not host this data in case somebody outside the university uploaded a malicious file. However the UB Library similarly struggles to obtain geospatial information about the local community and has referred patrons towards this geospatial repository. We hope that creating this repository will provide a forum to allow citizen scientists to compose their own analyses and see linkages within the geospatial information that describes their own community.

Acknowledgements

Dr. Aaron Krolikowski has been an essential part of this research by providing connections to the Buffalo community as well as collaborating on the direction and scope of the project. Support for the first phase of this research was generously provided by the Civic Engagement and Public Policy Research Initiative at the University at Buffalo. Shalmalee Vishwasrao served as a student research assistant and Christopher Fagiani provided technical assistance with establishing a working systems architecture.

References

City of Buffalo, 2015, Adopted Budget. ed. Ms office. Buffalo, NY.

Crampton JW, Krygier, John, 2005, An introduction to critical cartography. *ACME: an International E-journal for Critical Geographies,* 4(1)**:** 11-33.

Johnson JA, 2014, From open data to information justice. *Ethics and Information Technology,* 16(4)**:** 263-274.

Johnson PA, 2016, Models of direct editing of government spatial data: challenges and constraints to the acceptance of contributed data. *Cartography and Geographic Information Science,* 1-11.

Krolikowski A, Cotton, D, 2015, Putting the sharing economy to work. In *TEDxBuffalo*.

Sieber RE and PA Johnson, 2015, Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly,* 32(3)**:** 308-315.

A 3D Virtual Environment for Spatio-Temporal Analysis: Theoretical Approach, Proof of Concept, and User Study

R.N. Stewart[1], C. Wilkerson[1], E. Ragan[1], M. Agreda[1], D. White[1], S. Duchscherer[1], J. Piburn[1]

[1]Oak Ridge National Laboratory, P.O. Box 2008, 1 Bethel Valley Rd, Oak Ridge, TN 37831
Email: stewartrn@ornl.gov, cwilker7@vols.utk.edu, eragan@tamu.edu, mdagreda@yahoo.com, whiteda1@ornl.gov, piburnjo@ornl.gov

Abstract

We present a 3D virtual environment leveraging 3D game dynamics for statistical analysis on spatio-temporal (ST) data. We present a theoretical construct, a proof-of-concept implementation (STWorld), and preliminary results from a human-computer interaction (HCI) study. Results include a novel integration of the Unity game engine with open source tools PostGreSQL, R, and the D3.js graphics library. Preliminary results suggest that STWorld may be judged more enjoyable and incur better procedural memory retention than traditional dashboard interfaces. We conclude with findings on gender bias and the gap between gamers and non-gamers.

1. Introduction

Applying statistical tools to spatio-temporal (ST) data using traditional statistical packages requires access and understanding of statistical routines, data manipulation, navigation of user interfaces, and basic programming skills. These are obstacles for casual analysts impeding education and application. Web-based tools such as the World SpatioTemporal Analytics and Mapping Project Tool (Stewart *et al*. 2015) have helped in mitigating these. Alternatively, educational gaming research has demonstrated a strong and positive connection between gameplay, enjoyment, learning, and memory retention of new content and procedures. Jabbar and Felicia (2015) provide a review of the extensive body of work in this domain, and Armes *et al*. (1999) explored engaging statistical visualizations in virtual 3D space. In this paper, we examine the potential for leveraging first-person gameplay dynamics for analysing spatiotemporal data and raise the following questions:

1) Can we leverage benefits of 3D gaming for the study of ST data?
2) How does efficiency in learning and exploration compare to a traditional interface?
3) How does memory retention compare to a traditional web-tool?
4) Are effects influenced by gender or gaming experience?

We present a novel theoretical construct for analytics in 3D space, implement a novel proof-of-concept called STWorld, and report on a small user study aimed at these questions.

2. Theoretical Approach

We propose a 3D virtual environment populated with familiar objects whose real-world purpose and behaviour are replaced by abstract data and analytical operations. The user enters a room with a first-person perspective, where cardboard boxes represent PostGreSQL databases, a gun used to query data sets, a magnifying glass that performs statistical analysis, and walls that become interactive graph and map spaces (Figure 1). The intent is for these repurposed objects to serve as comfortable, memorable metaphors for otherwise abstract data operations. By distributing abstract analytics as objects in a room, we aim to improve memory retention of how to locate, access, and analyse data (Ragan *et al*., 2012). While not a game in the strictest sense, we repurposed foundational game knowledge for ST analytics by adopting

common 3D game dynamics including spatial navigation, selection of tools, pursuit of end goals, and typical game controls (Figure 1f). We developed a proof-of-concept to implement and test these concepts.

3. ST World: A Proof of Concept

The desktop 3D virtual environment, STWorld, was built using the Unity engine to incorporate and connect mainstream open source assets including the PostGreSQL database, the D3.js visualization library, and the R statistics package (Figure 1a). To our knowledge, the introduction of these approaches in a first-person environment is completely novel. Global ST data from the CIA World Factbook (1989-2012) was added to a PostGreSQL database and served to STWorld along with R statistical functions via a RESTful API. The D3.js library presents interactive visualizations. A standard controller (Figure 1b) allows users to move about the room, pick up tools, connect to data, and choose from available countries, attributes, years, and statistical functions. In typical game play, users enter the room and use a tool (gun or magnifying glass) to choose a database (shelved boxes, Figure 1c). The box is placed on a table and opened to provide access to CIA data organized by geographic regions shown as labeled spheres (Figure 1d). Selecting a sphere and projecting it onto the wall reveals its content in graph and map forms (Figure 1e). Users can then select areas of the map to view and compute statistics (Figure 1f).

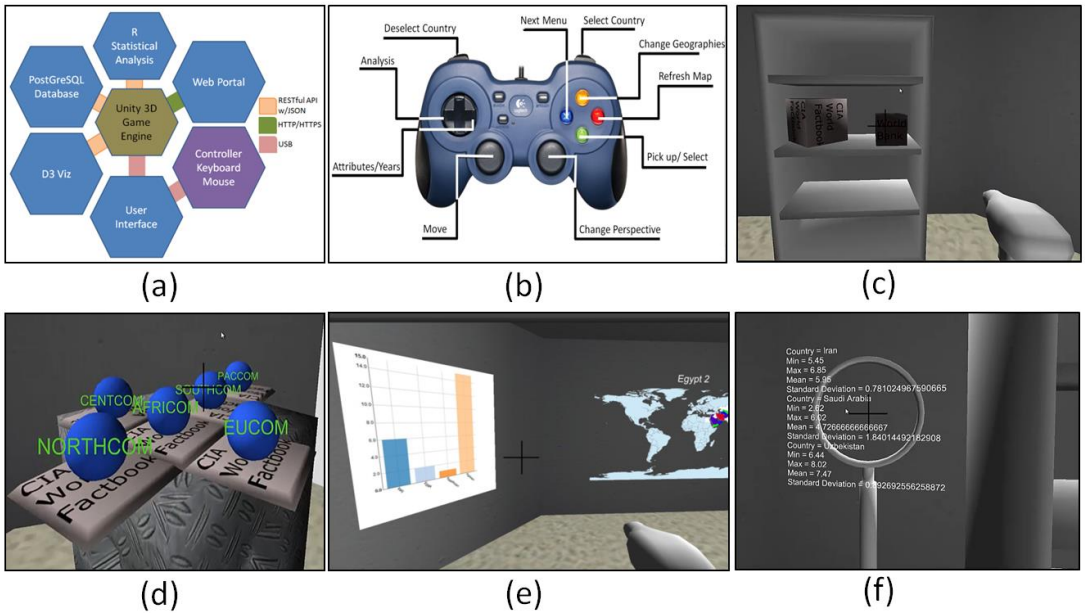


Figure 1. a) Architecture diagram, b) Game controller, c) Database representation, d) Data queries, e) Graphs and maps, f) R statistics

4. HCI Results

An HCI comparison of STWorld and WSTAMP (web tool interface) was designed to compare several factors including usability, recall, and enjoyment for users with STEM backgrounds (anticipated user community). Information was collected on participant age, gender, video game experience, education, highest math class completed, and confidence of data analysis. An open call for participants yielded 20 males and 10 females including staff and interns at Oak Ridge National Laboratory. Ages and gender skewed young (e.g. 90% under 30) and male (66%). We note the female ratio matched the national ratios for gender gaming (ESA, 2015). Participants first took a cube comparison test from ETS Kit of Factor-Referenced Cognitive Tests (Ekstrom 1976) to assess spatial ability. After a short tutorial, participants completed three tasks using both interfaces (ordering was counterbalanced). Tasks such as "*Produce a*

graph of Mexico's "birth rate" over time" were designed to produce unambiguous outcomes. Completion time was recorded for each task. One week later, participants were asked to return to repeat the exercise to assess memory retention.

STWORLD users were initially slower and had higher variance in task completion times, but were able to learn quickly becoming competitive with WSTAMP by Task 3 (Figure 2a). Gamers (60% females, 90% males) found STWORLD more enjoyable while non-gamers preferred WSTAMP (Figure 2b). A week later, 10 males and 10 females elected to return to complete the study. Performance comparisons suggest retention in STWORLD was slightly better with 6% less average time per task compared to 6% longer in WSTAMP.

Within STWORLD, non-gamers narrowed the skills gap by Task 3 (Figure 2c) suggesting a tractable learning curve. Because females were unable close the gap (Figure 2d), task completion times were regressed against gender, age, game play frequency, self-reported data skills, and ETS scores to explain variation in performance. Self-reported video game skill (1-10) was also collected but collinear (variance inflation = 2.1) with gender, so not included. See Table 1 for results.

Table 1. Model Results

Independent Variable	Task 1	Task 2	Task 3
Intercept	38.40	53.71	83.30
Gender (0-female, 1-male)	-39.95	-41.25*	-31.06*
Age	5.21	2.58	0.63
Video Gaming Per Week	0.23	-0.57	-0.63
First-Person Gaming	0.38	1.62	-1.03
Data Skill (1-10)	1.77	3.08	3.09
EKT Score	0.68	1.30	-0.87
R-squared	0.18	0.22	0.28

No variables were significant in the first task—probably because users were still learning the particular interface. With more familiarity in Tasks 2 and 3, gender bias (as seen in Jensen and Castell, 2010) emerges even within a post-secondary STEM cohort.

5. Summary

STWorld represents a first step in developing a 3D interface for spatio-temporal analysis. Gamers were able to transfer existing gaming skills (navigation, tool selection, etc.) into a spatio-temporal analytics context with greater enjoyment and better memory retention than with the web tool. Non-gamers quickly closed the gap, but gender bias was detected even in this small study. Next steps include exploring mitigation of gender bias, incorporating game incentives (e.g., scoring), and web deployment for larger studies.

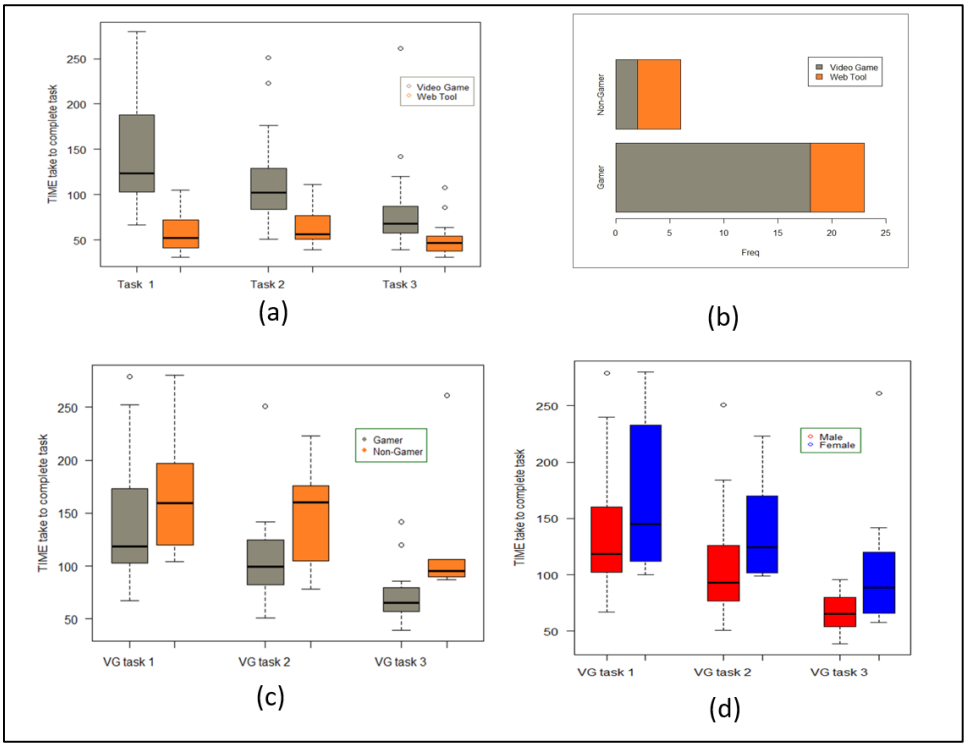


Figure 2. a) Task completion time, b) Enjoyment, c) Completion time by gaming, d) Completion time by gender

Copyright

Acknowledgement

Support for this work was provided by the National Geospatial-Intelligence Agency and the U.S. Department of Energy SULI program.

References

Arms L, Cook D, and Cruz-Neria C, 1999, The benefits of statistical visualization in an immersive environment. In: Virtual Reality, Proceedings, IEEE.

Ekstrom, R.B., 1976. Kit of factor-referenced cognitive tests. Educational Testing Service.

ESA, 2016, Essential Facts: Sales, Demographic, and Usage Data. Entertainment Software Association.

Jabbar A and Felicia P, 2015, Gameplay Engagement and Learning in Game-Based Learning: A Systematic Review, *Review of Educational Research* 85(4): 740-779.

Jenson and Castell, 2010, Gender, Simulation, and Gaming: Research Review and Redirections, *Simulation and Gaming,* 41(1): 51-71.

Ragan E, Sowndararajan A, Kopper R, and Bowman D, 2010. The Effects of Higher Levels of Immersion on Procedure Memorization Performance and Implications for Educational Virtual Environments, *Presence: Teleoperators & Virtual Environments,* 19(6): 527-543.

Stewart R, Piburn J, Sorokine A, Myers A, Moehl J, and White D, 2015. World Spatiotemporal Analytics and Mapping Project (WSTAMP): Discovering, Exploring, and Mapping Spatiotemporal Patterns Across the World's Largest Open source Datasets In: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences II-4/W2.

Extracting Accurate Building Information from Off-Nadir VHR Images

A. Suliman, Y. Zhang, R. Al-Tahir

Department of Geodesy and Geomatics Engineering, University of New Brunswick, Canada, E3B 5A3
Email: {a.suliman; yunzhang; riad.altahir}@unb.ca

1. Introduction

Since buildings are the most prominent class in urban areas, updated building information plays an important role in urban planning and management applications. The most cost-effective and broadly available geo-data for mapping building information is the very high resolution (VHR) satellite images. Thus, building detection in VHR images has been an active research area for the remote sensing community during the last two decades (Doxani, Karantzalos, and Tsakiri-Strati 2015; Khosravi, Momeni, and Rahnemoonfar 2014).

VHR optical images are the 2D perspective projection of the real surface. Hence, they lack the elevation information. Since buildings are elevated objects, image and elevation data co-registration is required for accurate and reliable detection results. However, several problems are introduced when such data types are integrated (Dowman 2004).

Elevation data are typically generated by the photogrammetric approaches and LiDAR (light detection and ranging) technology. Both of these sources provide the height information of the tops of surfaces such as buildings or trees. Hence, they result in digital surface models (DSMs). DSMs are generated as an orthographic projection (nadir view), while VHR images are usually acquired as perspective projections with across-track/along-track angle (off-nadir view). This projection difference makes accurate co-registration very difficult to achieve.

The co-registering methods of image and elevation data are extensively reviewed in our previous work (Suliman and Zhang 2015). In that work, we introduced the line-of-sight DSM (LoS-DSM) solution to overcome most of the limitations in the reviewed methods and provide pixel-by-pixel co-registration. In our recent work, some modifications have been made to the original algorithm where an improved efficient algorithm based on disparity information has been developed (Suliman and Zhang 2016).

In this study, we demonstrate the applicability of the recently developed disparity-based LoS-DSM solution for building detection in off-nadir VHR satellite images using stereo-based elevation data. The elevation data generated by the improved algorithm are of minimized terrain-relief effects. Hence, the need for calculating the aboveground building heights should be reduced in the dense urban areas that are characterized by moderate terrain-relief variations.

This paper is organized as follows: the proposed method is briefly described in Section 2, the optical data and the achieved results are provided in Section 3, and finally the conclusions are drawn in Section 4.

2. The Proposed Method

The method proposed is flowcharted in Figure 1. The method has four steps that are described as follows:

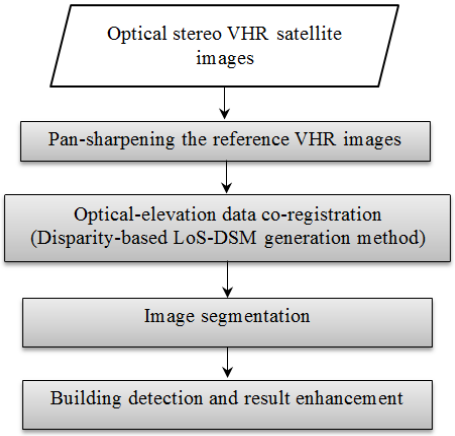


Figure 1. Flowchart for the proposed building detection method.

(1) Image Pan-sharpening
VHR images are acquired with panchromatic and multispectral bands. Since the ground resolution of the multispectral bands is one-fourth of that for the panchromatic one, these bands need to be pan-sharpened to improve the segmentation and detection results. The UNB pan-sharpening technique introduced by Zhang (2004) is selected for this step.
(2) Disparity-based LoS-DSM Generation
To achieve accurate and efficient image-elevation co-registration and minimizing the terrain-relief effects for the generated stereo-based elevation data, the disparity-based LoS-DSM generation algorithm is selected to be executed. The algorithm is described in our recent work (Suliman and Zhang 2016).
(3) Image Segmentation
To reduce the complexity of the reference VHR image, an image segmentation technique should be used. We propose the use of the multi-resolution image segmentation technique introduced by Baatz and Schäpe (2000) since it has proven as one of the most appropriate techniques for segmenting VHR images in urban areas.
(4) Building Detection and Enhancement
Since the improved LoS-DSM algorithm minimizes the terrain-relief variations, a direct threshold can be applied to detect the off-terrain objects. However, trees and building façades will also be included in this detection, and hence they should be removed.
Fortunately, tree objects can be detected by vegetation indexes. Thus, the Normalized Difference Vegetation Index (NDVI) is selected. The results will still suffer from occlusion due to building façades. Visibility checks for occlusion detection are reviewed by Egnal and Wildes (2002). Accordingly, the left-Right visibility check is selected in this study for generating an occlusion mask to detect and remove the building façades. In addition to the NDVI and occlusion masks, a few morphological operations can be applied to enhance the detection result.

3. Data and Results

The dataset used for this research is a subset of VHR multi-view stereo images that are acquired by the linear array sensor of the WorldView-2 satellite. These images are acquired off-nadir over a dense urban area in Rio de Janeiro, Brazil (2010). Figure 2(a) illustrates the off-nadir VHR reference image.

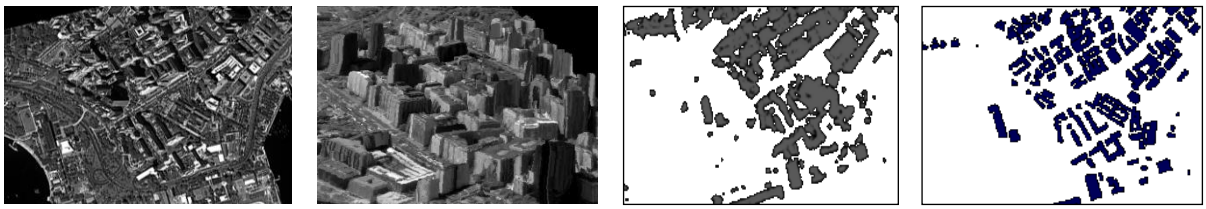
Following the steps identified in the previous section, the reference image was first pan-sharpened using the UNB fusion technique. This step allows using the spectral information in

the removal of vegetation, as well as producing a better segmentation result. Afterwards, the disparity-based LoS-DSM generation algorithm was executed. This algorithm guaranteed achieving accurate co-registration of image and elevation data. A visual inspection confirmed the successful achievement of accurate co-registration, as shown in Figure 2(b).

Next to this result, the pan-sharpened scene was segmented using the multiresolution image segmentation technique. The off-terrain segments were thresholded based on an elevation value close to zero. This is due to the remarkable terrain-relief minimization achieved in the improved algorithm. The un-enhanced detection result is presented in Figure 2(c).

To enhance the building map, two masks were generated: the vegetation mask and occlusion mask. The vegetation mask was generated by thresholding NDVI values calculated based on the pan-sharpened scene, while the occlusion mask was generated based on the left-Right visibility check. After applying these two masks and executing some finishing operations, the final enhanced building detection map was achieved as shown in Figure 2(d).

To evaluate the accuracy of the building detection result, the typical performance evaluation measures for building detection methods were used. These measures are described and used in Suliman and Zhang (2015). Based on these measures, the overall quality measure was found to be more than 90% relative to manually generated reference data. The detection result was almost 95% complete and correct.



(a) (b) (c) (d)

Figure 2. Data used in the test and the achieved results. (a) Reference VHR image, (b) isometric rendered view showing the quality of image-elevation data co-registration, (c) off-terrain level objects detected, and (d) the final enhanced building detection map.

4. Conclusions

This research demonstrates the applicability of the improved algorithm for generating LoS-DSM elevation data through an elevation-based building detection in off-nadir VHR satellite imagery acquired over a dense urban area.

The improved LoS-DSM algorithm was executed over a test dataset. The achieved image-elevation co-registration was very successful based on a visual assessment. Then, the generated and co-registered elevation data were applied in elevation-based building detection. The achieved building map was enhanced based on vegetation and occlusion masks as well as some morphological operations. The quality of the detection was evaluated based on manually generated reference data. The overall detection quality was found to be more than 90% with almost 95% of complete and correct detection. This level of performance in such a challenging dense urban area proves the high success of the disparity-based image-data co-registration as well as the applicability of the developed LoS-DSM elevations to detecting building objects even in off-nadir VHR satellite images acquired over dense urban areas.

We have noticed that the generated vegetation index may produce inaccurate results in the cases of roof gardens and buildings with high NDVI values. Thus, our future work will address this type of limitation.

Acknowledgements

This research is funded in part by the Libyan Ministry of Higher Education and Research (LMHER) and the Canadian Research Chair (CRC) program. The optical data used in this research are provided by Digital Globe for the IEEE-IGARSS 2011's Data Fusion Contest.

References

Baatz M, and Schäpe A, 2000, Multiresolution segmentation: an optimization approach for high quality multi-scale image segmentation. *Angewandte Geographische Informationsverarbeitung XII*, 58(2000):12–23.

Dowman I, 2004, Integration of LIDAR and IFSAR for mapping. *International Archives of Photogrammetry and Remote Sensing*, 35(B2):90–100.

Doxani G, Karantzalos K, and Tsakiri-Strati M, 2015, Object-based building change detection from a single multispectral image and pre-existing geospatial information. *Photogrammetric Engineering and Remote Sensing*, 81(6):481–489.

Egnal G, and Wildes R, 2002, Detecting binocular half-occlusions: empirical comparisons of five approaches. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1127–1133.

Khosravi I, Momeni M, and Rahnemoonfar M, 2014, Performance evaluation of object-based and pixel-based building detection algorithms from very high spatial resolution imagery. *Photogrammetric Engineering and Remote Sensing*, 80(6):519–528.

Suliman A, and Zhang Y, 2015, Development of line-of-sight digital surface model for co-registering off-nadir VHR satellite imagery with elevation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(5):1913–1923.

Suliman A, Zhang Y, 2016, Disparity-based generation of line-of-sight DSM for optical-elevation data co-registration to support building detection in off-nadir VHR satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*. (Under Review)

Zhang Y, 2004, Understanding image fusion. *Photogrammetric Engineering & Remote Sensing*, 70(6):657–661.

A Density-Based Spatial Flow Cluster Detection Method

Ran Tao[1], Jean-Claude Thill[1]

[1]Dept. of Geography & Earth Sciences, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC
Email:{ rtao2; jfthill}@uncc.edu

Abstract

Understanding the patterns and dynamics of spatial origin-destination flow data has been a long-standing goal of spatial scientists. In this paper we introduce a density-based cluster detection method tailored for disaggregated spatial flow data. The basic idea is to first measure flow density considering both endpoint coordinates and flow lengths, and combine it with state-of-art density-based clustering methods. We experiment with a carefully designed synthetic dataset. The results prove that our method can effectively extract flow clusters from various situations encompassing varied flow densities, lengths, hierarchies and, at the same time, avoid issues of Modifiable Areal Unit Problem (MAUP) of flows endpoints, loss of spatial information, and false positive errors on short flows.

1. Introduction

Spatial flows, also known as spatial interactions (SI) between georeferenced places, have been an enduring study object in a wide range of research fields. With the widespread adoption of location-aware technologies and the global diffusion of geographic information systems (GIS), spatial interaction data have been enriched in several respects including volume, type, availability, ubiquity, and spatiotemporal granularity (Yan and Thill 2009; Guo *et al*. 2012). While it brings unprecedented opportunities to improve our understanding of SI processes and thus enriching SI theories, it also brings the analytical challenge of developing more data-drive approaches tailored for SI data (Yan and Thill 2009).

As a common data mining technique, cluster detection has proved useful in exploratory analysis of large sets of spatial flows. One approach measures the spatial relationships among origins and destinations, respectively, before combining them, as the basis for clustering flows. Here, spatial relationships can be contiguity or proximity of origin or destination regions (Guo 2009; Zhu and Guo 2014). However these methods are sensitive to uneven distribution and ad hoc zoning definition of flow endpoints; besides they are prone to false positive errors on short-distance interactions. Another type of methods use flow geometry to bundle nearby ones (Cui *et al*. 2008). While the results usually have desirable visual clarity, these methods compromise through loss of valuable spatial information. In this paper we introduce a new method that not only can extract spatial flow clusters from various situations including varying flow densities, lengths, hierarchies, but also avoids problems like MAUP, false positive errors, and loss of information.

2. Methodology

Of various clustering methods, we choose to design our flow clustering method in the density-based tradition because of its capability to discover clusters of arbitrary shape and to filter out noise. Moreover, density-based methods like OPTICS (Ankerst *et al*. 1999) can effectively reveal hierarchical structures in the data since its byproduct, the reachability plot, is convertible to a dendrogram (Sander *et al*. 2003; Campello *et al*. 2013). Hereafter, we first introduce the proximity metric tailored to spatial flows; then we explain the clustering method step by step.

2.1 Flow Proximity Metric

Cluster analysis critically rests of an appropriate distance metric. Most methods use the Euclidean distance by default without further discussion. Regarding spatial flow data, there exists no 'natural' metric. Tao and Thill (2016) introduced a metric integrating both endpoint coordinates and flow length. The distance between two flows F_i (starting from O_i, ending at D_i) and F_j (from O_j to D_j) is calculated as:

$$FD_{ij} = \sqrt{({d_{Oij}}^2 + {d_{Dij}}^2)/(L_i L_j)^{\alpha}} \qquad (1)$$

Where d_{Oij} and d_{Dij} refer to the Euclidean distance between origins O_i and O_j, and between destinations D_i and D_j, respectively. L_i and L_j are flow lengths. The rationale is that this metric integrates all the spatial elements of a flow, i.e. a pair of endpoints, length, and direction (implicitly). The numerator leverages the accuracy of endpoint coordinates and captures the variation of distances continuously and consistently. The denominator assigns advantage on longer flows given that under most circumstances spatial interaction between distant locations is scarcer due to the "friction of distance" between origin and destination. The exponent α offers flexibility to account for this effect and by default it equals to one.

2.2 Density-based Flow Cluster Detection

Two classical measures of density-based clustering methods are derived from the metric described above. The core distance CoreD (Ester *et al*. 1996) refers to the distance between an object to its *kth* nearest neighbor (2). It measures local density. A small CoreD suggests tight connections to neighbors, thus a likely belongingness to a cluster. Following Ankerst *et al*. (1999), we do not set a search radius threshold for CoreD, like DBSCAN does (Ester *et al*. 1996), as it is usually arbitrary. The minimum cluster size *k* is the only parameter in this method, which is a classic smoothing factor in density estimation. The other measure is the mutual reachability distance MReachD (Campello *et al*. 2013), calculated by (3); it measures the spread between pairs of vertices and serves to separate objects that do not belong to the same cluster.

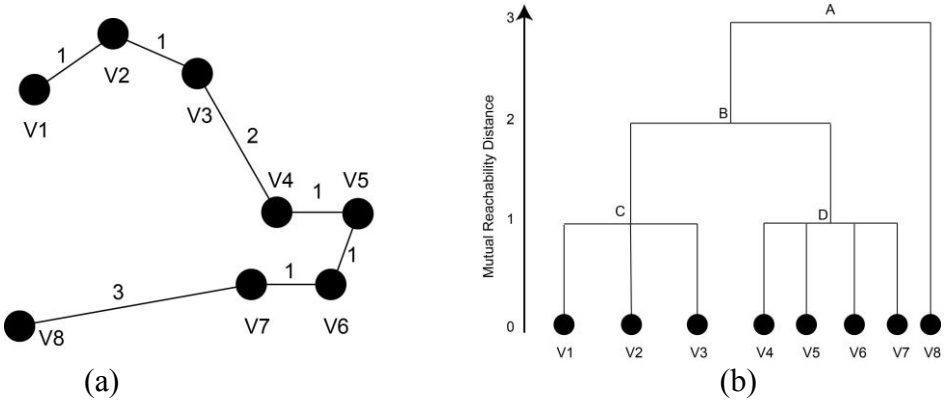
$$CoreD_i = FD_{i,kth\ nearest\ neighbor\ of\ i} \qquad (2)$$

$$MReachD_{ij} = \max(CoreD_i, CoreD_j, FD_{ij}) \qquad (3)$$

A minimum spanning tree (MST) is built in which vertices are the flow objects, which are connected by edges with weight equal to the MReachD between them. In practice, we build this tree by sequentially adding the least-weight edge that connects the current tree to a vertex not yet in the tree, starting from an arbitrarily selected vertex. Figure 1a illustrates a simple case of MST containing eight flow vertices (V1 to V8). Then by sorting the edges of the tree by increasing MReachD value, we can convert it to a dendrogram that connects all vertices in a single hierarchical structure (Figure 1b). However, this hierarchy contains all flow objects without differentiating them. We need a further step to discriminate vertices belonging to a cluster from noise.

We walk through the dendrogram in reverse, from the highest MReachD, and decide at each split whether it should be removed. We use the minimum cluster size *k* as the criterion. If the two children sets of a split are both greater than *k*, we maintain this split as both children sets can be stand-alone clusters. On the other hand, if one of the children sets contains fewer than *k* vertices, we remove this split from the dendrogram, drop the small children set, maintain and keep processing the larger one. We iterate this process until no more split can be removed. In the example of Figure 1, if we set $k = 3$, only split A is removed along with noise V8, while the remaining seven vertices form two clusters; Setting $k = 4$, there would be only one cluster. After processing the whole hierarchy, we end up with a smaller tree with only clusters remaining. We can visualize it as a hierarchical cluster tree starting with the whole dataset, then dropping noise

vertices or splitting to smaller branches at each distance level. The final result would only show the clusters with height and width representing density level and size, respectively.



(a) (b)

Figure 1. (a) Example of MST and (b) dendrogram

2.3 Main Steps of Algorithm

1. Calculate an N by N flow distance matrix with FD (Equation 1);
2. Calculate CoreD values with a selected *k* (Equation 2);
3. Calculate MReachD values (Equation 3);
4. Build a MST based on MReachD and sort it to obtain a dendrogram;
5. Iterate through dendrogram from the highest MReachD. If a children set is smaller than *k*, label and drop it as noise, remove the split and keep processing the large set; if both children sets meet standard *k*, keep the split, label and keep processing both sets.
6. Visualize the final hierarchical cluster tree.

3. Experiments and Discussions

To test our approach we designed a synthetic spatial flow dataset (Figure 2a) of 3,000 flows in various group configurations, labelled from 1 to 8. Groups 1 and 8 both consist of randomly distributed flows, except that the latter is very compact. Groups 2 to 7 are groups of clustered flows with similar direction. The legend indicates the color and size of each group of flows.

Figures 2b, 2c, and 2d are the resulting hierarchical cluster trees with k = 50, 100, and 250, respectively. Overall, the correctness is good as 100% of groups 2 to 7 flows are identified as clustered, while 96% of groups 1 and 8 are removed as noise. In particular, it avoids false positive errors due to short flows like group 8, which was a downside of previous studies using endpoint information separately. Meanwhile, it correctly identifies another group of short flows (group 7) as cluster. The result is not sensitive to the value of minimum cluster size k. For instance the results with k = 50 and 100 are almost identical. However, if a cluster must have at least 250 flows, group 4 is no longer a cluster; it is merged with its close neighbor group 3.

Reporting the inverse MReachD as vertical axis, we can determine at what density level each cluster is identified and at which level it splits to smaller clusters or disappears. This is of great help to reveal the hierarchical structure. For example groups 2, 3, 4, 5 form a single cluster at a lower density level, then split as individual clusters at higher densities. In the real world, this could correspond to clustered flows between two metropolitan areas, within which there are several small flow clusters between districts of each metropolitan area. The cluster size is visualized by width and gradient color. Unlike other clustering methods providing a single flat participation cluster result, we believe it is better to provide the full information and let users decide, since the scale or resolution of "cluster" may be elusive. Some may think group 6 is a cluster given its large size, while others may think only those dense enough (groups 3, 4, 5) are clusters.

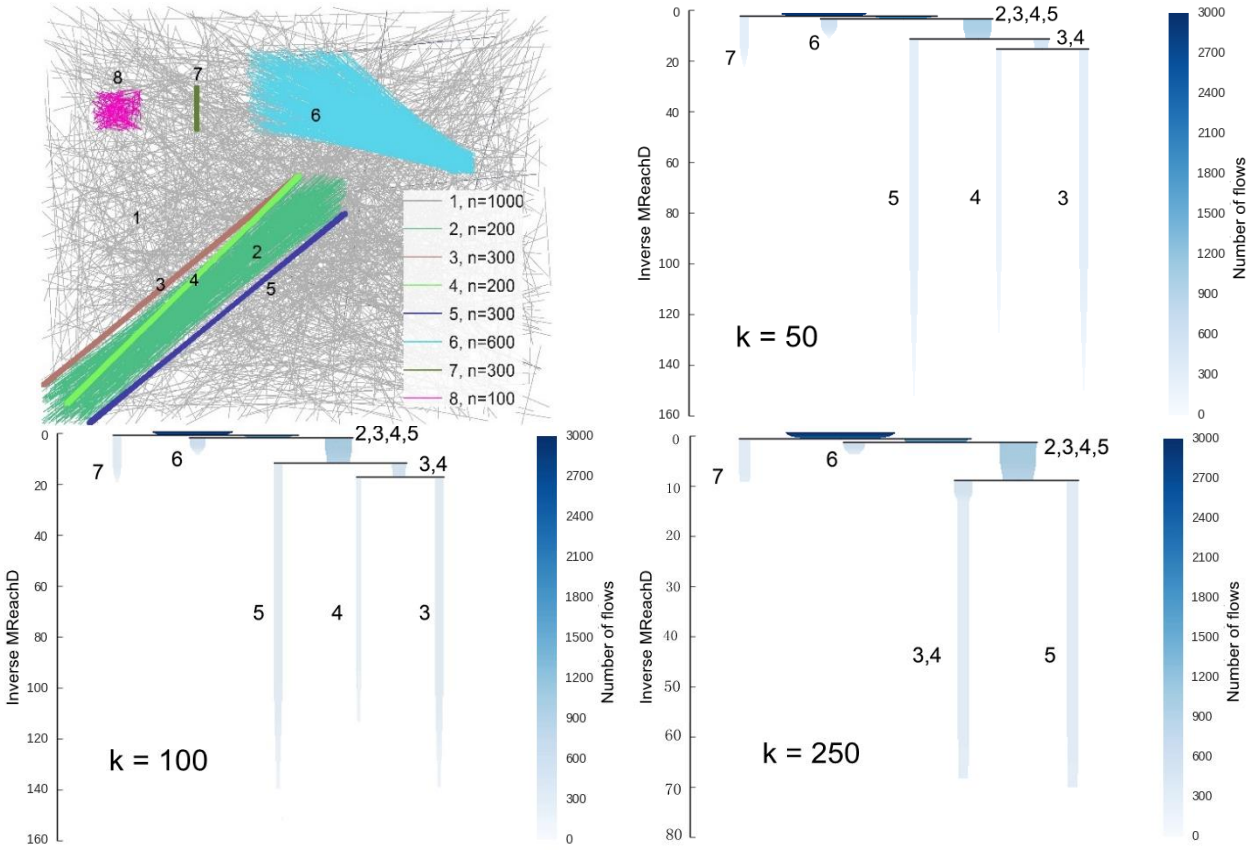


Figure 2. (a) Synthetic flow data (b) hierarchical cluster tree, k=50 (c) hierarchical cluster tree, k=100 (d) hierarchical cluster tree, k=250

4. Conclusions

We developed a density-based clustering approach for disaggregated spatial flows. With a spatial proximity metric tailored for flow data, we extend the state-of-the-art density-based clustering method to the spatial flow context. Our approach effectively identifies clusters of flows of varying lengths and densities and reveals hierarchical structures. It overcomes drawbacks such as loss of information, false positive errors on short flows, MAUP of flow's endpoints.

References

Ankerst M, Breunig MM, Kriegel HP and Sander J, 1999, OPTICS: ordering points to identify the clustering structure. *ACM Sigmod Record* 28(2): 49–60.

Campello RJ, Moulavi D, Sander J, 2013, Density-based clustering based on hierarchical density estimates. *Advances in Knowledge Discovery and Data Mining*, 160–172.

Cui W, Zhou H, Qu H, Wong PC and Li X, 2008, Geometry-based edge clustering for graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6): 1277–1284.

Ester M, Kriegel HP, Sander J and Xu X, 1996, A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD*-96 (34): 226–231.

Guo D, 2009, Flow mapping and multivariate visualization of large spatial interaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6): 1041–1048.

Sander J, Qin X, Lu Z, Niu N and Kovarsky A, 2003, Automatic extraction of clusters from hierarchical clustering representations. *Advances in Knowledge Discovery and Data Mining*, 75–87.

Tao R and Thill JC, 2016, Spatial cluster detection in spatial flow data. *Geographical Analysis*.

Yan J and Thill JC, 2009, Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B*, 36(3): 466–486.

Zhu X and Guo D, 2014, Mapping large spatial flow data with hierarchical clustering. *Transactions in GIS*, 18(3): 421–435.

An Algorithm for Empirically Informed Random Trajectory Generation Between Two Endpoints

G. Technitis[1], R. Weibel[1], B. Kranstauber[2,3], K. Safi[2,3]

[1] Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland
Email: georgios.technitis@geo.uzh.ch
[2] Department of Migration and Immuno-ecology, Max Planck Institute for Ornithology, Am Obstberg 1, 78315 Radolfzell, Germany
[3] Department of Biology, University of Konstanz, Konstanz, Germany

Abstract

We present a method for enabling the researcher to create empirically informed, and thus realistic, random trajectories between two endpoints. The method used relies on empirical distribution functions, which define a dynamic drift expressed in a stepwise joint probability surface. We create random discrete time-step trajectories that connect spatiotemporal points while maintaining a predefined geometry, often based on real observed trajectory. The resulting trajectories can be used a)to generate null models for hypotheses testing, b)as a basis for resource selection models, through the integration of spatial context and c)to quantify space use intensity.

1. Introduction

Random trajectories have been increasingly used in movement ecology since their introduction in the early 1980s (Kareiva and Shigesada 1983), gaining significant popularity in the last two decades (Turchin 1998). A wide range of case studies have used the concept, addressing multiple questions related to movement and space use. The majority of the examples found in the literature, however, share one characteristic: the movement has only one restrictive point, the start. Consequently, the simulation is forced to start at a specific location, but can then move according to the set conditions in the given space. In the real world however, this is not always useful: when studying migration patterns (Codling *et al.* 2010), nest borrowing (Waldeck *et al.* 2008), or fusion of high and low frequency GPS points, etc. the ability to specify the ending point is crucial. Technitis *et al.* (2015) introduced RTG, an algorithm that enables the user to create randomly varying, possible trajectories between endpoints, based on principles of Time Geography.

In this paper we substantially extend this algorithm. We present a methodology to connect two endpoints by generating empirically informed random trajectories, respecting characteristics of the moving object. Our approach is based on core theoretical concepts of Time Geography in combination with the Random Walk movement model, and most importantly, we use empirical data to inform our modelling process.

2. Background

Space-time prisms (STP) assist us in calculating the points accessible in space, given the time budget and the maximum speed of an agent (Kuijpers, *et al.* 2010). The calculated path space (in three dimensions defined by x,y and t), and more specifically its 2-D spatial projection, also known as potential path area (PPA), is a homogenous area within which the trajectory lies. The concept of the STP is very intuitive, although it accounts only for the maximum speed of the mover, gives no information regarding the preference of the mover within the given boundaries, and the result is an area, not an individual trajectory.

Bartumeus *et al*. (2005) highlighted the need for movement ecology to add directional persistence into movement modelling, in order to reproduce realistic animal movement, as noted previously by others (Kareiva and Shigesada 1983; Bovet and Benhamou 1988).
Aiming at this gap, Fleming et al. (2014) described a framework that supports the estimation of auto-correlated movement processes, which was later used in home range estimation. Finally, Technitis et al. (2015) presented an algorithm capable of efficiently generating random trajectories between a given origin and destination, with the least bias possible, within the bounds of the STP, honoring speed and time-budget limitations. The significant assumption of this algorithm is that for each step all space-time reachable points are equally probable to be selected. The trajectories derived from this algorithm are all possible, yet not all of them realistic, as they ignore typical movement characteristics of the moving object.
In summary, the direction that movement modeling is taking in ecology seems clear: starting from random walk models, these were successively extended by STP principles and point-to-point constraints. However, what is still missing is the integration of empirically informed movement parameters that can lead to realistic trajectories.

3. Algorithm

Our algorithm generates trajectories with given set of movement characteristics between two points, in discrete time-steps. The main motivation for creating probable trajectories is that these points represent an arbitrary pair of consecutive fixes of a trajectory, typically placed at a significant distance due to coarse sampling rate. Both points exert an effect on the agent's movement, though of different nature: the probability based on the starting point(A) expresses the 'local' decisions of a moving individual, such as the step-length and direction used for each relocation, performing a correlated random walk with fixed starting point, that assumes no a priori knowledge about the context of the movement (e.g. resource distributions).

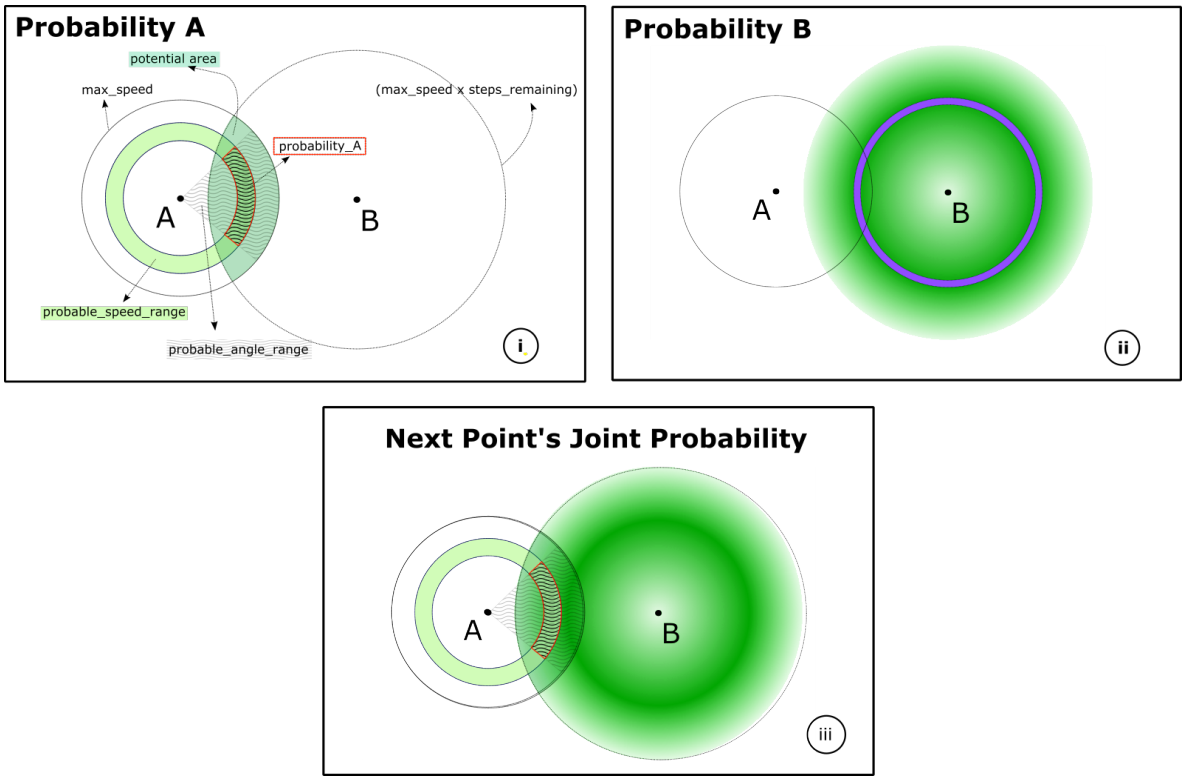


Figure 1. Workflow of generating the relocation of the next point of a trajectory.

On the other hand the probability based on the ending point(B), is a gravitational-type reminder, that the mover should head towards the desired endpoint. Depending on the time left and the current location this force is adjusted and acts as a dynamic drift parameter, applying the necessary bias towards the destination. The model can be described by a mean-reverting Ornstein-Uhlenback process in which "individuals [are] drifting randomly but attracted to an average point" (Smouse et al. 2010). The intensity of the attraction becomes higher, as time is running out, ensuring that the mover is within reach of the destination at any point in time. The flow of the methodology has three steps. First we pre-process the recorded data and calculate the summary statistics of the movement characteristics. Next, starting from the origin, Figure 1.i, we calculate for each time step the probability surface based on the movement characteristics extracted. Then, we calculate the attraction to the destination for the entire study area (Figure 1.ii) given the remaining number of steps to the end point, resulting in a probability surface for each time step. The number of time-steps(n) is the quotient of the duration of the walk and the simulation time interval set by the user. The last step is to combine the two surfaces into a joint probability (Figure 1.iii) out of which we sample the next location of the individual. The procedure is repeated over all time-steps, generating in a new trajectory.

4. Results and Discussion

The proposed algorithm was evaluated in depth on both synthetic and real data. We show an example using synthetic data, where we created three template trajectories using a correlated random walk, out of which our algorithm derived the empirical distributions of movement characteristics.

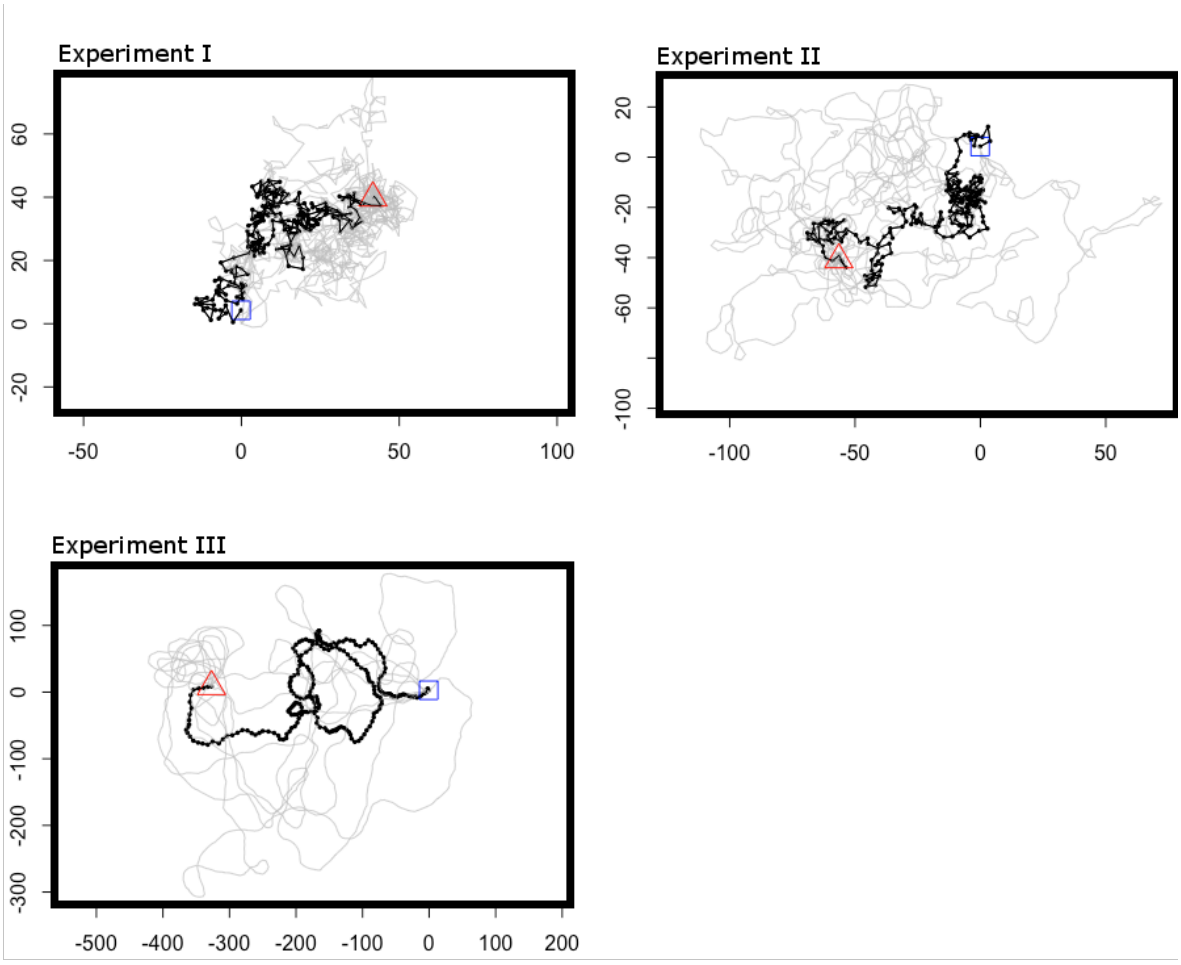


Figure 2. Sample of the results generated (gray) out of three initial datasets (black).

Each template resembles a different behavior of a potential animal, namely migration, foraging and opportunistic behavior. Figure 2 shows the simulated trajectories in gray and the created trajectories in black. While these illustrations suggest in a visual way that the simulated trajectories indeed resemble the original ones in their characteristics (though each takes a different path, as expected), we also conducted an in-depth statistical evaluation to establish that the empirical distributions of the generated trajectories on average match those of the original templates. The next step for this work is to optimize the algorithm's performance and incorporate the context's effect on the mover's behaviour.

Acknowledgements

This research represents part of the PhD project of the first author. Funding by the Swiss State Secretariat for Education, Research and Innovation (SERI) through project CASIMO (C09.0167) is gratefully acknowledged.

References

Bartumeus F, da Luz M G E, Viswanathan G M and Catalan J, 2005, "Animal Search Strategies: A Quantitative Random-Walk Analysis." Ecology 86 (11): 3078–3087.

Bovet P and Benhamou S, 1988, "Spatial Analysis of Animals' Movements Using a Correlated Random Walk Model." *Journal of Theoretical Biology* 131: 419–433. doi:10.1016/S0022-5193(88)80038-9.

Codling E, Bearon R , and Thorn G J, 2010, "Diffusion about the Mean Drift Location in a Biased Random Walk." *Ecology* 91 (10): 3106–3113. http://.ncbi.nlm.nih.gov/pubmed/21058570.

Fleming C, Calabrese J, Mueller T, Olson K, Leimgruber P and Fagan W F, 2014, "Non-Markovian Maximum Likelihood Estimation of Autocorrelated Movement Processes." *Methods in Ecology and Evolution* 5 (5): 462–472. doi:10.1111/2041-210X.12176.

Kareiva P M, and Shigesada N, 1983, "Analyzing Insect Movement as a Correlated Random Walk." *Oecologia* 56 (2-3): 234–238. doi:10.1007/BF00379695.

Kuijpers B, Grimson B and Othman W, 2010, "An Analytic Solution to the Alibi Query in the Space–time Prisms Model for Moving Object Data." *International Journal of Geographical Information Science* 25 (2): 1–13. doi: 10.1080/13658810902967397.

Smouse P E, Focardi S, Moorcroft P R, Kie J K, Forester J D and Morales J M, 2010, "Stochastic Modelling of Animal Movement." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1550): 2201–2211. doi:10.1098/rstb.2010.0078.

Technitis G, Othman W, Safi K and Weibel R, 2015, "From A to B, Randomly: A Point-to-Point Random Trajectory Generator for Animal Movement." *International Journal of Geographical Information Science* 8816 (November): 1–24. doi:10.1080/136588100240930.

Turchin P, 1998, *Quantitative Analysis of Movement: Measuring and Modeling Population Redistribution in Animals and Plants*. *Sinauer*.

Waldeck P, Andersson M, Kilpi M and Öst M, 2008 "Spatial Relatedness and Brood Parasitism in a Female-Philopatric Bird Population." *Behavioral Ecology* 19 (1): 67–73. doi:10.1093/beheco/arm113.

Crowd Sensing System for Public Participation

M. Tenney[1], R. Sieber[1], G.B. Hall[2]

1 McGill University, Montreal, QC, Canada
Email: matthew.tenney@mail.mcgill.ca, renee.sieber@mcgill.ca

2 Esri Canada, Toronto, ON, Canada
Email: bhall@esri.ca

Abstract

We propose a crowd sensing system to capture certain dynamics of public participation in a city. Crowd sensing systems (CSS) attempt to capture the opinions of local publics from web-resources. We define our CSS using a spatially-situated social network graph where users along with different variables, such as time, location, social interaction, service usage, and human activities can be studied and used to identify experts or influential citizens who are relevant to municipal affairs.

1. Introduction

In the fields of engineering and the computational sciences, the term crowd sensing represents a popular area of research (Cardone et al. 2013). Similar to citizen sensing, urban sensing and participatory sensing, we broadly define crowd sensing systems (CSS) as being an integrated hardware and software architecture designed to collect user-generated content for a specified topic, issue or theme. In this paper, we introduce a portion of our conceptual CSS, which describes several social and spatial interactions within a local population (i.e., connections between individuals and locations of communities-of-interest), establishes place-based topics across a city from user-generated content (e.g., geotagged posts from social media), and identifies various forms of activity across specific geographies (e.g., patterns of urban travel). The CSS combines methods of natural language processing, spatial analysis, and graph theory to create a data structure with possible value when used to inform local decision makers.

Our work builds on smart city initiatives, data-science and Web 2.0 literature that seek to revise traditional forms of public participation (Cardone et al. 2013). In particular, these difficulties can be assuaged by integrating data-driven techniques that automatically extract "similar" information (i.e., topical pubic opinions) from user-generated content.

2. Crowd Sensing Systems as Tripartite Network

Public participation in municipal affairs is often seen as a product of stakeholders and interest groups, which are spatially distributed across a city. Choosing to model public participation digitally, requires representing relationships between structured and unstructured content, deriving explicit and implicit social interactions, and inferring frames of context through shared interests and co-location. Like other social networks, our CSS network graph *(G)* contains nodes and edges $G = (N, E)$. The graph is further divided into three subgraphs containing unique node and edge types defined as $U = (Un, Ue)$, $C = (Cn, Ce)$ and $T = (Tn, Te)$, where Un, Ue are nodes and edges of the user profiles in subgraph U; user-generated content forms the subgraph C consisting of nodes Cn and edges Ce; and a "geotopics" subgraph T contains spatially located nodes Tn with temporally weighed edges Te. These connected subgraphs as shown in Figure 1, can form spatially-situated networks constructed from what we call Social Signals (SocSigs, see below). SocSigs information contained in each of the connected sub-graphs include a user-network U (i.e., social-network among

citizens with ties representing relationships from social affiliations or topical/interest similarities), content-network *C* (i.e., unstructured text) and a topic-network *T* (i.e., content nodes derived from secondary data). This latter subgraph contains geographical ties to relevant locations within the city, as well as two extra sets of bridge edges representing semantic similarity to data in the content network and/or related to the interests of users in the user network.

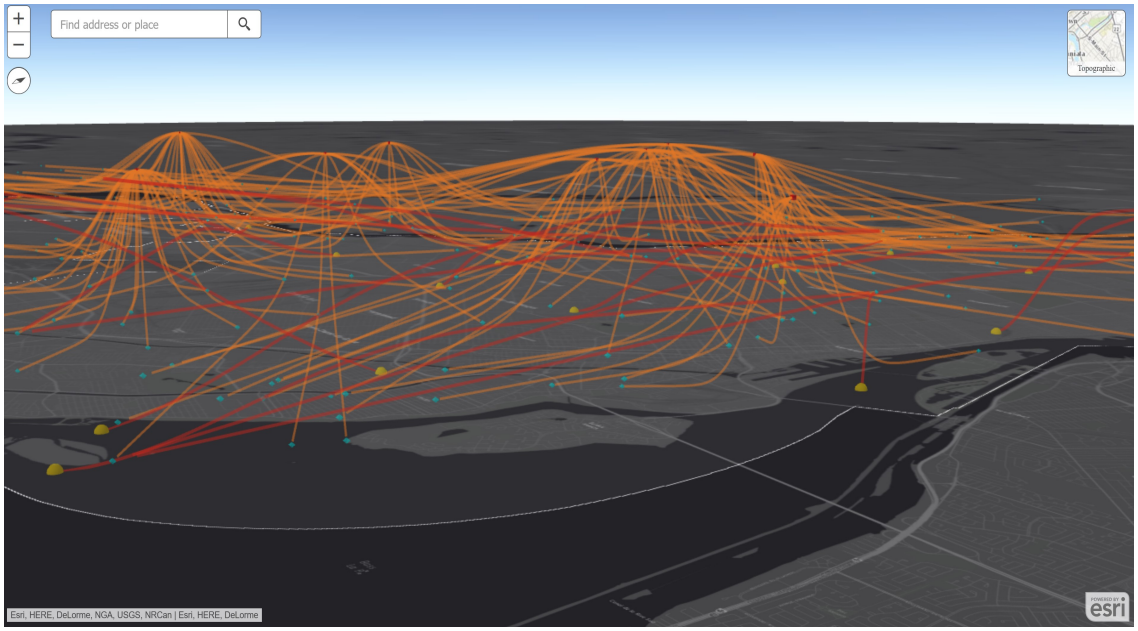


Figure 1 Example of CSS data-structure situated over Montreal, QC. Red nodes are Users in subgraph *U*, with content connections to subgraph *C* (blue) shown in orange, and aggregated 'geotopic" nodes of *T* seen in yellow.

All edges within a single sub-graph are undirected. Bridge edges connecting different sub-graphs are defined as directed-edges so to limit the connectivity of *G*. Using both undirected and directed edges allows for calculating different network measures on specific graph components (De Meo et al. 2014). For example, network communities can be found either by using only connections in the *U* sub-graph or by including all "in-edges" to the content-graph *C* as available paths between users. A mixed-edge graph design containing undirected and directed edges should provide a network-structure that is sensitive to variances in interaction-flows at both local and global levels (De Meo, Ferrara, Fiumara, & Provetti, 2014).

2.1 Relevant Social Signals: A Balance of Context and Content

SocSigs are informative signals that can directly or indirectly provide contextual meaning for interactions, relationships, and behaviours observed from user-generated content (Sheth 2009; Golbeck 2013). We chose four SocSigs variables relevant to our efforts of capturing interests held by a local network members. The first is *content*, which contains unstructured text (e.g., social media posts and status updates). The second is *users*, who are seen as the producers and consumers of the content. We include available characteristics about them, as well as connections to other people and content (e.g., "friends" or "followers", "Likes" or "Shares"). The third is *space-time*, which joins location and time attached to a collected dataset where geotagged content is collected at (*x*,*y*). In this instance, the locational content is time stamped with the time it is created. Fourth, is *strength* which includes the frequency of spatial and social connections among different users, their locations, and topics found in their content.

2.2 Citizen Centrality and Social Influence

Centrality, in graph theory, is a measure of a node's importance to the structure of a graph (Scott and Carrington 2011). Centrality has been used for identifying influential people (i.e., opinion leaders) in various kinds of social networks (Golbeck 2013). It is relevant to public participation because participatory activities can be understood, in part, as multiple actors interacting in relational systems. The centrality of the interactions about a user-node within a network resembles concepts found in participation literature like "opinion-leaders", gatekeepers, and stakeholders (Dubois and Gaffney 2014).

Centrality measures within our CSS represent the importance of a node in each of the sub-graphs without the inclusion of bridge links to other graph components. We use centrality to find salient content, users, and locations. The more central a node is, the more it can be said to represent important topics or people relative to other topics and people. Node influence determines how a combination of connections both within a sub-graph and to other components represent the leading topics important to citizens and opinion-leaders (Dubois and Gaffney 2014), and provides a means to estimate how the opinions of one user may affect the views of others.

2.3 Community Detection and Description

Community is often viewed as possessing a certain physicality, for example, a jurisdictional bounding of city blocks that comprise a neighbourhood. Individuals in this CSS can become community members by expressing shared interests, behaviours, and affiliations throughout the evolution of a network (e.g., increased social similarity between content or co-location patterns). In social network analysis, a community is a set of nodes with strong connections and that contain frequent interactions between members (Fortunato 2010). Community detection in our CSS attempts to decompose a complex network into groups of nodes and edges that are densely connected. Using either direct edge-connections or by including similarity measures between graph-components, grouped elements are considered to have similar interests including topics, activities, and locations (Clauset, Newman, and Moore 2004). Communities can overlap if they comprise a threshold number of nodes or edges that are members of two or more communities.

3. Conclusion

Automatically connecting citizens and governments is a form of automated public participation, which can be provided by a CSS (Cardone et al. 2013). We see a CSS as a computational instrument, composed of computers, sensors, software and algorithms. The instrument can automatically harvest posts, locations, times, and connections among streams of citizen's data to derive insights on the public intent. This process aligns with a big data and smart city vision as data provide access to localized "Citizen Sensor Networks" to computationally facilitate public participation (Koch et al. 2013). Future work will be the critical investigation of these systems, the algorithms applied to these network structures and datasets, and the implications of transitioning to a "coded form" of public participation.

Acknowledgements

We are grateful for the support from the following funders: SSHRC grant 895-2012-1023 "How the geospatial web 2.0 is reshaping government-citizen interactions" and Mitacs Accelerate PhD fellowship co-funded by Esri Canada Limited.

References

Burke, J. A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M. B. (2006). Participatory sensing. Center for Embedded Network Sensing.

Cardone, G., Foschini, L., Bellavista, P., Corradi, A., Borcea, C., Talasila, M., & Curtmola, R. (2013). Fostering participaction in smart cities: a geo-social crowdsensing platform. IEEE Communications Magazine, 51(6), 112–119.

Clauset, A., Newman, M. E. J., & Moore, C. (2004). Finding community structure in very large networks. Physical Review E, 70(6), 66111.

De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2014). Mixing local and global information for community detection in large networks. Journal of Computer and System Sciences, 80(1), 72–87.

Dubois, E., & Gaffney, D. (2014). The Multiple Facets of Influence Identifying Political Influentials and Opinion Leaders on Twitter. American Behavioral Scientist.

Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3–5), 75–174.

Gabrys, J. (2014). Programming Environments: Environmentality and Citizen Sensing in the Smart City. Environment and Planning D: Society and Space, 32(1), 30–48.

Golbeck, J. (2013). *Analyzing the social web* (First edition). Waltham, MA: Morgan Kaufmann is an imprint of Elsevier.

Koch, F., Cardonha, C., Gentil, J. M., & Borger, S. (2013). A Platform for Citizen Sensing in Sentient Cities. In J. Nin & D. Villatoro (Eds.), Citizen in Sensor Networks (pp. 57–66).

Scott, J., & Carrington, P. J. (2011). The SAGE Handbook of Social Network Analysis. SAGE Publications.

Sheth, A. (2009). Citizen Sensing, Social Signals, and Enriching Human Experience. IEEE Internet Computing, 13(4), 87–92.

Curating Transient Population in Urban Dynamics System

Gautam S. Thakur, Kevin A. Sparks, Robert N. Stewart,
Marie L. Urban, Budhendra L. Bhaduri
The Geographic Information Science and Technology (GIST) Group
Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831 USA
Email: {thakurg, sparksa, stewartrn, urbanml, bhaduribl}@ornl.gov

Abstract

Research efforts in population modelling has proven its efficacy in understanding the basic information about residential and commercial areas, as well as for the purposes of planning, development and improvement of the community as an eco-system. Limited by our current ability to capture the dynamics of population change at a finer resolution of space and time, more or less, such efforts assume static nature of population. Today, more and more people are becoming mobile, traveling across borders impacting the nuts and bolts of our urban fabric. Unfortunately, our current efforts are being surpassed by the need to capture such transient population. It is becoming imperative to identify and define them, as well as measure their dynamics and interconnectedness. In this work, we intend to research urban population mobility patterns, gauge their transient nature, and extend our knowledge of their visited locations. We plan to achieve this by designing and developing novel methods and using VGI data that models and characterizes transient population dynamics.

1. Introduction

Over the last decade, worldwide percentages of transient population i.e. population on-the-move have grown linearly, on account of socio-economic development and rising global conflicts (Worldbank 2016). Characterizing such movement patterns are vital to understand the dynamics of an urban system. Beyond that, privacy-preserved and discrete insight into the diurnal activities of people, where they are, where they go, and how much time they spent at certain locations is vital for first responders in times of emergencies, and for urban planners to develop robust future cities. Traditionally, approaches involved in curating such information have heavily relied on census data, surveys, and simulations models. While they provide an excellent source of baseline statistics - for logistical purposes, they rarely collect the spatio-temporal distribution relating to the dynamics of population movements. Furthermore, they omit non-residential locations, such as business districts, parks, and museums. Social media(Hawelka et al. 2014; Frias-Martinez et al. 2012), cellular data(Di Lorenzo & Calabrese 2011; Calabrese et al. 2010; Calabrese et al. 2011; Calabrese et al. 2013; Isaacman et al. 2012) have proved their efficacies in representing global patterns of human mobility. However, much of the current research has not focused on curating population distribution on a temporal scale (e.g. Census data focuses on residential or night time population). This work attempts to provide a set of guidelines that help infer the change in population numbers at different temporal scales. In this work, first we propose *Transient Population Dynamics Model* that attempts to describe and model the mobility patterns of active population. In addition, it provides approaches to identify and estimate transient population and aid in discovering new locations and the duration of their visits. Second, we utilize this model in estimating the transient population across Australia and compare the differences against respective census data. Also, we intend to discover frequently visited locations in outback and metropolitan areas. We believe this work will generate enthusiasm among researchers and make a case for the importance of curating transient population.

2. Model for Transient Population

The Transient Population Model (TPM) attempts to provide insight into the dynamics of an urban system by estimating population mobility and activities governing them. The purpose of the model is to - i) Identify transient population; ii) Estimate transient population; iii) Discover locations (facilities) containing transient population; and iv) Estimate duration of transient activity. For a given geographical region with population N= *{1,2,3…,n}*, we define the subset of this population $\tilde{N}$=*{1,2,3,…,ñ}* as transient; total active = η, such that $N \geq \tilde{N}$, moving among a set of locations L=*{1,2,3…,l}* (total of Γ locations) in that region at any given time. An individual $\tilde{\mathrm{N}}_i$ is associated with its base location(e.g. home) is transient, if moving from location $l_j \rightarrow l_k$ or already moved to a new location l_k.

2.1 Identify transient population:

The transient population is identified based on the movement patterns between a pair of locations or while away from their base location. Thus, person $\tilde{\mathrm{N}}_i$ is transient when,

$$\boldsymbol{T}(\tilde{\mathrm{N}}_i) \rightarrow \begin{cases} \mathbf{1}; & \boldsymbol{if\ l_j \rightarrow l_k} \\ \mathbf{0}; & \boldsymbol{otherwise} \end{cases} \quad \textbf{(1)}$$

Where *T(.)* is the transient function, representing the movement of $\tilde{\mathrm{N}}_i$.

2.2 Estimate transient population

The aggregation of movements of individual $\tilde{\mathrm{N}}_i$ over an observed time provides the estimate of its mobility patterns in the given geographical region. The total number of movement patterns of $\tilde{\mathrm{N}}_i$ over a time interval t is given by,

$$M(\tilde{\mathrm{N}}_i) = \beta \sum_{s=0;j,k=1;j\neq k}^{\Gamma} (\tilde{\mathrm{N}}_i)_{t_s}^{l_j \rightarrow l_k} \quad \textbf{(2)}$$

for M is the movement pattern profile for $\tilde{\mathrm{N}}_i$ and is β constant to reduce ping-pong effect (micro and short-term mobility)(Chiou 2007). Thus, cumulative transient estimates ***(W)*** can be calculated as the agglomeration of all $\tilde{\mathrm{N}}_i$ over the time t.

$$\boldsymbol{W}(M(\tilde{\mathrm{N}}_i)) = \beta \sum_{i}^{\eta} \sum_{s=0;j,k=1;j\neq k}^{\Gamma} (\tilde{\mathrm{N}}_i)_{t_s}^{l_j \rightarrow l_k} \quad \textbf{(3)}$$

2.3 Discover locations containing transient population

Here we discover a set of destination locations where transient population appears over time (exclude base/home location of the population). The motivation to identify such locations provide an insight into the activity patterns, which are otherwise not captured through census survey. In order to differentiate from the base location, we make an explicit assumption that transient location is characterized by a greater number of unique visitors. The agglomeration (Q) of transient population for a location l_k

$$Q(l_k) = \sum_{i=0;j=0,j\neq k}^{\eta,\Gamma} (\tilde{\mathrm{N}}_i)_{t_s}^{l_j \rightarrow l_k} \quad \textbf{(4)}$$

2.4 Estimate duration of transient activity

The duration of transient activity is the sum of the time duration of movement of individuals in transient. For simplicity, we aggregate the total amount of time individuals' movements

$$\Theta(\tilde{N}) = \sum_{t} \sum_{;j,k=1;j\neq k}^{\Gamma} (\tilde{N}_i)_{t_s}^{l_j \rightarrow l_k} \quad (5)$$

3. Results and Analysis

We have used the proposed model and data from Twitter Streaming and Facebook to estimate transient population and discover locations (facilities) in Australia. Geo-located tweets and check-in information from Facebook Graph API is used in this analysis.

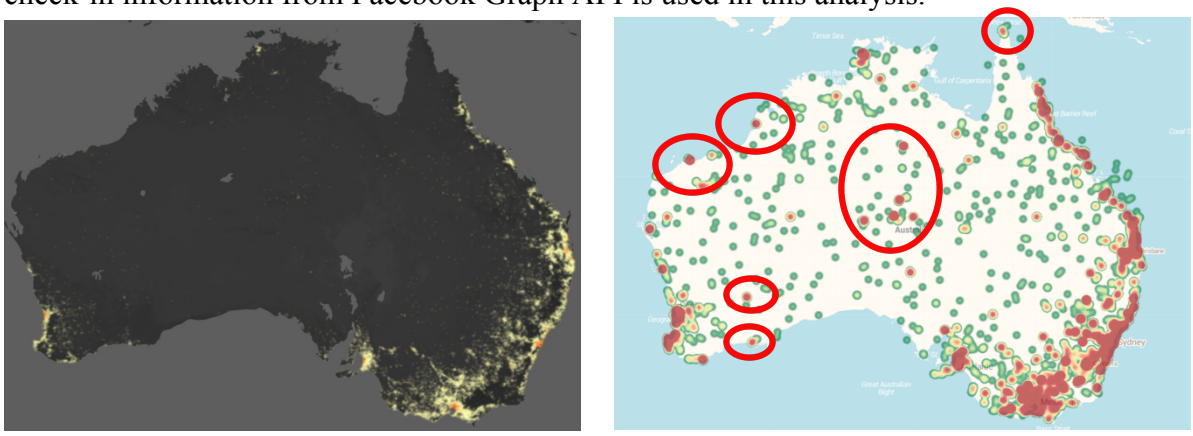


Figure 1: Comparing the Census (a) vs. transient population (b) in Australia

We compare our results to the census of Australia in Figure 1. The observed transient population is shown in population heat map in (b). These statistics are missing in the census map (a). Our preliminary analysis demonstrates the discovery of hidden patches of geographical regions that are not represented traditional techniques (e.g. census and surveys).

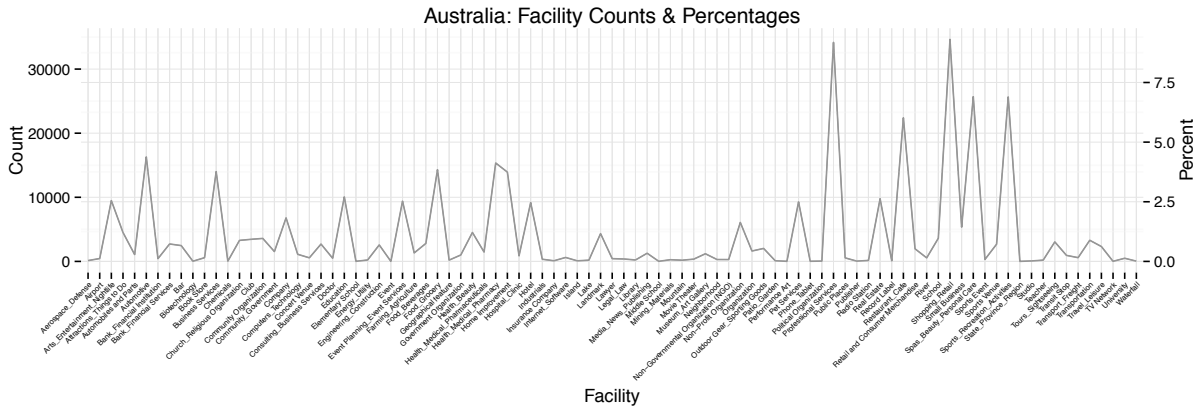


Figure 2: Transient locations discovered in Australia.

We explore each of these regions to generate a high resolution dataset of locations and population. In Figure 2, we show a list of locations and their frequency in Australia, which are mostly commercial and have people actively visiting them. In all, we have discovered more than 100k unique facilities in Australia and clustered them in 50+ categories.

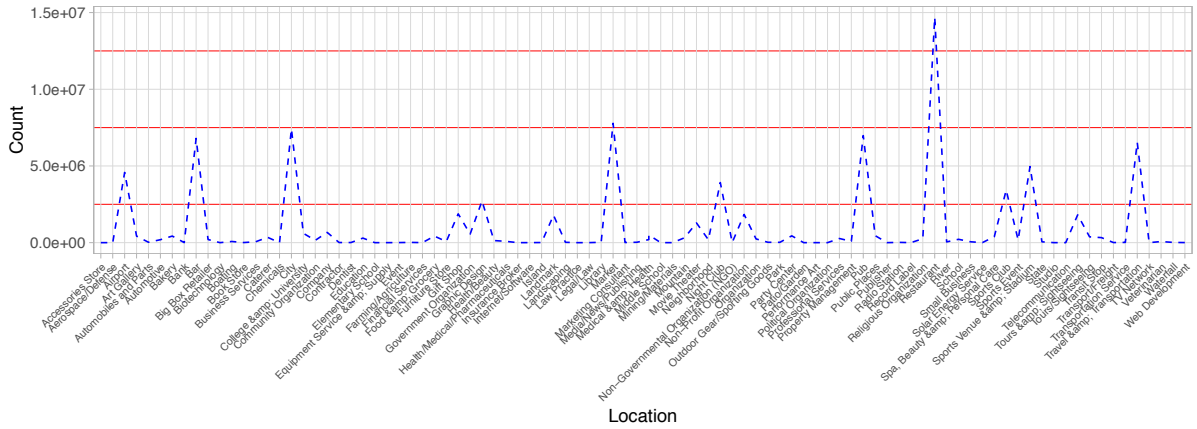


Figure 3: Transient population in Australia.

Figure 3 shows locations and approximate number of people for Australia. 'Restaurants', 'Market Places', and 'Pubs' have recorded the highest number of active population visits during the observed time period. Thus the proposed work also provides insights into ranking location types based on transient population.

4. Conclusion

This research provided an insight into the nature of transient population – their dynamics, mobility patterns, and discovering locations where such population might trend. In addition, we proposed a model to discover and quantify population's structure and emergence. We used VGI data from several social media sources to demonstrate our approach for studying transient population and emerging locations in Australia. In future, we would like to expand on this idea, by first improving the temporal signatures at even finer resolution (of day and hours), and second, to capture seasonal variations appearing over the years. We hope our work will provide a more realistic and accurate approach for generating population signatures and to aid in the analysis, simulation, and design of future urban systems.

Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

References

Calabrese, F. et al., 2011. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1), pp.141–151. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5594641.

Calabrese, F. et al., 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26, pp.301–313.

Calabrese, F., Di Lorenzo, G. & Ratti, C., 2010. Human mobility prediction based on individual and collective geographical preferences. In *13th International IEEE Conference on Intelligent Transportation Systems*. IEEE, pp. 312–317.

Chiou, T., 2007. Method for detecting and reducing ping-pong handover effect of cellular network.

Frias-Martinez, V. et al., 2012. Characterizing Urban Landscapes Using Geolocated Tweets. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, pp. 239–248.

Hawelka, B. et al., 2014. Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3), pp.260–271.

Isaacman, S. et al., 2012. Human mobility modeling at metropolitan scales. In *Proceedings of the 10th international conference on Mobile systems, applications, and services - MobiSys '12*. New York, New York, USA: ACM Press, p. 239.

Di Lorenzo, G. & Calabrese, F., 2011. Identifying human spatio-temporal activity patterns from mobile-phone traces. In *2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 1069–1074.

Worldbank, 2016. World Development Indicators and International tourism, number of arrivals. , 1(1), p.1.

Searching for Common Ground (Again)[1]

J. Thatcher[1], L. Bergmann[2], D. O'Sullivan[3]

[1]University of Washington Tacoma, 1900 Commerce Street, Tacoma, WA 98402
Email: jethatch@uw.edu
[2]University of Washington, Box 353550, Seattle, WA 98195
Email: luke.bergmann@gmail.com
[3]University of California Berkeley, 504 McCone Hall #4740, Berkeley, CA 94720
Email: dosullivan@berkeley.edu

Abstract

At over twenty-five years old, GIScience has been successful academically and institutionally. However, its relationship to one of its 'natural' homes, the discipline of Geography, has often been troubled and uncertain. We suggest that from the founding of GIScience, its close association with Geographical Information Systems (GIS) has contributed to an acceptance of an absolute coordinate space encoded as (x, y[, z, t]) as both a relatively unproblematic and dominant representation of geographical space. We briefly consider how this situation may have arisen, perhaps as an unintended consequence of an originally tactical disciplinary positioning move. However, our purpose here is *not* criticism, but to highlight the many other more minor strands within GIScience, which can provide fertile common ground for renewed conversations between GIScience and Geography. We suggest congruences between less dominant strands of research in GIScience and theoretical concepts in Geography, as an invitation to constructive collaborations.

1. Introduction

"GIScience," as a term, is around a quarter of a century old. It has become institutionally and academically valorized through multiple conferences, journals, and textbooks. We argue that from its inception as distinct from GISystems (GIS), GIScience (GISci) has been framed in ways that have partially delimited the spatial thinking that occurs within it. Specifically, by framing GISci as fundamentally concerned with "spatial information" and "spatial analysis" of such information (Goodchild 1992, 2006), GISci has tended to accept space as absolute, transforming it into data whose 'atomic' units are measurements within a coordinate system of the form (x, y [, z, t, etc.]) (Goodchild et al. 2007), and de-emphasizing alternative conceptualizations of space, time, and process. While this reflects GISci's construction as a "science of *geographical information*" (Goodchild 1992: 38, emphasis ours), we suggest that, in taking a specific concept of spatial information as given, GISci may have unnecessarily distanced itself from broader research in computational geography and in geographic thought.

To make this argument, first, we briefly examine how GISci was framed in an important early conceptualization, focusing on definitions of geographic information. Next, we provide evidence suggestive of continuing ties between these early framings and ongoing research in the field. Finally, we suggest that a reconsideration of the conceptualizations of both space and spatial processes found in broader computational and quantitative geography as well as some theoretical realms of (human) Geography could be valuable to GISci.

[1] Our title is a nod to *A Search for Common Ground* (Gould and Olsson 1982) and the follow up *A Ground for Common Search* (Golledge et al. 1986) which were among the last attempts to bridge the philosophical divides between quantitative and 'critical' geography before the 1990s 'science wars' and, more recently, 'critical GIS'.

2. A Science of Geographical Information

Accounts parsing the histories of GIS and GISci are available elsewhere (e.g., Clarke and Cloud 2000; Schuurman 2000; Mark 2003). If we accept the 1994 definition of GISci offered at the founding of UCGIS which points to "the development and use of theories, methods, technology, and data for understanding geographic processes, relationships, and patterns" (in Mark 2000: 51), then the distinction between 'research about GIS' (development) and 'research with GIS' (use) is clear from the outset (Goodchild 1992: 2006). However, here, we focus on articles by Mike Goodchild from the early 1990s which are perhaps the most influential in the emergence of a 'self-conscious' "science of geographical information" (Goodchild 1992: 38). In emphasizing this strand of GISci's history, we can present only a partial story[2]. Our intent is not a "just so" history, but an analytic reflection on what some consider its foundational moments (see Mark 2003 and elsewhere) and their influence.

This dominant strand suggests that GISci should take "geographical information" as given and as the natural focus of a new science. Even as it was intended to signal a difference between the rote handling of spatial data and the scientific endeavor that motivated such inquiries, to avoid the community being reduced to the "United Parcel Service of GIS" (Goodchild 1992, 31), a focus on understanding spatial data has remained dominant. GISci *has* fostered explorations into what spatial information might consist of (Couclelis 1992; Raper and Livingstone 1995; Goodchild *et al*. 2007; Galton and Mizoguchi 2009), as well as how it might be stored and analyzed; yet, most work in GISci remains far more circumscribed by assumptions about how space can be represented in data than is desirable for GISci to be more "cross-disciplinary" (Mark 2000). While other strands are found in the GISci literature the claim (under the heading 'Geographical Reality') that "the fundamental element of geographical information is the tuple $T = <x, y, z_1, z_2, ..., z_n>$" (Goodchild in Frank and Goodchild 1990) goes largely unchallenged, even as it has since been elaborated (Cova and Goodchild 2002, Goodchild et al. 2007). The limits of this singular perspective have long been recognized. Andrew Frank (in Frank and Goodchild 1990) points to *Graphs*, *Cognitive spaces* and *Imaging schema* as other spatial concepts worthy of expression in data models. Even so, Goodchild's formulation remains dominant (but see Miller and Wentz 2003).

It is difficult to 'prove' this argument, without close reading of key contributions. Here, we present preliminary bibliometric analysis of four leading GISci journals[3], using Web of Science data. CitNetExplorer's citation-network clustering technique (van Eck and Waltman 2014) identifies seven large clusters of research (although we name only six) (Fig 1). There is little here to suggest a sustained research theme of alternative models of space. As Mark and Frank note, "except for geodesics on cost surfaces, non-Euclidean geometries have not made significant inroads into mainstream geographic models" (1996, 17), a claim substantially true today, which also hints at the crux of the problem: the lock-in effects of existing GIS, regardless of the efforts of GIScientists (Miller and Wentz 2003). Of course, computation does require reduction and abstraction of events to textual, numeric, or otherwise programmatically understood representations (Ullman 1997, Drucker 2009). We are not suggesting that one particular spatial representation is "wrong," as that work has been

[2] For example, during the same period, the more open-ended 'spatio-temporal reasoning in geographic space' (Frank et al. 1992), was emerging and soon to become the Conference on Spatial Information Theory (COSIT).

[3] We retrieved 3109 items from *Cartography and Geographic Information Science*/*Systems*, *Geoinformatica, International Journal of Geographical Information Science*/*Systems* and *Transactions in GIS* between 1990 and 2016, which cited a further 2511 items. Clustering these and removing uncited items yielded 3973 items.

fruitful. Rather, we suggest this understanding of space has been prioritized in GISci since its founding.

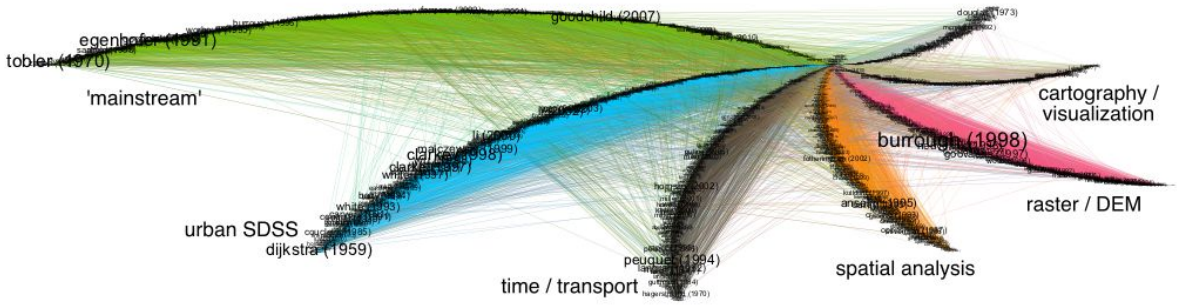


Figure 1. Research clusters for 3973 articles in GIScience.

3. FInding Common Ground (Again)

If GISci is really never again to be "quite the comfortable retreat for the technically minded" (Goodchild 2006, 687) then, while avoiding the oft-revisited, sometimes unproductive debates between and amongst GIS/ci and Critical GIS practitioners (Wright *et al.* 1997; Pickles 1995), we would suggest that there are many lines of research in contemporary GISci that align well enough with theoretical frameworks and concepts in geography to provide firm ground for renewed, closer and above all *constructive* engagement between the fields. A preliminary sketch of this ground is provided in Fig 2.

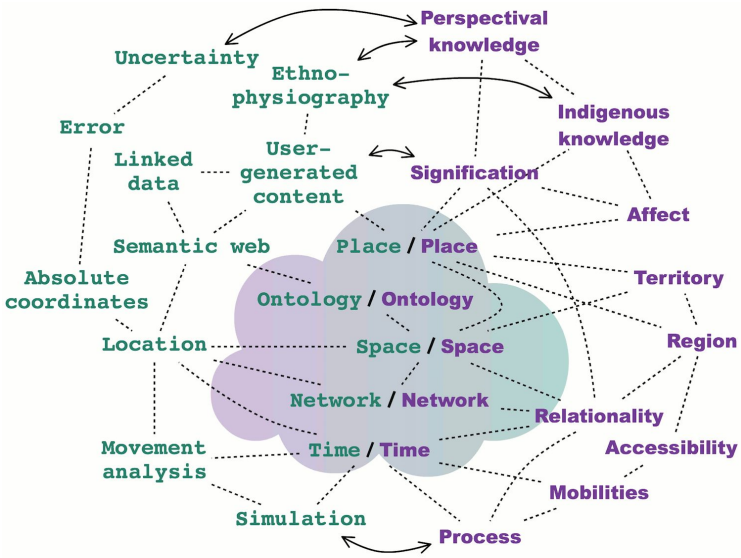


Figure 2. A sketch of common ground. This diagram is not intended to be definitive!

Space does not permit us to expand on all of these congruences (see also Duckham and Sharp 2005). Instead we note that while spatial representations in GIS/ci remain absolute, human geography increasingly takes relational space as the norm (Harvey 2006; Jones 2009). While human geographers often position their critiques of absolute space and its quantitative representation outside of GIS/ci debates and histories, it is unclear why GIS/ci should neglect relative conceptions of space, when they are so relevant to the systems we wish to engage. Computational challenges are not what they were, and much remains to be done, theoretically and in terms of applications, to demonstrate the feasibility and practical utility of such

approaches for mainstream GIS/ci (O'Sullivan et al. forthcoming). We suggest that this is where we might find rewards to renewed consideration of theoretical concepts and empirical examples from geography, even if much relevant research has not yet been formalized as it would need to be in order to become mainstream GISci. At GIScience 2016, we hope to provoke new thinking and new conversations along these lines.

References

Clarke KC, and Cloud JG, 2000, On the origins of analytical cartography. *Cartography and Geographic Information Science* 27(3): 195-204.

Couclelis H, 1992, People manipulate objects (but cultivate fields): beyond the raster-vector debate in GIS. In: AU Frank, I Campari and U Formentini (eds), *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space,* Springer: 65–77.

Cova TJ and MF Goodchild, 2002, Extending geographical representation to include fields of spatial objects. *International Journal of Geographical Information Science*, 16(6): 509–532.

Drucker J, 2009, *SpecLab: Digital Aesthetics and Projects in Speculative Computing*, U Chicago Press.

Duckham M and Sharp J, 2005, Uncertainty and geographic information: Computational and critical convergence. In: Fisher PF, Unwin D (eds) *Re-presenting GIS*. Wiley, Chichester. England: 113-24.

Frank AU and Goodchild MF, 1990, Two perspectives on geographical data modelling. *NCGIA Technical Reports* (90-11), available at http://escholarship.org/uc/item/7zn585sw (accessed August 2016).

Frank AU, I Campari and U Formentini (eds), 1992. *Theories and methods of spatio-temporal reasoning in geographic space*. Berlin ; New York: Springer-Verlag.

Galton A, and Mizoguchi R, 2009, The water falls but the waterfall does not fall: New perspectives on objects, processes and events. *Applied Ontology*, 4(2): 71-107.

Golledge RG, Couclelis H and Gould P (eds), 1988, *A Ground for Common Search*. Santa Barbara Geographical Press, Goleta, CA.

Goodchild MF, 1992, Geographical information science. *International Journal of Geographical Information Systems,* 6(1): 31-45.

Goodchild MF, 2006, Geographical information science: fifteen years later. In: Fisher PF (ed), *Classics from IJGIS: Twenty years of the International Journal of Geographical Information Science and Systems,* CRC Press, Boca Raton: 199-204.

Goodchild MF, Yuan M and Cova TJ, 2007, Towards a general theory of geographic representation in GIS. *International Journal of Geographical Information Science*, 21(3): 239–260.

Gould P and Olsson G (eds), 1982, *A Search for Common Ground*. Pion, London.

Harvey D, 2006, *Spaces of Global Capitalism: A Theory of Uneven Geographical Development*, Verso, London.

Jones M, 2009, Phase space: geography, relational thinking, and beyond. *Progress in Human Geography* 33(4): 487–506.

Mark DM, 2000, Geographic Information Science: Critical Issues in an Emerging Cross-Disciplinary Research Domain. *URISA Journal,* 12(1): 45-54.

Mark DM 2003, Geographic Information Science: Defining the Field. In: Duckham M, Goodchild, MF, Worboys M (eds), *Foundations of Geographic Information Science*. Taylor & Francis, London: 3–18.

Mark DM and Frank AU, 1996,Experiential and formal models of geographic space. *Environment and Planning B: Planning and Design*, 23(1): 3-24

Miller HJ and Wentz EA, 2003, Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers* 93(3): 574–594

O'Sullivan D, Bergman LR and Thatcher JE, forthcoming, Spatiality, maps, and mathematics in critical human geography:Toward a repetition with difference. *The Professional Geographer*

Pickles J (ed), 1995, *Ground Truth*. Guilford Press, New York, NY.

Raper JF, and Livingstone D, 1995, Development of a geomorphological data model using object-oriented design. *International Journal of Geographical Information Systems*, 9: 359-83.

Schuurman N, 2000, Trouble in the heartland: GIS and its critics in the 1990s. *Progress in Human Geography*, 24(4): 569-590.

Ullman E, 1997, *Close to the Machine*, City Lights Books, San Francisco, CA.

van Eck NJ and Waltman L, 2014, CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8(4): 802-823.

Wright D, Goodchild MF, and Proctor D, 1997, Still hoping to turn that theoretical corner. *Annals of the Association of American Gerographers*, 87(2): 373.

Refugee Spatial Awareness: Evidence from Za'atari

Brian M. Tomaszewski[1] and Kenneth J. Tomaszewski[2]

1Department of Information Sciences and Technologies, Rochester Institute of Technology, 152 Lomb Memorial Drive, Rochester, NY 14623 USA
bmtski@rit.edu

2KJT Group, 6 East St, Honeoye Falls, NY 14472 USA
ken@kjtgroup.com

Abstract

This research considers the relationship between spatial cognition and situation awareness, or "spatial awareness" in a refugee camp. Specifically, we present some of the first research on spatial awareness using empirical data collected in the Za'atari Syrian refugee camp in Jordan. Results showed clear spatial awareness differences between male and female in terms of camp infrastructure completeness. This result is likely based on the underlying cultural dynamics of the refugee population. We outline areas for future refugee spatial awareness research such as community asset mapping. A better understanding of how refugees maintain spatial awareness in camp settings can inform GIScience research aimed at (a) identifying new methodologies and educational pathways for supporting spatial awareness in long-term displacement situations, (b) refugee camp design, and (c) space/time representations to ultimately improve the lives of people that are forced to leave their homes and countries due to natural disasters or armed conflicts.

1. Introduction

Forced displacement, whether by natural disasters or armed conflict, is a growing global problem. It raises important challenges for maintaining understanding of the spatial, temporal and thematic aspects of one's circumstances, or spatial awareness (SA). Being forced to leave one's region (such as a town or village) as well as immediate environment (neighbourhood, home), forces the displaced to obtain spatial awareness within new environments.

In this paper, we begin to examine camp-scale spatial awareness for refugees. By "camp-scale", we mean the spatial extent of a settlement specifically built to support survivors of a natural disaster or armed conflict (or both). We present some of the first empirical research on refugee spatial awareness via mental mapping evidence obtained in the Za'atari Syrian refugee camp of Jordan - one of the world's largest refugee camps, hosting an estimated 79, 900 persons of concern who are survivors of the Syrian civil war and where SA a recurring problem [1] (Figure 1).

Figure 1. Location of Za'atai Syrian Refugee Camp in Jordan.

2. Theoretical Framework - Spatial Awareness: Situation Awareness and Spatial Cognition

We use the term spatial awareness (SA) to describe the intersection of situation awareness and spatial cognition. Situation awareness has been defined as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future " [2:36]. Situation awareness is a dynamic concept, changing as situations evolve over time and new circumstances occur. Spatial cognition is the idea of how humans think about the world around us and our interactions with the environment [3]. For the research reported here, SA was used to explain and interpret refugee perceptions of their environment within a space/time context.

3. Methods

We utilized a mental mapping exercise (MME) to establish baseline evidence to investigate refugee spatial awareness. Mental maps are a well-established method for collecting data about SA. Participants were given a pen and blank pieces of paper. After a brief explanation of the idea of a mental map in both English and Arabic, they were then asked to draw the "map in their head" or when they think about Za'atari as a place, what that would look like. Participants were not shown an existing mental map to prevent bias when creating their own mental maps.

Thirty five mental maps were generated. Nine respondents listed their gender as female, seventeen listed their gender as male, and the remaining thirteen listing no gender. Twenty map feature categories ranging from general criteria such as number of building to specific criteria such identification of specific buildings on the mental maps such as United Nations High Commission for Refugees (UNHCR) were initially established by the lead author who had visited the camp on two occasions and was familiar with the camp context. Thematic codes were developed to interpret the mental maps based on (a) the aforementioned initial criteria and (b) existing mental map literature [4]. SA thematic codes related to camp infrastructure completeness such as details about roads, buildings and camp administrative areas known as sectors are reported here. Each mental map was then independently reviewed by two annotators who assigned the mental maps into the thematic codes by recording frequency counts of specific mental map features found on the mental maps.

4. Results

Table 1 shows baseline total results of camp infrastructure completeness.

Table 1. Camp infrastructure completeness totals.

Counts	Number of Streets	Number of Buildings	Number of Districts
Female (n=9)	26	13	0
Male (n=17)	66	383	67
Total (n=26)	**92**	**396**	**67**

Mental map results related to camp infrastructure completeness showed clear differences between males and females. Males showed more detailed views of the camp when compared with females. For examples, on average, males showed five streets compared with three streets for females. It was informally communicated to the lead author by a camp staff member during a field visit to Za'atari that the traditional Islamic culture of many of the Syrian refugees living in the camp restricts the movements of females. The distinctions become greater when looking at the number of buildings that were identified where on average, males would depict fourteen buildings compared with seven buildings for females.

Figure 2 shows a proto-typical example of mental map created by a female that visually demonstrates these points. What makes this mental map proto-typical of female maps is the limited number of places (four) shown. We further examined female mental mapping during a field exercise where a young women was given a GPS and asked to map her daily life at the camp. The map produced during this field exercise was almost identical the mental maps in terms of a limited number of features (four) within a relatively small, bounded area (Figure 3).

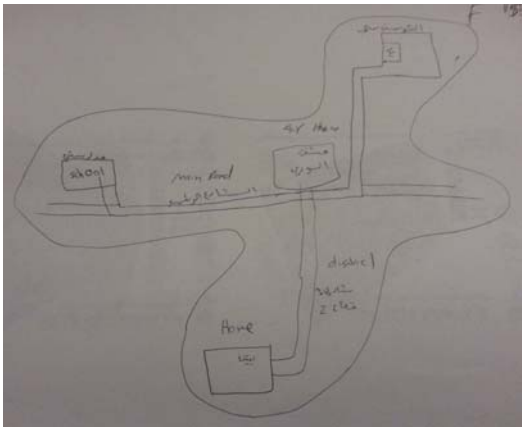


Figure 2. Prototypical female metal map.

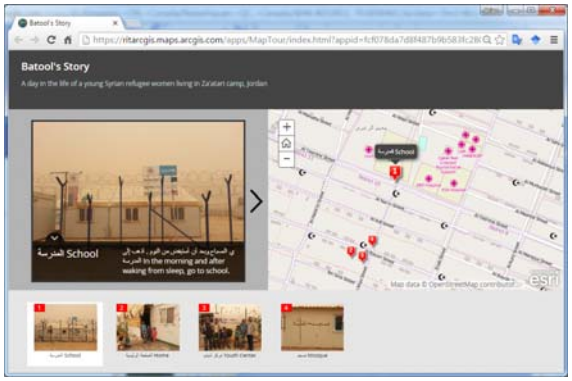


Figure 3. Story map of a young Syrian refugee women.

Perhaps the most striking contrast with camp infrastructure completeness by gender were the numbers of camp districts (an administrative unit) shown on the maps. In this case, males, on average, identified three districts where females identified zero districts. For example, mental maps created by males show often showed a top down view of the entire camp (Figure 4). Mental maps drawn by males are very reflective of the UNHCR map given to refugees when they first arrive in Za'atari (Figure 5).

Figure 4. Proto-typical mental map drawn by a 23 year old male refugee. Numbers show the various districts within the camp.

Figure 5. UNHCR Map given to refugees when they first arrive in the camp.

5. Future Work, Summary and Conclusions

Mental map results clearly indicated differences in gender in terms of camp infrastructure completeness. This result is likely based on the underlying cultural dynamics of the refugee population. One area for future work on refugee spatial awareness would be refugee community asset mapping or how the refugees utilize and think about refugee camp space for daily living and problem solving. For example, Figure 6 is a community asset map made by a male Za'atari refuge who was given a GPS device and mapped spatial community assets or what he considered his 'safe zone' (i.e., the immediate area where he lives) and his 'neighborhood' (i.e., his closest people).

Figure 6. Community asset mapping of refugee safe zones and neighborhoods.

Understanding refugee SA has potential for identifying new methodologies and educational pathways for supporting SA in long-term displacement situations [5]. Furthermore, a better understanding of refugees SA in camp settings can inform research on refugee camp design [6], and ultimately lead to improvement of the lives of displaced people.

Acknowledgements

This research was funded through a grant from the US National Science Foundation and the project US-Jordan planning visit: Integrated wireless and geographic information infrastructure for security in a multi-organizational context (NSF IIA- 1427873).

References

1.Tomaszewski, B., Mohamad, F.A., Hamad, Y.: Refugee Situation Awareness: Camps and Beyond. Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech2015, Boston, MA (2015)
2.Endsley, M.R.: Toward a theory of situation awareness in dynamic systems. Human Factors 37, 32-64 (1995)
3.Mark, D.M.: Human spatial cognition. Human factors in geographical information systems 51-60 (1993)
4.Nishimoto, S.: Evaluating Mental Maps. (2012)
5.National Academy of Sciences: Disaster Resilience: A National Imperative. National Academies Press (2012)
6.Kennedy, J.: Challenging camp design guidelines. Forced Migration Review 23, 46-47 (2005)

Machine Learning on Spark for the Optimal IDW-based Spatiotemporal Interpolation

Weitian Tong[1], Jason Franklin[1], Xiaolu Zhou[2], Lixin Li[1*], Gina Besenyi[3]

[1]Department of Computer Sciences,
[2]Department of Geology and Geography,
Georgia Southern University,
P.O. Box 7997, Statesboro, GA 30460, USA
Emails: {wtong; jf00936; xzhou; lli}@georgiasouthern.edu

[3]Clinical and Digital Health Sciences, CAHS,
Augusta University, Augusta, GA 30912, USA
Email: gbesenyi@augusta.edu

Abstract

To improve current spatiotemporal interpolation methods for public health applications (Li *et al.*, 2010), we combine the extension approach (Li and Revesz, 2004) with machine learning methods, employ the efficient k-d tree structure to store data, and implement our method on Apache Spark (Spark, 2016). Preliminary results demonstrate the computational power of our method, which outperforms the previous work in terms of speed and generates comparable results in terms of accuracy (Li *et al*., 2014). Future research will continue exploring this method to improve the interpolation accuracy and efficiency, with the long term objective of establishing associations between air pollution exposure and adverse health effects.

1. Introduction

To implement the spatiotemporal interpolation method, Li and Revesz (2004) proposed an *extension approach*, which resolves the spatiotemporal interpolation into a higher-dimensional spatial interpolation by treating time as an *asymmetric* dimension in space. Unfortunately, modern work on spatiotemporal interpolation (Pebesma, 2012; Gräler *et al*., 2013; Losser *et al*., 2014; Li *et al*., 2014, *etc*) utilizes simplistic methods to scale the range of the time dimension. In recent work, Li *et al.* (2014) extended the inverse distance weighted (IDW) method (Shepard, 1968) to model the $PM_{2.5}$ exposure risk by scaling the time domain with a parameter c, which is a similar concept to the *spatiotemporal anisotropy parameter* (Gräler *et al*., 2014).

In applying the *extension approach* to the spatial IDW method to interpolate the spatiotemporal data, we arrived at the following formulae

$$w(x, y, ct) = \sum_{i=1}^{n} \lambda_i w_i, \qquad \lambda_i = \frac{(1/d_i)^p}{\sum_{k=1}^{N} (1/d_i)^p},$$

where $w(x, y, ct)$ represents the unknown value to be calculated at the un-sampled location (x, y) and time instance t, c is the spatiotemporal anisotropy parameter, p is the exponent that influences the weighting of w_i, and n is the number of nearest neighbors. Applying k-fold cross validation (k-CV) to the training set can discover the optimal parameters c, p and n for this data set in order to estimate the daily $PM_{2.5}$ concentration values at unknown points. Building upon this work, our method parallelizes the implementation of the original IDW algorithm using

* Correspondence Author

Apache Spark (Spark, 2016) (Figure 1), which is a lightning-fast cluster computing framework and represents the avant-garde of big data processing tools.

In short, our parallelized IDW process broadcasts structured data (the k-d tree) to worker nodes for distributed nearest-neighbor queries and, thus, rapid estimation of pollution levels at unmeasured locations. Naturally, the accuracy of the IDW method for pollution level estimation depends on certain model parameters. Previous work (Li *et al*., 2014) might search a few dozen parameterizations because of limited computational resources. Our system can search tens of thousands of parameterizations in a manageable amount of time. We attempt to learn the best model parameters with brute force, and Apache Spark allows for the application of tremendous force.

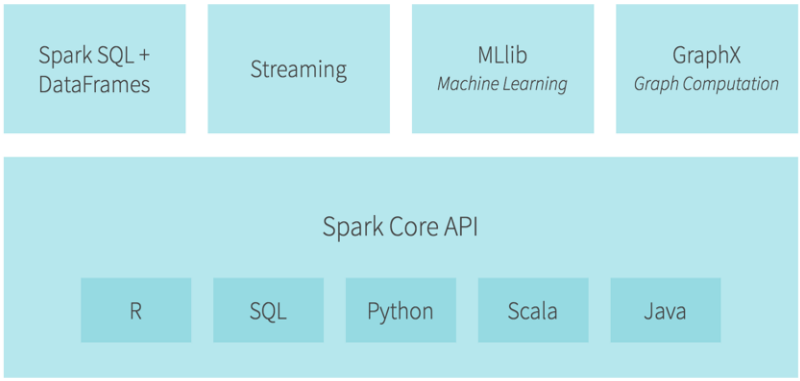


Figure 1. Spark Ecosystem (Spark, 2016)

2. Data Sets

To demonstrate the efficacy and efficiency of our new method, we explored three daily $PM_{2.5}$ data sets for comparison with the results from Li *et al.* (2014). The first data set was air pollution data from the EPA's AQS (Air Quality System) which provided 146,125 $PM_{2.5}$ measurements collected at 955 monitoring sites on all 365 days of the year 2009 (Figure 2).

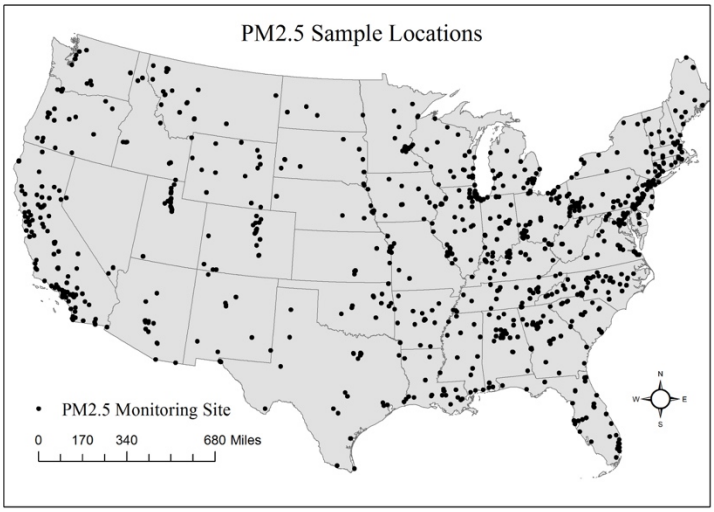


Figure 2. $PM_{2.5}$ Sample Locations

The second and third data sets contain centroid locations of 3109 counties and 207,630 census block groups in the contiguous U.S., respectively. Census block groups (the smallest geographical unit for which the Census Bureau publishes sample data) contain roughly 600~3000 people and are commonly used spatial units to explore population health variables (Iceland and Steinmetz, 2003, Krieger *et al*., 2002).

We train our IDW-based model to estimate the daily $PM_{2.5}$ concentration values in 2009 at the centroid locations (the second and third data sets) for the contiguous U.S., using the existing $PM_{2.5}$ measurements (the first data set) as the training set.

3. Preliminary Results

Two pilot experiments using our general approach have been built. Preliminary results demonstrate that our method and implementation is extremely fast compared to previous work while achieving better prediction accuracy. Since our current method follows the work by Li *et al.* (2014), the main contributions are the larger learning ranges for the parameters in the model and the employment of the cutting-edge technique – Spark. The experiment details are shown as follows. The summaries also include runtimes on Spark with the equivalent time it would have taken in a sequential procedure (a statistic tracked by Spark).

Experiment 1: We exactly follow the work by Li *et al.* (2014), where the *spatiotemporal anisotropy parameter* c was fixed as 0.1086 and 45 parameter configurations were selected for inspection. The learning task in our system only took 2.3 minutes on Spark (70 minutes in sequential time). Since Li *et al.* (2014) did not provide time consumptions for this learning process, we are not able to compare our result with their outcome. For the prediction stage, where the daily $PM_{2.5}$ concentration values at the centroids of counties and census block groups are estimated, our implementation only took about 8% of Li *et al.* (2014)'s record.

Experiment 2: Instead of fixing the *spatiotemporal anisotropy parameter* c, we search for the optimal value. The parameters considered here include c, n, and p. Furthermore, each parameter configuration was run across three 10-CV partitions with the resulting error statistics averaged (to reduce the effect that using a particular partition might have). This set up amounted to 16,848 configurations that were tested in 144.6 hours (4,694.3 hours in sequential time). As expected, the prediction accuracy, measured by MARE, is increased from 1.2058 to 0.3791. This result is actually better than the current best accuracy under the 10-CV, which is 0.3866 and was produced by Li et al. (2012)'s shape-function-based method.

We are confident that our experiments will efficiently learn the optimal parameters, and thus improve the estimation accuracy of the interpolation model, helping us to definitively establish more accurate associations between air pollution exposure and adverse health effects.

4. Future Work

Future research will extend our machine learning approach on Spark in the following four directions: (1) scanning a wider parameterization space and further optimizing search methods for the parameter configurations, (2) exploring alternate machine learning methods such as *leave-one-out cross validation* and *random forest*, (3) attempting other spatiotemporal methods such as *shape function* and *Kriging* based methods, and (4) analyzing other data sets such as real-time hourly air pollution data from the AirNow government website service that provides hourly updates of pollution measurements data from sites across North America.

Acknowledgements

We would like to thank Brandon Kimmons, Director of Computational Research Technical Support at Georgia Southern University, for helping us set up Spark. Franklin, Tong and Zhou were supported in part by funds from the Office of the Vice President for Research & Economic Development at Georgia Southern University.

References

Gräler B, Rehr M, Gerharz LE and Pebesma E, 2013. Spatio-temporal analysis and interpolation of PM10 measurements in Europe for 2009. *ETC/ACM Technical Paper*.

Iceland, J, and Steinmetz, E, 2003. The effects of using census block groups instead of census tracts when examining residential housing patterns. *Bureau of the Census*.

Krieger, N, Chen, JT, Waterman, PD, Soobader, MJ, Subramanian, SV, and Carson, R, 2002. Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? The Public Health Disparities Geocoding Project. *American journal of epidemiology*, 156(5):471--482.

Li L and Revesz, P, 2004. Interpolation methods for spatiotemporal geographic data. *Computers, Environment and Urban Systems*, 28:201–227.

Li L, et al., 2012. Estimating Population Exposure to Fine Particulate Matter in the Conterminous U.S. using Shape Function-based Spatiotemporal Interpolation Method: A County Level Analysis. *GSTF: International Journal on Computing*, 1:24–30.

Li L, Zhang, X and Piltner, R, 2010. An application of the shape function based spatiotemporal interpolation method on ozone and population exposure in the contiguous U.S. *Journal of Environmental Informatics*, 12:120–128.

Li L, Losser T, Yorke C and Piltner R, 2014. Fast Inverse Distance Weighting-based spatiotemporal interpolation: a web-based application of interpolating daily fine particulate matter $PM_{2.5}$ in the Contiguous U.S. using parallel programming and k-d tree. *International Journal of Environmental Research and Public Health*, 11(9): 9101-9141.

Losser L, Li L and Piltner R, 2014. A spatiotemporal interpolation method using radial basis functions for geospatiotemporal big data. *In Proceeding of the 5th International Conference on Computing for Geospatial Research and Application*, Washington DC, USA, 17-24.

Pebesma E, 2012. Spacetime: spatio-temporal data in R. *Journal of Statistical Software*, 51(7):1–30.

Shepard D, 1968. A two-dimensional interpolation function for irregularly spaced data. *In Proceedings of the 23nd National Conference ACM*, 517-524.

Spark, 2016. https://databricks.com/spark/about.

Spatial accuracy measures of soft classification in land cover

N. Tsutsumida[1] and A. Comber[2]

[1]Graduate School of Global Environmental Studies, Kyoto University
Email: naru@kais.kyoto-u.ac.jp

[2]Department of Geography, University of Leeds
Email: a.comber@leeds.ac.uk

Abstract

Accuracy of land cover maps is important for map users. The soft classification of land cover has been developed for avoiding mixed pixel problem, however the proportional map is traditionally assessed only by a global measure, such as R-squared and root mean square error (RMSE), lacking local information of accuracy. We developed a way of local measures of accuracy employed by a geographically weighted (GW) model. GW-Rsquared and GW-RMSE are locally assessed a soft classification map of urban agglomeration as a case study. Lower accuracies are found at the edge of urban boundary surrounding the core of the urban area and such local information is valuable for a deeper understanding of spatial accuracy.

1. Introduction

Land cover maps are important for those who are interested in climate change, biodiversity and anthropogenic impacts on terrestrial environments and the accuracy of the map is an important consideration. Traditionally land cover maps classified as categorized classes (hard classification) are assessed by building a confusion matrix that compares predicted and observed classes, with predicted classes derived from the classification and observed classes from independent validation data. Measures of user, producer and overall accuracy and kappa index are calculated from the matrix. The reliability of land cover data classified using continuous measures such as fuzzy set memberships (soft classification) is frequently assessed using measures such as R-squared and root mean square error (RMSE) (Chen *et al.* 2010; Tsutsumida *et al.* 2016; Yuan *et al.* 2008). However these measures only provide global measures of reliability and accuracy, and they do not take spatial configuration into account. Local assessments would be valuable for a deeper understanding of spatial accuracy.

2. Background

Spatial accuracy assessments for hard classification of land use and land cover have been considered by some previous studies (Foody 2005; Pontius *et al.* 2011). In particular, geographically weighted (GW) logistic regression model have recently been developed (Comber *et al.* 2012; Comber 2013). These generate local confusion matrices at discrete location in the study area and generate spatially distributed estimates (surfaces) of user, producer and overall accuracy. However, little work has focused on spatially distributed accuracy measures for soft classifications. In this study we develop the spatial accuracy measures of soft classification by determining R-squared and RMSE locally. A GW model is applied to develop such measures spatially.

3. Materials

A map of fractional impervious surface area (ISA) in Jakarta metropolitan areas, the biggest urban agglomeration in Indonesia, in 2012 was used in this study (Figure 1). This map was

one of annual ISA maps inferred by a random forest regression over space and time during the period 2001-2013 produced by Tsutsumida *et al.* (2016). The ISA rate was documented in this map in the range of 0-100% according to the result of the random forest regression. Randomly distributed independent 401 validation samples constructed by the visual inspection of very high resolution satellite images in Google Earth were used for the accuracy assessment. The global R-squared and %RMSE were 0.557 and 20.18, respectively.

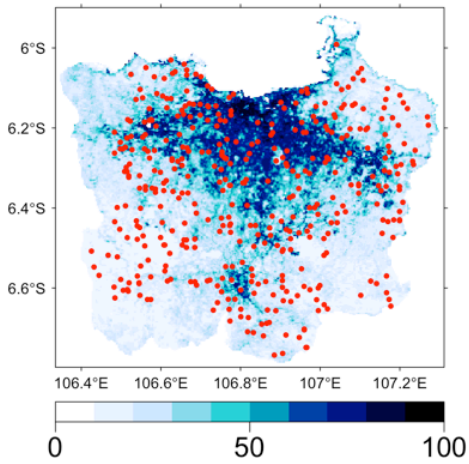


Figure1. Estimated proportion of impervious surface areas and validation samples in the study area.

4. Methodology

4.1 Geographically weighted accuracy measures

In this study, R-squared and %RMSE were considered and incorporated into geographically weighted models. GW-Rsquared can be explained using GW-mean as follows:

$$\text{GW-Rsquared: } Rsq_{(x_i,y_i)} = 1 - \frac{\sum_{j=1}^{n} \omega_{ij}(y_j - x_j)^2}{\sum_{j=1}^{n} \omega_{ij}\left(x_j - m(x_i)\right)^2}, \quad (1)$$

here $m(x_i)$ is GW-mean explained as

$$\text{GW-mean: } m(x_i) = \frac{\sum_{j=1}^{n} \omega_{ij} x_j}{\sum_{j=1}^{n} \omega_{ij}}, \quad (2)$$

x_i and y_i are is observed and predicted value at any location i, respectively, and ω_{ij} accords to a kernel function.

GW-RMSE can be written as:

$$\text{GW-RMSE: } rmse_{(x_i,y_i)} = \frac{\sqrt{\sum_{j=1}^{n} \omega_{ij}(y_j - x_j)^2}}{\sum_{j=1}^{n} \omega_{ij}}, \quad (3)$$

4.2 Bandwidth selection

The kernel type and bandwidth should be determined before implementation of any GW model. A kernel is selected from one of a number of distance functions such as Gaussian, exponential, box-car, bi-square, or tri-cube. The bi-square has widely used due to its simplicity (Harris *et al.* 2014). The kernel gives null weights when the distance of observation is greater than b. The weight decreases as the distance of observation point from the centre of the kernel increases until this distance corresponds to b (Gollini *et al.* 2013). The distance b can be specified either as a fixed distance or a fixed number of considered

data in a kernel. As the validation samples are distributed randomly and irregularly, an adaptive kernel specifying a fixed number of data points was used in this study. While approaches have been developed for automated bandwidth selection with GWmodel and other GWR implementations, procedures for determination of bandwidth optimality for other models such as GW-Rsquared and GW-RMSE, are lacking. Thus, this analysis tested several adaptive kernels with sizes of 5, 10, 15, and 20% of the data points. The results of using 10% of validation samples are shown here because local variations were described well.

5. Results and Discussions

The generation of spatial accuracy surfaces describing the spatial distribution of GW-Rsquared and GW-RMSE are shown in Figure 2. The classification can be regarded as being more reliable in areas where higher values of GW-Rsquared or lower values of GW-RMSE exist. High accuracy was suggested both in the northern-west part of the study area, however the local estimations of accuracy in these locations may be due to the lack of validation data compared to other locations. Lower local accuracies were found at the urban boundary surrounding the core of the urban area by GW-RMSE. This is a complex urban frontier between urban/non-urban areas where it is difficult to estimate impervious surface cover proportions.

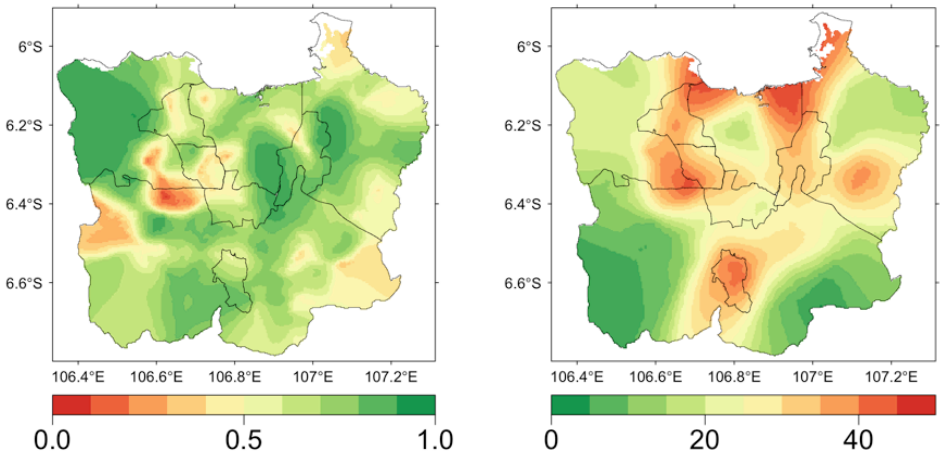


Figure 2. Spatial accuracy measures of proportional impervious surface map estimated by GW-Rsquared (left) and GW-RMSE (right).

6. Conclusions

The extension of global accuracy measures into spatial ones is beneficial to understanding where land cover soft classification maps are accurate or not. In this case study, lower accuracies were found on the urban frontier. Such findings have not identified before. Other measures often used for the assessment of soft classification such as mean absolute error will developed in future work for the general use of accuracy assessments to be more informative.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 15K21086 and SPIRITS program in Kyoto University.

References

Chen J, Zhu X, Imura H, and Chen X 2010, Consistency of accuracy assessment indices for soft classification: Simulation analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, *65*(2): 156–164.

Comber AJ 2013, Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. *Remote Sensing Letters*, *4*(4): 373–380.

Comber AJ, FisherP, Brunsdon C and Khmag A 2012, Spatial analysis of remote sensing image classification accuracy. *Remote Sensing of Environment*, *127*: 237–246.

Foody GM 2005, Local characterization of thematic classification accuracy through spatially constrained confusion matrices. *International Journal of Remote Sensing*, 36(6): 1217-1228.

Gollini I, Lu B, Charlton M, Brunsdo C and Harris P 2013, GWmodel: an R Package for Exploring Spatial Heterogeneity using Geographically Weighted Models. *arXiv*, *1306.0413*.

Harris P, Clarke A, Juggins S, Brunsdon C and Charlton M 2014, Enhancements to a Geographically Weighted Principal Component Analysis in the Context of an Application to an Environmental Data Set. *Geographical Analysis*, *47*: 146–172.

Pontius R and Millones M 2011, Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15): 4407–4429.

Tsutsumida N, Comber A, Barrett K, Saizen I and Rustiadi E 2016, Sub-Pixel Classification of MODIS EVI for Annual Mappings of Impervious Surface Areas. *Remote Sensing*, *8*(2).

Yuan F, Wu C and Bauer ME 2008, Comparison of Spectral Analysis Techniques for Impervious Surface Estimation Using Landsat Imagery. *Photogrammetric Engineering & Remote Sensing*, *74*(8): 1045–1055.

Assessing Spatiotemporal Agreement between Multi-Temporal Built-up Land Layers and Integrated Cadastral and Building Data

Johannes H. Uhl[1], Stefan Leyk[1], Aneta J. Florczyk[2], Martino Pesaresi[2], Deborah Balk[3]

[1] University of Colorado Boulder, Department of Geography, Boulder, CO 80309, U.S.A.
Email: {johannes.uhl; stefan.leyk}@colorado.edu

[2]European Commission – Joint Research Centre (JRC), Institute for the Protection and Security of the Citizen (IPSC), Global Security and Crisis Management Unit, 21027 Ispra, Italy
Email: {martino.pesaresi; aneta.florczyk}@jrc.ec.europa.eu

[3] City University of New York, Institute for Demographic Research and Baruch College, New York, NY 10010, U.S.A.
Email: deborah.balk@baruch.cuny.edu

Abstract

There is an increasing availability of multi-temporal land use and built-up land datasets. However, little research has been done regarding the spatiotemporal uncertainty of these data products. In this work we present an approach that has the potential to be applicable for spatiotemporal evaluation of the novel Global Human Settlement Layer (GHSL) created by automatic classification of global collections of Landsat data recorded in the epochs 1975, 1990, 2000, and 2014. The proposed approach produces the reference data by integrating publicly available parcel and building data and derives agreement statistics with the GHSL.

1. Introduction

The Global Human Settlement Layer (GHSL) project aims to assess the human presence in the planet by estimating the amount of built-up area based on remote sensing data and census data (Pesaresi *et al.* 2015). In Pesaresi *et al*. (2013) the GHSL information production workflow was tested for sensors ranging between 0.5 and 10m spatial resolution, which usually perform very well in detection of built-up areas but are typically constrained regarding data access, redistribution rights, and are typically not available consistently for long periods of time. For these reasons, these data are difficult to use for spatiotemporally uniform and systematic extraction of built-up areas on global, national or even regional level. Therefore, the GHSL workflow was tested with global collections of publicly available Landsat imagery collected in the past 40 years (Pesaresi *et al.* 2016). The Landsat GHSL dataset is available as seamless global mosaic at high spatial resolution (approx. 38m) and for various points in time (1975, 1990, 2000, 2014, see Figure 1a). GHSL data offers promising opportunities for population projections, disaster management and risk assessment (Freire *et al.* 2015; Freire *et al.* 2016), as well as for analysing and modelling urban dynamics and land use change.

Before such novel data products can be made available to the research community, an extensive quality assessment is required. However, such assessments are difficult due to the lack of reliable reference data particularly for earlier time periods and in less developed regions. In this paper we present and discuss a novel approach to evaluate multi-temporal spatial data on built-up land such as GHSL or developed land cover classes using publicly available parcel (cadastral) data integrated with building footprints in the U.S.

The aim of this study is to establish a protocol for future testing the accuracy of fine-scale multi-temporal products derived from automatic classification of satellite data featuring the presence of built-up areas. The selection of the test areas discussed here are driven and constrained by the availability of the reference data and consequently the results derived are

not statistically representative of the whole information contained in the GHSL products but merely serve the purpose of testing the proposed approach and inherent data sensitivities.

2. Data and Method

Open data policy makes cadastral, tax assessment and occasionally building data publicly available – often as GIS-compatible format – for many regions in the U.S. Furthermore, parcel data usually contains rich attribute information related to the type of land use, characteristics of the structure and the year when a structure in a parcel has been established (built year). Based on this information snapshots of built-up parcels can be created for the same points in time used in GHSL (Figure 1b). Building outline data are becoming increasingly available and are used here to spatially refine the geometries of parcel units to the built-up areas. This refinement is very effective in rural areas where parcel units can have large area extents. The integration process of parcel data and building data is complex and has to resolve geometric and topological conflicts to be meaningful. Based on the building footprints enriched with parcel information, the built year attribute is used to create reference snapshots of refined built-up areas that correspond to the GHSL time periods (Figure 1c).

In this study, reference datasets were created for 19 administrative regions (e.g., counties and cities) within the U.S. where the required data are publicly available. The reference layer for each study region was rasterized to create a GHSL-compatible spatial resolution (Figure 1d) and compared pixel-wise to the GHSL built-up areas. Confusion matrices were built to derive several measures of agreement (see e.g., Fielding and Bell 1997) that quantify the overall classification agreement between reference data and GHSL in each study region and for each GHSL category (i.e., four time periods and a not built-up class). Agreement measures were created for areas that changed from not built-up to built-up within each time period, and for the overall (cumulative) built-up area identified for each point in time.

Figure 1. (a) GHSL built-up areas, (b) parcel-based reference data, (c) reference data based on building outlines, and (d) final reference dataset rasterized and resampled to GHSL resolution for a subset of Boulder County (Colorado).

Agreement statistics are expected to vary significantly between urban and rural areas. To examine such trends, the percentage of urban area reported in the 2010 U.S. census county summaries was used to classify the 19 study areas into regions of predominantly urban and rural character. Using thresholds of urban area coverage ranging from 30% to 90% (increments of 10%), the behaviour of agreement measures over time was observed for urban and rural study regions.

3. Results

Several agreement measures were calculated a) regarding the area that changed from not built-up to built-up in each time period, and b) regarding the overall (cumulative) built-up area labelled at each point in time. As an example, the evaluation of changes in built-up land shows that User´s accuracy decreases from recent to earlier epochs but increases for the built-up areas before 1975 (Figure 2a). Since the earliest time period usually covers more built-up area than more recent time periods, the GHSL classifier seems to perform better for built-up land prior to 1975 than for small proportions of change to built-up land in recent time periods. For the cumulative built-up land at each point in time, User´s accuracy tends to decrease for earlier time periods (Figure 2b).

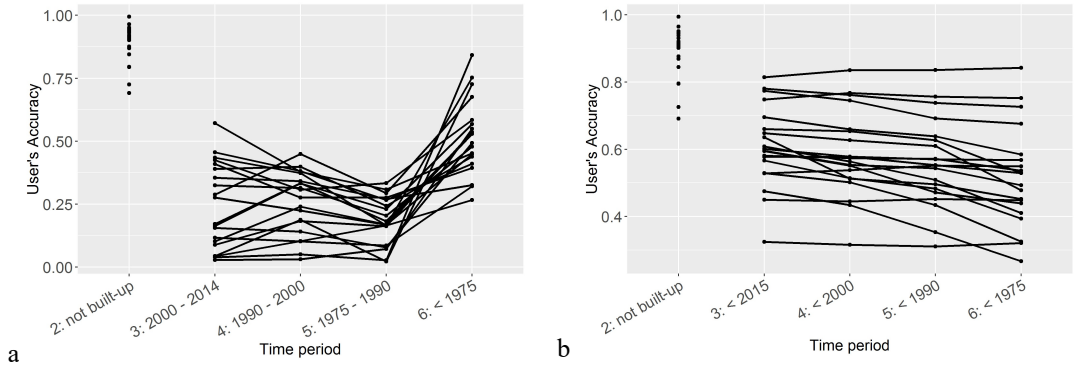
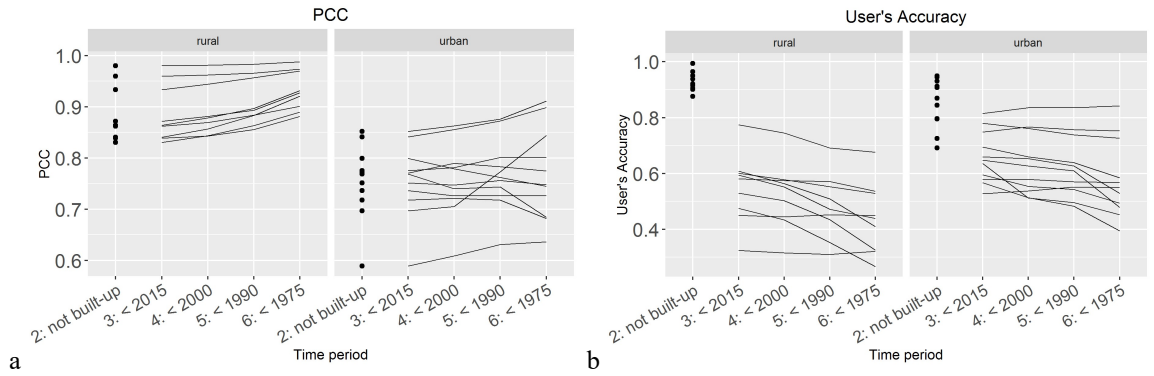


Figure 2. Behaviour of User's accuracy for GHSL built-up labels over time: (a) for changes in built-up area per time period, and (b) for overall (cumulative) built-up areas for each point in time.

As mentioned, the behaviour of GHSL agreement measures over time was analysed separately for urban and rural areas. For a threshold of 50% of urban area coverage the accuracy measures show a different behaviour in urban and rural areas. While the trends of these measures over time are similar for urban and rural areas, the magnitudes differ, considerably. For example, PCC is higher in rural areas than in urban areas (Figure 3a). This effect is particularly strong for the cumulative built-up areas rather than changes in built-up land over separate time periods (not shown). User's accuracy, as another example, shows a similar trend, it is lower in earlier epochs however, it shows higher magnitudes in urban areas than in rural areas (Figure 3b). The results varied with changing threshold for urban/rural distinction but showed similar trends.



a b

Figure 3. Temporal behavior of (a) PCC and (b) User's accuracy for GHSL overall (cumulative) built-up areas in urban and rural study areas using a threshold of 50% urban area coverage.

4. Discussion

The proposed approach allows assessing spatiotemporal agreement between remote sensing-derived built-up labels such as GHSL and publicly available reference data, such as parcel data including built year information and building outline data. It can be observed that agreement measures vary from recent to earlier epochs, and that these variations show different patterns in urban and rural regions. In addition, different agreement levels are obtained for the assessment of cumulative built-up area for each point in time versus the change in built-up land between two points in time. These results give room for further exploration, interpretation and analyses. However, for an adequate interpretation of the agreement measures, thematic uncertainty in the abstract definition of built-up land in GHSL and temporal inaccuracy of the built-up year in the reference data, as well as spatial uncertainty due to displacement errors between satellite-derived labels and the reference data need to be investigated and properly modelled. In addition to these uncertainties, the sensitivity of the agreement measures to the urban/rural thresholds will need to be tested in more detail including aspects of spatial variation in the agreement measures.

Once these issues are examined a first quantitative evaluation of the GHSL will be performed to provide a broader understanding of inherent uncertainty across space and time in GHSL built-up area and inform the future user community on fundamental quality aspects if GHSL is applied to similar settings in other countries where no validation data may be available.

References

Fielding, AH and Bell JF, 1997, A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental conservation*, 24(01), 38-49.

Freire S, Florczyk A, Ehrlich D and Pesaresi M, 2015, Remote sensing derived continental high resolution built-up and population geoinformation for crisis management. *Geoscience and Remote Sensing Symposium (IGARSS)*, *2015 IEEE International*, 2677-2679.

Freire S, MacManus K, Pesaresi M, Doxsey-Whitfield E and Mills J, 2016, Development of new open and free multi-temporal global population grids at 250 m resolution. *Proceedings of the 19th AGILE Conference on Geographic Information Science*. Helsinki, Finland, June 14-17, 2016. (Accepted)

Pesaresi M, Ehrlich D, Ferri S, Florczyk A, Carneiro Freire SM, Halkia S, Julea AM, Kemper T, Soille P and Syrris V, 2016, Operating procedure for the production of the Global Human Settlement Layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. *JRC Technical Report EUR 27741 EN*; doi:10.2788/253582 (online)

Pesaresi M, Ehrlich D, Ferri S, Florczyk A, Freire S, Haag F, Halkia M, Julea AM, Kemper T and Soille P., 2015, Global Human Settlement Analysis for Disaster Risk Reduction. The *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*; 40(7):837.

Pesaresi M, Huadong G, Blaes X, Ehrlich D, Ferri S, Gueguen L, Halkia M, Kauffmann M, Kemper T, Lu L, Marin-Herrera MA, Ouzounis GK, Scavazzon M, Soille P, Syrris V and Zanchetta L, 2013, A global human settlement layer from optical HR/VHR RS data: concept and first results. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5), 2102-2131.

The Cognitive Aspect of Place Properties

Maria Vasardani, Martin Tomko, Stephan Winter

The University of Melbourne, Parkville, 3010, VIC, Australia
Email: {mvasardani, tomkom, winter}@unimelb.edu.au

Abstract

The need to computationally handle the cognitively grounded concept of *place* is fundamental for spatial human-computer interaction. However, there is thus far no consensus about a formal definition of *place*. In this paper, we explore the feasibility of a *constructor* of an abstract data type *place* by exploring a cognitively supported set of properties of place. We study the applicability of Alexander's 15 structural properties of a whole to inform a place property parser of natural language place descriptions.

1. Introduction

The concept of *place* is grounded in common sense and evoked in a variety of contexts. It is fundamental to spatial human communication and increasingly to spatial human-computer interaction. The concept of place becomes central to Geographic Information Science when location information about emergencies needs to be extracted from witnesses' descriptions near real-time. Understanding the cognitive aspects behind natural language (NL) place descriptions is an essential first step for formalising the concept of *place*, so that that it can be used in spatial reasoning and decision making.

Past attempts towards a broadly accepted definition of *place* have not been successful (see Cresswell, 2014). Vasardani and Winter (2015) argued that rather than providing precise but variable definitions of *place* according to application domain and context, it is possible to identify places through a set of properties encoding the concept. As a starting point, they suggested Alexander's (2002) 15 structural properties that characterise a whole and examined how they correspond to properties of the various applications of *place* as studied throughout GIScience. In this work, we set out to explore whether a sub- or superset of these properties is cognitively supported, with the hope that this set can then be used by a *place constructor*—a generator of computational representations of place instances, operating as a function of place properties (i.e., attributes). We use textual place descriptions as a source of these properties.

2. Relevant work

Salient locations that stand out from the *ground* become *places,* thus instances of *objects,* when applying Kuhn's (2012) terminology of core concepts of spatial information. As such, place instances have identity, exist in space, and exhibit spatial, temporal and thematic properties. Arguably, places exhibit a subset or superset of the properties described in (Vasardani and Winter, 2015). Here, we examine which of these 15 place wholeness properties should be part of a cognitively grounded place constructor.

At a basic level, place constructors from text can take the form of parsers. For example, when parsing placenames from geotagged social media text, spatial clusters of placenames form candidate footprints, indicating a consensus about the existence and extent of *places*. These places with no exact boundaries, but rather a fuzzy membership function (Hollenstein and Purves, 2010; Pasley, 2008). Similarly, when extracting <*locatum*, *relation*, *relatum*> triplets using NL parsing methods (Khan *et al.*, 2013), the parser constructs a single place

occurrence based on relationships to other places, but without concern for their grounding in spatial reference systems.

3. Method and Data

Fourteen graduate students of the University of Melbourne completed the following tasks:

1. Imagine a friend is visiting Melbourne for the first time. Provide them with a written description of three meeting places—a place in urban Melbourne, a place in the rural area, and an indoor place. The purpose is for them to be able to recognize the place as they wait for you to arrive.
2. Highlight which (if any) of the 15 properties of Table 1 in (Vasardani and Winter, 2015) you can identify in your written descriptions.
3. Provide any additional properties you can detect in your place descriptions that cannot be classified according to the given set.

The data collection process resulted in 16 urban, 13 rural and 13 indoor place descriptions, as one participant provided three urban place descriptions, instead. In addition, we analyzed the properties of 45 university campus descriptions collected in a previous experiment (Vasardani *et al*., 2013).

4. Analysis and Discussion

Figure 1a shows the most identified properties in the set of urban, rural and indoor descriptions from Task 2, while Figure 1b summarizes the frequencies of each of the 15 properties identified in the set of campus descriptions. These latter descriptions were about a bigger place—a whole campus in contrast to a meeting place. They were also more numerous, hence the number of properties identified.

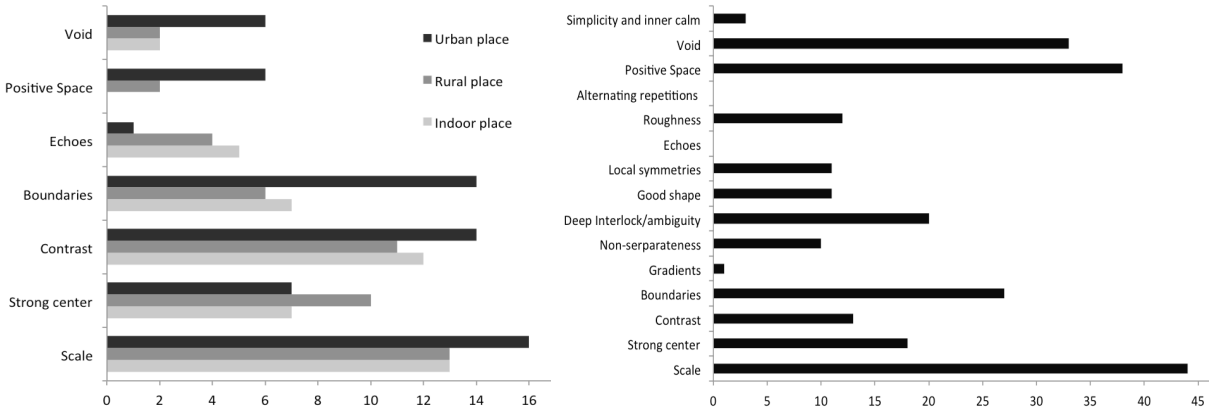


Figure 1: (a) Most identified properties in the place descriptions—urban, rural and indoor; and (b) properties identified in the campus descriptions.

The properties of *scale, strong center, contrast, boundaries, positive space* and *void* are a pronounced subset in both data sets. The intrinsic *scale* property is never specifically referred to, but rather extracted from the resolution of each description type. It was collectively at street and building levels (as per (Richter *et al.*, 2013)) for the urban, rural and campus descriptions, and room/building level for the indoor descriptions. A *pronounced center* appears in most rural descriptions and is well represented in the campus descriptions. The *boundaries* of place seem to play a more significant role in the urban and indoor descriptions. This is not surprising as urban places are structurally denser, thus delineating the boundaries may help separating individual places. In the less dense rural areas, strong centers help provide an identity. Boundaries were also frequently mentioned in the descriptions of the

University of Melbourne city-center campus in the form of distinct street boundaries. In indoor descriptions, boundaries seem essential for the identification of distinct places.

In both sets of descriptions people showed a strong preference for *contrast* features of similar scale (see Winter and Freksa's (2012) contrast sets). According to Tomko and Winter's (2013) formalisation of Lynch's (1960) city elements, these contrast features are non-accessible landmarks in urban, indoor and campus descriptions, perceived mostly as reference points. However, the same contrast features represent accessible nodes in rural places, mostly as the start or end points of transition media (e.g., train, bus, or tram stops). Similarly, in urban descriptions streets are often perceived as edges, separating places, while in rural descriptions streets are mostly paths, connecting distant places.

The property of *positive space*, or the figure-ground relation between artefacts and ground, is most pronounced in descriptions of urban environments and of the campus. Perhaps more surprising is that respondents also referred to *void*, empty spaces of a place (e.g., grassy areas and squares) as often as to buildings.

Amongst properties missing from Alexander's set, place *affordances* were frequently mentioned (Task 3). Hence, functional characteristics should complement structural characteristics in a place formalization (Ortmann and Kuhn, 2010). Some participants referred to signs as part of their place descriptions, and suggested that a separate property should be included. One could argue, however, that signs are either already covered by the placename itself (when the sign is about the place), or they are just another type of landmark, belonging to a specific contrast set.

5. Conclusions

We set out to explore whether there is a cognitively supported set of place properties that could be used to inform a place constructor—in its simplest form, a natural language text parser. To assess our hypothesis we asked 14 participants to think about the properties they use to describe different types of places and compare them against Alexander's 15 structural properties of a whole. We also examined a number of university campus textual descriptions against the same properties set.

The experiment showed people's preference for a subset of Alexander's set with the addition of *affordance*. This suggests that a place constructor should, at the very least, support values of the following properties: {*scale*, *strong center, contrast, boundaries, positive space*, *void*} + {affordance}, although in the lengthier campus descriptions additional properties occur. This exercise also reveals the synergies among different cognitively grounded theories pertaining to *place*. Lynch's elements of the city form can be associated with different properties, e.g., streets act as either edges or paths in urban or rural places, respectively. Elements of contrast sets in a variety of resolutions act as landmarks or nodes. The mention of void, empty spaces counterbalances the detailed descriptions of buildings or artefacts in a recognizable figure-ground relationship.

While preliminary and limited, these results indicate a possible place constructor informed by properties that stand out not only in this cognitive experiment, but also relate to basic place concepts in cognitive GIScience theories. Examination of larger and varied datasets of place descriptions is necessary before a universal place constructor can be proposed.

A place constructor would have to generate unique places. For a text parser that relies on property values, this requirement implies that unique combinations of property values need to be allocated to each place. While this is a necessary condition, it may not be a sufficient one for the creation of uniquely identified places. For instance, it is not yet clear whether and how intra- and inter-place spatial relations that can potentially assist in place identification, become part of a place constructor.

References

Alexander, C., 2002. The Nature of Order: An Essay on the Art of Building and the Nature of the Universe, Book 1 - The Phenomenon of Life. Routledge, Berkeley, CA.

Cresswell, T., 2014. Place: An Introduction. John Wiley & Sons.

Hollenstein, L., Purves, R., 2010. Exploring place through user-generated content: Using Flickr tags to describe city cores. *Journal of Spatial Information Science* 1, 21–48.

Khan, A., Vasardani, M., Winter, S., 2013. Extracting Spatial Information From Place Descriptions, in: Scheider, S. (Ed.), *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place*, COMP '13. ACM, New York, NY, USA, pp. 62–69.

Kuhn, W., 2012. Core concepts of spatial information for transdisciplinary research. *International Journal of Geographical Information Science* 26, 2267–2276.

Lynch, K., 1960. The Image of the City. MIT Press.

Ortmann, J., Kuhn, W., 2010. Affordances as Qualities, in: *Proceedings of the 6th International Conference on Formal Ontology in Information Systems* (FOIS 2010). IOS Press, pp. 117–130.

Pasley, R.C., 2008. Defining Imprecise Regions Using the Web, in: *Proceedings of the 2nd PhD Workshop on Information and Knowledge Management*, PIKM '08. ACM, New York, NY, USA, pp. 105–108.

Richter, D., Vasardani, M., Stirling, L., Richter, K.-F., Winter, S., 2013. Zooming In–Zooming Out Hierarchies in Place Descriptions, in: Krisp, J.M. (Ed.), Progress in Location-Based Services, *Lecture Notes in Geoinformation and Cartography*, Springer Berlin Heidelberg, pp. 339–355.

Tomko, M., Winter, S., 2013. Describing the functional spatial structure of urban environments. *Computers, Environment and Urban Systems* 41, 177–187.

Vasardani, M., Timpf, S., Winter, S., Tomko, M., 2013. From Descriptions to Depictions: A Conceptual Framework, in: Tenbrink, T., Stell, J., Galton, A., Wood, Z. (Eds.), Spatial Information Theory - *Proceedings of 11th International Conference*, COSIT 2013, Scarborough, UK, September 2-6, 2013., *Lecture Notes in Computer Science*, Springer International Publishing, pp. 299–319.

Vasardani, M., Winter, S., 2015. Place Properties, in: Onsrud, H., Kuhn, W. (Eds.), Advancing Geographic Information Science. GSDI Press, pp. 243–254.

Winter, S., Freksa, C., 2012. Approaching the Notion of Place by Contrast. *Journal of Spatial Information Science* 5, 31–50.

Spatial Data Considerations for a Trauma Transport Spatial Decision Support System

Y. Vasilyeva[1], M. J. Widener[1], Z. Ginsberg[2], S. Galvagno[3]

[1]University of Toronto - St. George, 100 St. George Street, Room 5037, Toronto, Ontario M5S 3G3, Canada
Email: katia.vasilyeva@mail.utoronto.ca, michael.widener@utoronto.ca

[2]Kettering Health Network - 8280 Yankee St., Centreville, OH 45458, USA
Email: zginsberg@gmail.com

[3]R. Adams Cowley Shock Trauma Center, University of Maryland - 22 S. Greene Street, Baltimore, MD 21201, USA
Email: sgalvagno@anes.umm.edu

Abstract

With the recent proliferation of sensing and tracking technologies in medical settings, hospitals have the ability to integrate real-time spatial and aspatial data to make important logistical decisions. One area in need of this is trauma transportation, where patient outcomes are sensitive to the mode of transportation and the time it takes to receive care. This research presents a spatial decision support system that uses real-time information to guide medical personnel responsible for making complicated transportation choices with a diverse set of dynamic and static spatial and aspatial variables.

1. Introduction

The use of real-time spatial information coupled with relevant health data to drive decisions during emergency medical scenarios is, in practice, uncommon (Raghupathi and Raghupathi 2014). This is especially true in the context of trauma transportation, where the decision process for selecting the mode of transportation (i.e. helicopter or ambulance) to an appropriate trauma centre in an acceptable amount of time, is relatively unsophisticated and results in questionable choices (Widener et al. 2015). It is important to note that in many severe trauma cases, patients may be many miles away from care, providing for a complex geographic context that includes a range of rural and urban environments. Additionally, trauma patient outcomes are sensitive to transport mode (Brown et al. 2012). Given the quickening pace of adoption of sensing and tracking technologies in the medical community (Raghupathi and Raghupathi 2014; Xu et al. 2014), there is an opportunity to further integrate geographic data to address these inherently spatial problems to maximize the survival of patients and minimize the resource burden on the health care system. While there has been some research using GISystems and Science to analyze trauma transport in the past (Widener et al. 2015), and older examples of GISystems guiding emergency medical services (Peters and Hall 1999), research considering the relationships between modern spatial data feeds relevant to transporting trauma patients has not been thoroughly developed.

This short paper presents a spatial decision support system (SDSS) that considers the various streams of spatial and aspatial data required to generate informed responses to trauma incidents. Special attention is paid to links and feedbacks between dynamically changing and static spatial

datasets, the order of processing these data, and, finally, how decisions are computed, visualized, and communicated via GISystems. With this SDSS, there is a framework that guides hospitals and emergency medical responders to optimal decisions for a diverse array of trauma scenarios, and takes advantage of the increasing number of hospital resources providing real-time information. Additionally, we will demonstrate the utility of the SDSS through testing a variety of trauma transport scenarios using historical and simulated data. However, because the data on trauma incidents were just acquired, only the SDSS and select preliminary results are presented below. Scenario tests are anticipated to be completed by July 2016.

2. Data

While the decision framework developed through this research project can be applied in most North American contexts, the data used to demonstrate the SDSS will be from Maryland, USA, where the helicopter emergency medical services (HEMS) system is publicly funded with a standardized protocol for deploying more expensive helicopter transportation.

A variety of data are used to inform the development of, and test, the SDSS. Table 1 outlines the static and dynamic variables considered in the SDSS, where "static" describes variables considered time-independent and "dynamic" describes variables considered time-dependent. The

Table 1. Variables relevant to the SDSS.

Static Variables	Dynamic Variables
Hospital locations	Patient condition
Helicopter landing sites	Traffic conditions
Helicopter bases	Weather conditions
Ambulance bases	Ambulance positions
Road network	Hospital capacity

first static variable that will be considered is the locations of all trauma centres, accounting for their capacities, described by the American College of Surgeons. Other static variables considered are helicopter and ambulance base locations. Suitable helicopter landing sites are determined by calculating the slope of sufficiently large cleared patches of land (preliminary sites are seen in Figure 1). Additionally, the road network is used to calculate optimal routes for ground ambulances.

Important dynamic variables relevant to trauma transport decisions are patient condition, traffic conditions, weather, ambulance positions, and hospital capacity. After data are processed and incorporated, various dynamic values for these variables will be simulated within the GIS to explore the SDSS's output.

3. Spatial Decision Support System Map

Current protocol for trauma patient transport begins after ground emergency medical services (GEMS), has arrived at the site. Upon evaluation of the patient, the GEMS crew determines whether the patient should be driven to a trauma centre or flown via HEMS. However, the process for determining the appropriate mode of transport can be somewhat arbitrary. Widener et al. (2015) show a significant proportion of patients were flown to the trauma centre despite the

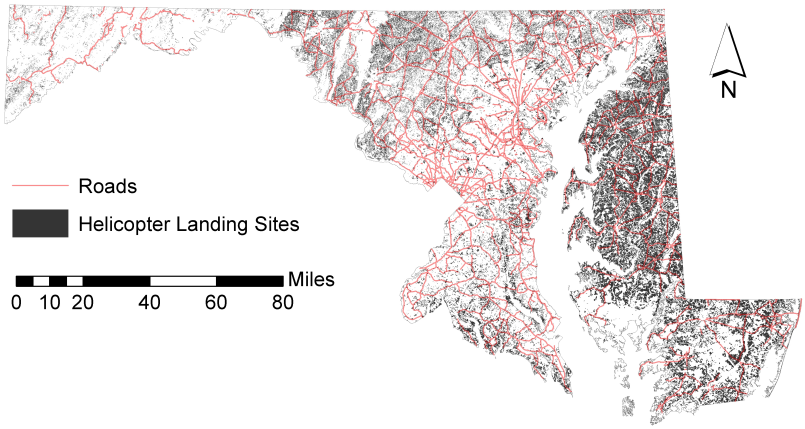


Figure 1. Potential helicopter landing sites with major roads.

travel cost via GEMS being lower. Given that trauma patient outcomes improve if they arrive at a trauma centre within one hour (known as the golden hour (Newgard et al. 2010)) it is critical that a standardized, but flexible, system for decision-making be developed.

The SDSS (Figure 2) is initiated via a user through a Graphical User Interface (GUI) that takes information on the patient's location, condition, and the time the incident was reported.

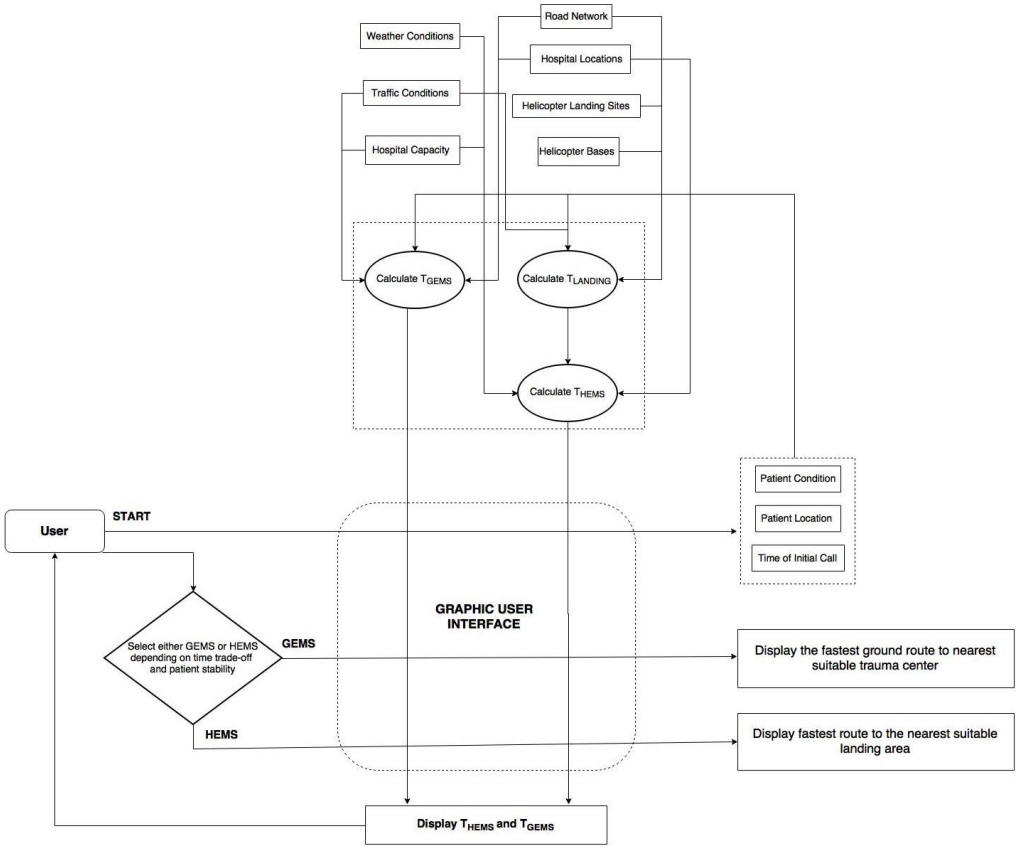


Figure 2. Model of the Trauma Transport Mode Selection SDSS.

With this information, T_{GEMS} and T_{HEMS} are both computed. Prior to computing T_{HEMS}, the travel cost of the time needed to transport the patient from the incident location to a suitable helicopter landing site, $T_{LANDING}$, is also calculated and considered.

Given that every trauma case is unique, the SDSS displays the computed GEMS and HEMS travel times to the EMS crew via the GUI and allows the EMS crew to select a mode depending on the patient's estimated stability and how valuable the EMS crew believes a time-tradeoff between the two modes might be. In this way, the SDSS handles all of the sophisticated aspatial and spatial data behind the scenes and produces a result that the EMS crew can easily combine with their expertise. When the EMS crew decides on a mode and communicates that decision to the SDSS via the GUI, the SDSS outputs one of two displays depending on the EMS crew's decision - if the EMS crew selects GEMS, then the output is a display of the fastest route to the nearest suitable trauma center. If the EMS crew selects HEMS, then the output is a display of the fastest route to the nearest suitable helicopter landing site.

This SDSS explicitly addresses the role of new streams of data and how this information can be processed into a meaningful product useful to an EMS crew. It effectively integrates complex real-time traffic and weather data (acquired through esri's World Traffic Service and the US National Oceanic and Atmospheric Administration), simulated real-time patient data, and static spatial datasets to demonstrate the need for such a decision support system, and delivers a prototype that will guide EMS crews in making complex and time-sensitive medical decisions. It provides for a flexible system that is modifiable for each unique trauma case, and is capable of communicating with EMS crews and hospitals about transportation logistics given current on-the-ground conditions and in-person assessments of the situation. Ultimately, this research demonstrates the utility of carefully handling and processing spatial data, in combination with aspatial information, to generate objective decisions in a variety of geographic and time-sensitive trauma scenarios.

References

Brown, Brandon S, Korby A Pogue, Emily Williams, Jesse Hatfield, Matthew Thomas, Annette Arthur, and Stephen H Thomas. 2012. Helicopter EMS transport outcomes literature: annotated review of articles published 2007–2011. *Emergency Medicine International* 2012.

Newgard, Craig D, Robert H Schmicker, Jerris R Hedges, John P Trickett, Daniel P Davis, Eileen M Bulger, Tom P Aufderheide, Joseph P Minei, J Steven Hata, and K Dean Gubler. 2010. Emergency medical services intervals and survival in trauma: assessment of the "golden hour" in a North American prospective cohort. *Annals of Emergency Medicine* 55 (3):235-246. e4.

Peters, Jeremy, and G Brent Hall. 1999. Assessment of ambulance response performance using a geographic information system. *Social Science & Medicine* 49 (11):1551-1566.

Raghupathi, Wullianallur, and Viju Raghupathi. 2014. Big data analytics in healthcare: promise and potential. *Health Information Science and Systems* 2 (1):3.

Widener, Michael J, Zac Ginsberg, Daniel Schleith, Douglas J Floccare, Jon Mark Hirshon, and Samuel Galvagno. 2015. Ground and helicopter emergency medical services time tradeoffs assessed with geographic information. *Aerospace Medicine and Human Performance* 86 (7):620-627.

Xu, Boyi, Li Da Xu, Hongming Cai, Cheng Xie, Jingyuan Hu, and Fenglin Bu. 2014. Ubiquitous data accessing method in IoT-based information system for emergency medical services. *Industrial Informatics, IEEE Transactions on* 10 (2):1578-1586.

Pedestrian Navigation Aids, Spatial Knowledge and Walkability

Jia Wang, Michael Worboys

Department of Mathematical Sciences, University of Greenwich,
30 Park Row, London SE10 9LS, UK
Email: {J.Wang, M.Worboys}@greenwich.ac.uk

Abstract

This study attempts to demonstrate the impact of pedestrian navigation aids on spatial knowledge acquisition and its link to walkability in an urban environment. Spatial knowledge is important for pedestrian travel. Rich spatial knowledge contributes to a good mental image of the walking environment, which consequently increases travel confidence and potentially allows more active walking. While there are plenty of studies on walkability, little work has been done on how navigation aids influence walkability. Using a pilot wayfinding experiment, we examined the effect on users' acquired spatial knowledge of two major pedestrian navigation aids used in London in comparison to direct experience of routes.

1. Introduction and Background

Walkability has become a widely discussed topic in urban and transportation planning and gained public interest since 2005. A walkable city that provides an accessible walking environment encourages more pedestrian walking. This results in benefits to the economy, improved public health and reduced ground emissions. Existing walkability studies focus on the assessment of street pattern, land use diversity and housing density (Frank *et al*. 2010), and relate to local routes and subjective pedestrian perceptions (Ewing and Handy 2006).

Spatial knowledge is important for pedestrian travel. Better spatial knowledge contributes to richer cognitive maps and thus allows improved understanding of the walking space. Acquired spatial knowledge of an environment can be differentiated depending on whether it comes from direct resources associated with travel experiences or from indirect resources such as signs and maps. Researchers conducted several experiments to compare spatial knowledge obtained from different resources (Ishikawa *et al*. 2008). The design and placement of signage systems clearly affect pedestrian orientation during their journeys (Arthur and Passini 1992), and thus have impact on spatial knowledge.

Recently various types of pedestrian navigation aids (PNA) have been developed as aids in wayfinding. These systems assist pedestrians in gaining the ability to get from one place to another, without getting lost (most of the time). Existing work focused mostly on GPS-based mobile devices (e.g., Huang *et al*. 2012). Little work has been done on how different PNAs, digital or non-digital, static or dynamic, influence spatial knowledge acquisition of pedestrians.

In this paper, we aim to demonstrate the influences of different types of pedestrian navigation aids on spatial knowledge acquisition. A navigation aid was assessed by its support of spatial knowledge acquisition of its users. We conducted a pilot wayfinding experiment to assess the two major PNAs used in London (Google Maps & Legible London) in comparison to direct experience of routes as a base line. The Legible London system is a citywide signage system for pedestrian wayfinding initiated by Transport for London in 2007. It is designed to help visitors and local residents to easily gain local spatial knowledge and so

to encourage more walking by providing time, neighborhood and transport information, 3D buildings, and heads up orientation (Transport for London 2007).

2. Method

We conducted an experiment to examine how navigation aids influence users' walking behavior and more importantly users' acquired spatial knowledge in comparison to direct experience. The two navigation aids evaluated in the experiment were GPS-based Google Maps and the Legible London signage system. The participants were split into three groups, Google Maps, Legible London and direct experience, in terms of the way to acquire spatial knowledge during wayfinding. With a similar design to Ishikawa *et al*. (2008), our experiment consisted of three tasks, namely sense-of-direction fill-out, wayfinding, and map sketching. We analyzed sketch map completeness and the accuracy of the qualitative sketch aspects (topology, orientation and order) proposed by Wang and Schwering (2015). Figure 1 gives an overview of the experiment design and workflow.

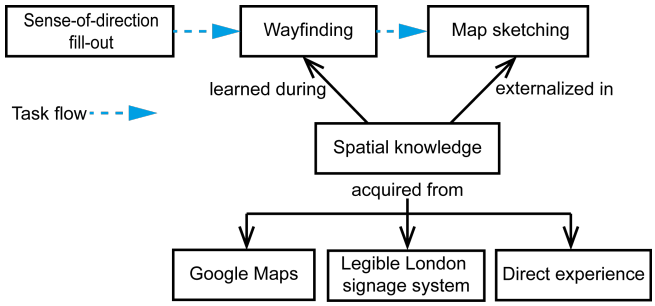


Figure 1. The experiment design and workflow.

Participants. Eight people, three men and five women, with average age of 30.3 years (*SD* = 3.58) took part in the experiment. None of them had visited the study area before the experiment. The participants were randomly assigned to one of the three groups: Google Maps (*n*=3), Legible London signage (*n*=3), and direct-experience (*n*=2).

Study area. The study area is St Christopher's Place in the West End of London. It is an open area with a mixture of shops, boutiques, restaurants and bars located just off Oxford Street. The study area is not visible when you walk along Oxford Street. So it has been described as "a hidden gem". Figure 2 (left) shows the start point (Bond street underground station) and the destination (the clothing shop called Jigsaw at St Christopher's Place) of the wayfinding task.

Figure 2. Study area (left) and a Legible London sign showing the study area (right).

Materials and Procedure. The first task (see Figure 1) required all participants to fill out the Santa Barbara Sense-of-Direction scale designed by Hegarty *et al*. (2002) as a self-report measure of environmental spatial ability. After filling out the scale, participants were taken individually to the start point and began wayfinding to the destination. Each participant was

followed by the experimenter and their wayfinding behavior was observed and recorded by using a behavior diary. The Google Maps group used the mobile GPS-based application as its navigation aid. The Legible London group was free to check any Legible London signs positioned in the study area. The three Legible London signs located near the start point are shown in Figure 2 (left). The direct experience group first walked the study area guided by the experimenter. They were then returned to the start point and asked to walk towards the destination without any navigation aid. Finally, all participants were asked to sketch the area they had travelled in a detailed manner on a piece of paper with routes highlighted.

3. Results and Discussion

Sense-of-direction fill-out task. Following the scoring procedure for the scale (Hegarty *et al.* 2002), each participant got a score where the higher the score the better the perceived sense of direction (Google Maps: M = 4.58, SD = 0.92; Legible London: M = 3.93, SD = 1.39; direction experience: M = 4.77, SD = 1.08). We did not find significant differences in the spatial ability of orientation between groups.

Wayfinding task. Figure 3 highlights the routes (in red) taken by the participants. Google Maps users all followed the same route recommended by the application (Oxford St. - James St. - Barret St.). They checked frequently their current positions on their screens and tried to align map features with their surrounding environment. Because the destination and user's current position were not always shown together on the screen, Google Maps users sometimes needed to zoom out the map to relocate themselves in relation to the destination. Legible London users made more stops, with each stop much longer than the other two groups, so they took the longest travel time. The 'heads up' style Legible London adopts makes it easy for users to understand their immediate environment. However, it took time to mentally rotate and align the signage maps with reality when the 'heads up' direction differed from the walking direction. Routes taken by Legible London users were restricted to the placement of the signs so they all look similar to the red route shown in Figure 3. Participants in the direct experience group both took the shortcut (via Gee's Ct.) without any stop so they were the fastest to reach the destination.

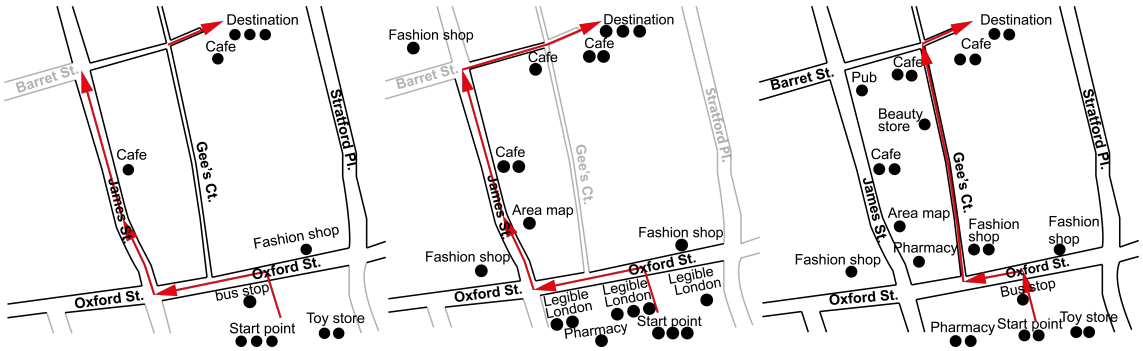


Figure 3. Routes and sketched buildings and streets.

Map sketching task. In general, the direct experience group created the most complete sketch maps in terms of the streets and buildings that were recalled and drawn; the Google Maps group drew the fewest buildings, and the Legible London group drew the most incomplete street network. In Figure 3, the number of black circles shows the number of participants who drew certain objects. Streets in grey were omitted by sketchers. The Legible London group included in their sketch maps the signs used during wayfinding. These signs became part of the study area and each sign played the role of a landmark – it provided orientation cues and memorable locations during wayfinding. The Legible London and Google Maps groups only drew the buildings on their routes while the direct experience

group also drew off-route buildings. This suggests that the former two groups acquired route knowledge and the latter group was able to learn more complex configurational survey knowledge that is not limited to any particular route. The Google Maps group performed better in learning the street network than the Legible London group. The reason could be that the Google Maps users learned the street layout by zooming out the digital map, and the signage users learned the route by recognizing in reality the relevant streets shown on signage maps and ignored irrelevant ones. As a result, the signage users could only reflect the streets constituting their travel paths on sketch maps.

We measured sketch map accuracy by using the approach proposed by Wang and Schwering (2015). The direct experience group made the most accurate sketch maps. Sketch maps from the Google Maps group were better able to represent the street network topology but worse in orientation and ordering in comparison to the Legible London group.

4. Conclusions

The results of our study suggest that the two navigation aids affect spatial knowledge in different ways. Due to the lack of exploration of the study area, the two types of PNA users had difficult in obtaining survey knowledge and learning how streets fit together so they could not take the best route to the destination. The Legible London group needed to work out a route by learning signage maps, which on one hand provided them local knowledge about their immediate environment but on the other hand prevented them from learning about their surroundings. The Google Maps group only needed to follow the automatic route guidance, which was (most of the time) quick and reliable. However, the limited screen size made it difficult to always relate the current position to the destination. Google Maps users also paid the least attention to their routes and surroundings. These two reasons caused Google Maps users to have the worst orientation during wayfinding.

The small size of this pilot study did not allow us to conduct significance and correlation tests. The continued and expanded formal experiment will include more participants from different backgrounds and more study areas of diverse spatial layouts. The formal experiment will also show how enhanced spatial knowledge could promote more active walking. Based on the formal experiment, the goal is to suggest a new aid to pedestrian navigation that makes cities more walkable by connecting users to their surroundings through enhanced spatial knowledge.

References

Arthur P and Passini R, 1992, *Wayfinding: people, signs, and architecture*. Mcgraw-Hill, Texas, USA.

Ewing R, Handy S, Brownson R, Clemente O and Winston E, 2006, Identifying and measuring urban design qualities related to walkability. *Journal of Physical Activity & Health*, 3 (suppl1): S223-S240.

Frank L, Sallis J, Saelens B, Leary L, Cain K, Conway T and Hess P, 2010, The development of a walkability index: application to the neighborhood quality of life study. *British Journal of Sports Medicine*, 44(13): 924-933.

Hegarty M, Richardson A, Montello D, Lovelace K and Subbiah I, 2002, Development of a self-report measure of environmental spatial ability. *Intelligence*, 30(5): 425-447.

Huang H, Schmidt M, and Gartner G, 2012, Spatial knowledge acquisition with mobile maps, augmented reality and voice in the context of GPS-based pedestrian navigation: Results from a field test. *Cartography and Geographic Information Science*, 39(2), 107-116.

Ishikawa T, Fujiwara H, Imai O and Okabe A, 2008, Wayfinding with a GPS-based mobile navigation system: a comparison with maps and direct experience. *Journal of Environmental Psychology*, 28(1): 74-82.

Transport for London, 2007, Legible London – A prototype wayfinding system for London. Retrieved from http://content.tfl.gov.uk/ll-yellow-book.pdf

Wang J and Schwering A, 2015, Invariant spatial information in sketch maps – a study of survey sketch maps of urban areas. *Journal of spatial information science*, 11, 31-52.

Characterizing place: an empirical comparison between user-generated content and free-listing data

F. M. Wartmann[1], C. Derungs[1,2], R.S. Purves[1,2]

[1]Department of Geography, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
Email: {flurina.wartmann; curdin.derungs; ross.purves}@geo.uzh.ch

[2]URPP Language and Space, University of Zurich, Freiestrasse 6, CH-8032 Zurich

Abstract

Methods to gather information from the public about place range from ethnographic approaches such as free listing to automatic extraction from user-generated content. We compared aspects of place (*location*, *locale* and *sense of place*) contained in free lists elicited from participants recruited on site with tags from georeferenced Flickr images. Using manual annotation we assigned content as toponyms (*location*), landscape elements (*locale*) and feelings (*sense of place*). Flickr tags contained more toponyms than free-listing data, but almost no information relating to feelings. Landscape elements were prominent in both data sets, with those captured by free lists and Flickr being cognitively more salient than those only captured by free listing, suggesting they represent basic levels. In Flickr, landscape elements consisted of basic levels in different languages (e.g. *mountain, Berg*), while free lists contained landscape elements both at the basic and subordinate level (e.g. *Arvenwälder*, arolla pine forests). We conclude that both methods yielded information about *locale*, with Flickr contributing basic levels and free lists also more detailed information, but that Flickr provided little information about *sense of place* compared to *in situ* free-listing elicitation with participants.

1. Introduction

The importance of taking into account meanings people assign to places in management and planning is increasingly being recognized (Jones 2007; Prieur *et al.* 2006). However, collecting such information from the public is non-trivial, and different approaches are used, ranging from ethnographic work to extraction from user-generated content. Given the diversity of approaches, it is crucial to compare different approaches and understand what types of information are captured by different methods such that research on place can effectively capture relevant information. In this paper, we compare free-listing elicitation with automated extraction of place descriptions from user-generated content. In free-listing experiments, participants are asked to freely name examples of, for instance, categories they associate with the landscape they currently find themselves in (Bieling *et al.* 2014; Wartmann *et al.* 2015). Another method increasingly used is to automatically extract information from user-generated content, often in the form of image tags (e.g. Jenkins *et al.* 2016; Purves *et al.* 2011; Rattenbury and Naaman 2009). We compared these two methods based on the information about place they contained.

2. Methods

We chose two sites in the Swiss Alps (Figure 1) where outdoor free listings on landscape categories had been conducted with 59 participants recruited on site (Wartmann *et al.* 2015). Terms were elicited with a Swiss German question that literally translates to: *'what is there for you in a*

landscape?', resulting in 332 unique terms. From a set of ~1 million georeferenced images from the Flickr platform we automatically extracted image tags using a 10km radius around our study sites. After filtering for bulk-uploads and removing tags only applied by a single user, we retained a total of 633 tags, each of which was applied as a tag by between 2 to 241 users.

Flims

Flickr tags	Free-listing data
Laax	*Berge (mountains)*
Flims	*Bäume (trees)*
Graubünden	*Felsen (rocks)*

Müstair

Flickr tags	Free-listing data
Graubünden	*Berge (mountains)*
Ofenpass	*Tal (valley)*
Müstair	*Wald (forest)*

Figure 1. Study sites and example data (Flickr tags with highest TF-IDF values and free-listing data with highest cognitive saliency index)

We compared the two data sets using three aspects of place as defined by Agnew (1987): *location* (a specific location identified through coordinates, or a name), *locale* (the actual elements that characterize a place, such as *streets*, *buildings*, *rivers*, or *forests*) and *sense of place* (the meanings and feelings people attach to a place). As a language-independent means of comparing the data sets, we manually annotated content types. We devised written annotation guidelines, using the content type 'toponym' for *location*, (e.g. *Flims, Buffalora*), 'landscape element' for *locale*, which included both natural (e.g. *lake, forest, mountain)* and anthropogenic elements (e.g. *road, restaurant*), 'feeling' for *sense of place* (e.g. *peace, quietness*), and 'other' for other content that mostly related to photography, qualities, and person or event names (e.g. *350d, blue, burtonopen*). Three researchers assigned content types using the guidelines and discussed unclear cases until all content was annotated. We calculated the absolute number and percentage of tags/free-listing elements per content type. For the type 'landscape element' we assessed correspondences between Flickr and free lists by considering terms that were identical, translations of each other (e.g. *river-Fluss*), or singular/plural forms (e.g. *river-Flüsse*). For Flickr tags we compared a spatial term frequency–inverse document frequency (TF-IDF, c.f. Rattenbury and Naaman 2009) and for free-listing data cognitive saliency values (Sutrop 2001) between corresponding and non-corresponding terms using Mann-Whitney-U tests to assess statistical significance at a significance level of 0.05.

3. Results and Interpretation

In the following, we briefly present our results with respect to the classification introduced above for Flickr tags and free-listing elements (Table 1). Toponyms are more prevalent in Flickr than our free listings. By contrast, landscape elements are in absolute terms numerous in both data sets, though relatively less prominent in Flickr. Almost no tags relate to feelings in Flickr tags extracted for the study sites, while our free-listing data contain 21 such terms, including *Geborgenheit* (emotional security), *mit sich eins sein* (to be at one with oneself), and *Erdverbundenheit* (connection to the Earth). Both data sets contain relatively high proportions of landscape elements.

Content classified as 'other' dominates Flickr, and includes a wide range of tags which related to, for example, qualities (e.g. *snow, cloudy*), activities (e.g. *hiking, skiing*), and camera metadata.

Table 1. Content types for Flickr tags and free-listing data for the study sites

	Flickr		Free lists	
Flims	(absolute)	(%)	(absolute)	(%)
toponyms	99	19.64%	7	2.41%
landscape elements	60	11.91%	134	46.05%
feelings	1	0.20%	11	3.78%
other	151	68.25%	2	47.77%
Val Müstair				
toponyms	67	31.46%	12	7.55%
landscape elements	28	13.15%	67	42.14%
feelings	0	0.00%	10	6.29%
other	57	55.40%	5	44.03%

As landscape elements are common in both datasets, we explore the degree of overlap between this content, and compare the nature of overlapping and non-overlapping tags/free-listing data. Figure 2 illustrates the properties of Flickr tags and free-listing elements for Val Müstair.

Figure 2. Corresponding and non-corresponding landscape elements in tags and free lists

A number of observations can be made. Firstly, Flickr is dominated by English, while our free-listing (since it was carried out in variant of German) only contains German terms. Secondly, only 8 Flickr tags are not replicated in the free-listing data, while 54 landscape elements were unique to free-listing data. Thirdly, we found correspondences between 19 Flickr tags with 13 free-listing landscape elements. The reason for the mismatch in these counts is illustrated in Figure 2, with for example matches between synonyms (e.g. *creek/stream* in Flickr with *bach* in free lists), as well as between singular/plural forms and different languages (e.g. *mountain/mountains/montagna/ montagne* with *berg/berge*). Since we hypothesized that landscape elements which were more salient in our lists were more likely to have correspondences with Flickr tags, we tested for significant differences in TF-IDF and cognitive salience values between corresponding and not corresponding landscape elements. For Flickr tags (TF-IDF) we found no significant relationship,

while for free-listing landscape elements cognitive salience values were significantly higher for corresponding landscape element for both Flims and Val Müstair.

4. Discussion

Our results confirm previous findings that Flickr tags are a good source for toponyms, that is, in Agnew's terms, for information about *location* (Sigurbjörnsson and van Zwol 2008), but we found very few tags relating to feelings or *sense of place*. Although the free-listing experiment was not aimed at specifically eliciting *sense of place*, participants provided such information. Both data sets contained landscape terms, but due to their specific elicitation, free lists contained more terms despite few participants. Landscape terms in free lists shared with Flickr were significantly more cognitively salient than non-shared terms, suggesting shared terms may be candidates for highly salient basic-level landscape categories (Tversky and Hemenway 1983) such as *river*, or *mountain*, while elements only included in free-listing data often appear to be typical subordinate categories (e.g. *rough pasture* or *mountain torrent*). The shared Flickr terms did not have significantly higher TFIDF-values than non-shared terms, indicating that, as potential basic-level categories, these terms are represented throughout the Flickr corpus (Rorissa 2008), and not particularly overrepresented in the study areas. Flickr was a reliable source of basic-level categories for *locale*, but contained little additional information about landscape elements not present in free-listing data. For future work we thus suggest triangulating methods, where *location* and basic-level categories for *locale* are extracted from user-generated content, with ethnographic methods providing more detailed information about *locale* and, using particular elicitation questions, about *sense of place*.

Acknowledgements

This work was supported by the cogito foundation (grant no. 15-129-R) and the Swiss National Science Foundation (SNF) through the project PlaceGen (grant no. 149823).

References

Agnew JA, 1987, *Place and politics. The geographical mediation of state and society*. Allen & Unwin, Boston, USA.

Bieling C, Plieninger T, Pirker H and Vogl CR, 2014, Linkages between landscapes and human well-being: An empirical exploration with short interviews. *Ecological Economics*, 105:19–30.

Jenkins A, Croitoru A, Crooks AT, Stefanidis A, 2016, Crowdsourcing a collective sense of place. *PLoS ONE*, 11(4).

Jones M, 2007. The European landscape convention and the question of public participation. *Landscape Research*, 32(5):613–633.

Prieur M, *et al.*, 2006, *Landscape and sustainable development-challenges of the European Landscape Convention.* Council of Europe Publishing, Strasbourg, France.

Purves RS, Edwardes A and Wood J, 2011, Describing place through user generated content. *First Monday*, 16(9).

Rattenbury T and Naaman M, 2009, Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1):1–30.

Rorissa, A, 2008, User-generated descriptions of individual images versus labels of groups of images: A comparison using basic level theory. *Information Processing & Management*, 44(5): 1741–1753.

Sigurbjörnsson B and van Zwol R, 2008, Flickr tag recommendation based on collective knowledge. In *Proceedings of WWW'08*. ACM Press, New York, USA, 327–336.

Sutrop U, 2001, List Task and a Cognitive Salience Index. *Field Methods*, 13(3):263–276.

Tversky B and Hemenway K, 1983, Categories of environmental scenes. *Cognitive psychology*, 15(1):121–149.

Wartmann FM, Egorova E, Derungs C, Purves RS, and Mark DM, 2015, More than a list: what outdoor free listings of landscape categories reveal about commonsense geographic concepts and memory search strategies. In Fabrikant SI, Raubal M, Bertolotto M, Davies C, Freundschuh S, and Bell S (eds.), *Spatial Information Theory. Lecture Notes in Computer Science,* Vol. 9368. Springer International, Cham, Switzerland, 224–243.

Outlier Detection in OpenStreetMap Data using the Random Forest Algorithm

Richard Wen, Claus Rinner

Department of Geography and Environmental Studies, Ryerson University
350 Victoria St., Toronto, Ontario, M5B 2K3, Canada
Email: {rwen, crinner}@ryerson.ca

Abstract

OpenStreetMap (OSM) data consist of digitized geographic objects with semantic tags assigned by the volunteer contributors. The tags describe the geographic objects in a way that is understandable by both humans and computers. The variability in contributor behaviour creates reliability concerns for the tagging quality of OSM data. The detection of irregular contributions may improve OSM data quality and editing tools. This research applies the random forest algorithm on geospatial variables in order to detect outliers without ground-truth reference data to direct human inspection. An application to OSM data for Toronto, Ontario, was effective in revealing abnormal amenity tagging of school and hospital objects.

1. Introduction

OpenStreetMap (OSM) is an online platform enabling registered volunteers to contribute geospatial data by digitizing point-, line-, or polygon-shaped geographic objects and annotating them with tags referring to common feature classes such as roads and restaurants (Haklay 2008). OSM tags are semantically structured as key-value pairs, where the key refers to a broad class of geographic objects and the value details the specific geographic object being tagged (Ballatore *et al*. 2013). Examples of tags are *amenity=school*, *highway=residential*, and *building=house*.

The open and flexible nature of OSM tagging leads to varying contribution behaviour by different communities (Mooney *et al*. 2010). The varying contribution behaviour creates concerns about the quality of OSM data and the community standards of OSM tagging. Quality control and corrections rely heavily on human interaction, which raises additional questions on the reliability of OSM data. Finally, the experience of the volunteer contributor has an effect on the tagging quality of each geographic object as experienced contributors are more familiar with the tagging norms of the area being edited. Although OSM is an effective and efficient platform for generating masses of geospatial data, it is plagued by reliability, quality, and completeness issues.

The aim of this paper is to examine the ability of an automated machine learning algorithm, the random forest algorithm, to support manual human inspection and minimize bias in OSM data editing. The use of an automated algorithm improves the detection of abnormal tagging behaviour, avoids the bias of human judgement, and reduces the time required to search through masses of tagged geographic objects. A combination of human knowledge and experience with the logical accuracy of machines could improve OSM tagging quality and standards, and enable the development of advanced editing tools.

2. Data and Methods

OSM data for the City of Toronto, Ontario, were downloaded from Mapzen Metro Extracts in the form of a GEOJSON file (Mapzen 2016). A GEOJSON file contains one or more spatial objects described by geometry types and properties in a key/value data structure (Butler *et al*. 2008). The OSM key category datasets amenities, places, transport areas, aero ways, transport points, and roads were selected to be used from the downloaded data. The selected data consisted of 70,535 geographic objects in the City of Toronto detailed in Table 1. The majority of geographic objects resided in the transport points and roads datasets. The data were projected from a geographic coordinate system (WGS 1984) into a planar coordinate system (NAD 1983 UTM Zone 17 North) for geometric calculations. A tag value is referred to as a tag in this paper.

Table 1. OpenStreetMap Data for City of Toronto, Ontario from Mapzen.

Key Category Dataset	Tag Values Available	Geometry Type	Count
Amenities	fire_station, fuel, hospital, library, police, school, townhall, university	Point	1507
Places	city, county, hamlet, locality, neighbourhood, suburb, town, village	Point	760
Transport Areas	aerodrome, apron, helipad, platform, station, terminal	Polygon	72
Aero Ways	runway, taxiway	Line	438
Transport Points	aerodrome, bus_stop, crossing, gate, halt, helipad, level_crossing, motorway_junction, station, subway_entrance, terminal, tram_stop, turning_circle	Point	21,309
Roads	disused, monorail, motorway, motorway_link, preserved, primary, primary_link, rail, secondary, secondary_link, subway, tertiary, tertiary_link, tram, trunk, trunk_link	Line	46,812

The methods first required the extraction of geospatially meaningful variables to describe the geometric characteristics and spatial relationships of each geographic object. The first set of geospatial variables were extracted by utilizing the geometric structures in the data. Area, length, and the number of vertices for each geographic object in the data were extracted as columns. These geospatial variables describe the geographic objects in terms of geometric characteristics such as size (area) and geometric complexity (vertices). Representative coordinates were also extracted in the form of the x- and y-coordinates of the centre of each object. These representative coordinates allowed the random forest model to utilize spatial patterns if they existed. The second set of geospatial variables were the distances to the nearest uniquely tagged geographic object for each individual geographic object. A column was created for each of the available tag values. These columns were referred to as the Distance to the Nearest Neighbour Tag (DNNT). An example of the DNNT concept is seen in Figure 1.

Redundant variables were removed by examining each variable for a high correlation (less than -0.7 and greater than 0.7) to another variable and removing the other highly correlated variables in a specified order. The order arranged the area, length and vertices first, followed by sorting the DNNT variables by their tag frequency. The result of the extracted geospatial variables after removing redundant variables is referred to as the input data in this paper.

Figure 1. Distance to the Nearest Amenity Tag for a Police Station Object.

Several random forest models were run on the input data to classify the tag value of geographic objects. A random forest consists of a number of decision trees built on subsamples of approximately two-thirds of the input data (Breiman 2001). The other one-third of the subsamples are used to calculate an out-of-bag error estimate by aggregating the predictive scores (Liaw and Wiener 2002). Each random forest model used balanced tag weights, penalizing misclassification of minority tags, to adjust for tag frequency imbalances in the data (Chen *et al*. 2004). A number of maximum split variables equal to the square root of the number of variables in the input data were used for each decision tree in the random forest models. Three random forests models were constructed to optimize the number of decision trees using 64, 96, and 128 decision trees as suggested by Oshiro *et al*. (2012) to determine the model with the lowest out-of-bag error estimate. The selected model with the lowest out-of-bag error estimate is referred to as the Tree Optimized Random Forest (TORF) model in this paper.

The TORF model was used to determine outliers in the input data by calculating proximity measures between two geographic objects (Louppe 2014) to produce proximity matrices for geographic objects inside each tag followed by calculating outlier measures (eq. 1) according to Breiman and Cutler (2004) for each geographic object.

$$outlier(n_c) = \frac{N}{\sum_{k_c}^{K}[proximity(n_c, k_c)]^2} \quad \text{(eq. 1)}$$

where n_c is a sample instance of tag c, k_c is all other sample instances of tag c, K is the total number of k_c, and N is the total number of n samples. The outlier measures were then normalized by subtracting every outlier value for n instances of each tag c by the median of all outlier measures inside the same tag c, and dividing by the absolute deviation from the median.

A geographic object was suspected of being an outlier if its normalized outlier measure was greater than 10.

The interpretation of outliers was enhanced by using local variable contribution increments (eq. 2) to calculate variable contributions (eq. 3) according to Palczewska *et al* (2014) for the outlier tags. The variable contributions were ranked to find the most influential variables.

$$LI_f^c = \begin{cases} Y_{mean}^c - Y_{mean}^p & if\ split\ of\ p\ is\ for\ f \\ 0 & otherwise \end{cases} \quad \text{(eq. 2)}$$

where LI is the local variable contribution increment, f is a variable, c is the child node, p is the parent node, Y_{mean}^c is the fraction of training samples in a child node, and Y_{mean}^p is the fraction of training samples in a parent node.

$$FC_i^f = \frac{1}{T}\sum_{t=1}^{T} FC_{i,t}^f \quad \text{(eq. 3)}$$

where FC_i^f is the variable contribution of a training sample, $FC_{i,t}^f$ is the sum of local variable specific contribution increments, f is a variable, T is the total number of trees in the forest, t is a tree in the forest, and i is a training sample.

3. Results

The TORF model was obtained from 128 trees, which provided the lowest out-of-bag error of 0.162 compared to 0.166 and 0.167 for 96 and 64 trees respectively. The schools in Figure 2 and the hospitals in Figure 3 had normalized outlier measures above 10. Closer inspection of the schools revealed that the schools are historical and were far away from bus stops. The hospitals were individual wings of Sunnybrook hospital, which were further away from secondary roads than normal.

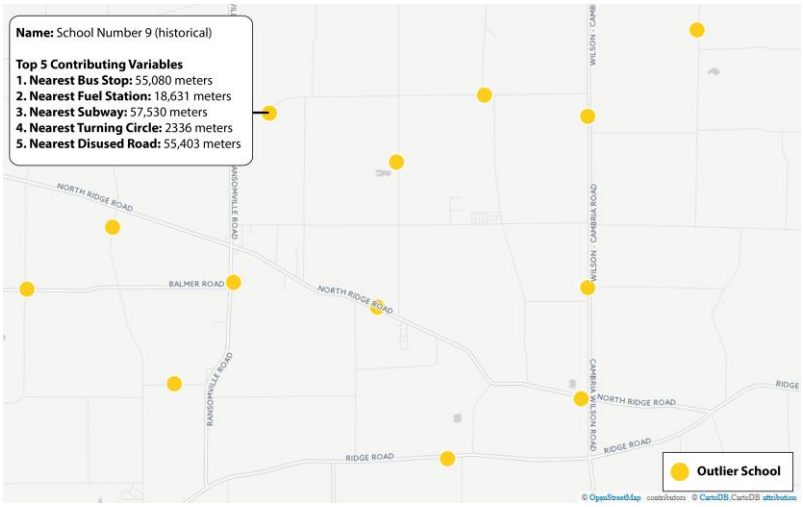


Figure 2. Detected Tag Outliers of Value School

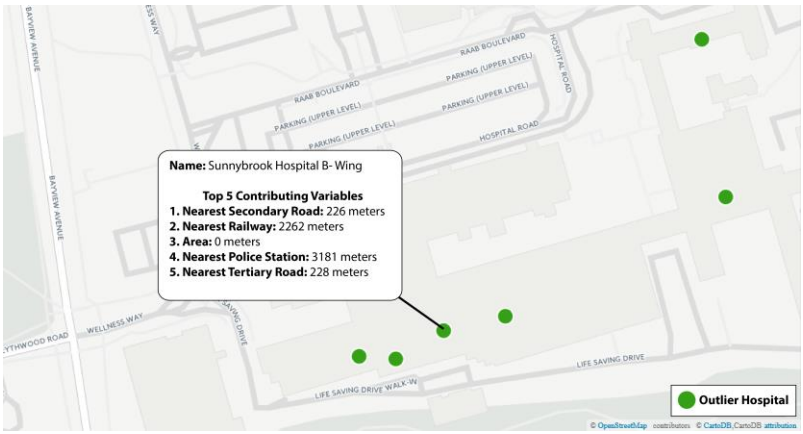


Figure 3. Detected Tag Outliers of Value Hospital

4. Conclusion

The use of random forests for outlier detection has the potential to support manual data cleaning efforts, where the discovery of potential outliers that may yield insight into the community tagging standards and errors in a study area. The outlier interpretation was also enhanced by the variable contributions, which may provide reasons for abnormal tags. However, satellite imagery and user editing history were not used, although they may have a significant impact on outlier detection. Only nearest neighbour objects were used, other spatial relations such as distance buffers should be tested in the future. Adding raster data, temporal data, and a variety of spatial relations to the random forest model could further improve the outlier detection of OSM tags.

Acknowledgements

To be added after review.

References

Ballatore A, Bertolotto M and Wilson DC, 2013, Geographic knowledge extraction and semantic similarity in OpenStreetMap, *Knowledge and Information Systems*, 37(1):61-81.

Breiman L, 2001, Random Forests, *Machine Learning*, 45(1):5-32.

Breiman L, Cutler A, 2004, Random Forests, Retrieved from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm#outliers

Butler H, Daly M, Doyle A, Gillies S, Schaub T and Schmidt C, 2008, The GeoJSON Format Specification, Retrieved from: https://datatracker.ietf.org/doc/draft-ietf-geojson/

Chen C, Liaw A, and Breiman L, 2004, Using random forest to learn imbalanced data, University of California, Berkeley.

Haklay MM, 2008, OpenStreeMap: User-Generated Street Maps, *Pervasive Computing*, 12-18.

Liaw A and Wiener M, 2002, Classification and Regression by randomForest, *R News*, 18-22.

Louppe G, 2014, Understanding Random Forests: From Theory to Practice, University of Liege, Belgium.

Mapzen, 2016, Metro Extracts, Retrieved from: https://mapzen.com/data/metro-extracts/

Mooney P, Corcoran P and Winstantly AC, 2010, Towards Quality Metrics for OpenStreetMap, *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, New York, USA, 514-517.

Oshiro, TM, Perez PS and Augusto J, 2012, How many trees in a random forest?, *Machine Learning and Data Mining in Pattern Recognition*, 7376: 154-168.

Palczewska A, Palczewski J, Robinson RM and Neagu D, 2014, Interpreting random forest classification models using a feature contribution method. *Integration of Reusable Systems*, 26:193-218.

Constructing a Routable Transit Network from a Real-time Vehicle Location Feed

Nate Wessel[1], Jeff Allen[1], Steven Farber[2]

[1]Department of Geography and Planning, University of Toronto, Sidney Smith Hall, 100 St. George Street, Toronto, ON M5S 3G3, Canada
Email: {nate.wessel; jeff.allen}@mail.utoronto.ca

[2]Department of Human Geography, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON M1C 1A4, Canada
Email: steven.farber@utoronto.ca

Abstract

Many transit agencies publish open access feeds of real-time fleet locations for use by application developers to deliver mobile navigation services. However, by collecting a transit data feed over time, we demonstrate how to recreate retrospective, routable transit networks that are useful in answering network performance and accessibility research questions. In this research we develop an end-to-end GIS toolchain for 1) downloading and storing a transit data feed, 2) coercing the collected spatiotemporal database into a retrospective transit schedule adhering to the General Transit Feed Specification (GTFS) data standard, and 3) creating a routable transit network with time-dependent travel times in OpenTripPlanner. We further demonstrate how this toolchain can be used to identify discrepancies between scheduled and actual travel times on the network and motivate the usefulness of this approach through an accessibility analysis.

1. Introduction

The emergence of the General Transit Feed Specification (GTFS) data standard, and the publication of GTFS packages by thousands of agencies worldwide has unleashed a flurry of tools development for researching transit networks (Hadas and Ranjitkar 2012), transit travel times (Farber et al. 2014), and time-dependent accessibility metrics (Fransen et al. 2015; Owen and Levinson 2015; Farber et al. 2016). One problem with travel time research based on GTFS schedules is that this format implicitly ignores inaccuracies in travel times due to, among other causes, operational delays, service interruptions or unrealistic schedules. Researchers requiring more accurate measures of transit travel times have relied on field measurements and large-scale simulation models of multimodal network assignments. But these sources of data are expensive to collect, may take years to implement, may not be very accurate, and are difficult to implement on a large network, in continuous time, and in perpetuity.

To account for these shortcomings, we put forward a new methodology that capitalizes on freely available vehicle location data feeds to produce a routable retrospective transit network using open source tools. This network contains actual travel times that are true to observed transit network performance, not the expected performance contained in the official transit schedules. In this paper we will describe the toolchain we developed and demonstrate its use in detecting service that differs from the schedule, as well as the implications of these differences for end-to-end travel times and time-dependent accessibility scores.

2. Data and Methods

This research uses two primary datasets, a current GTFS package for the Toronto Transit Commission (TTC) and a 48-hour extract of all vehicle locations reported by the TTC-NextBus

API. The GTFS package contains the scheduled stop times of every transit vehicle at every stop on every route in the TTC network, structured so as to allow routing between stops across the network. Our general approach is to produce a modified GTFS package with updated trips and stop times based on observations of the actual vehicle locations over the course of a day. The modified GTFS package can then be analysed in the same exact way as the scheduled data. The methodology is summarized in Figure 1.

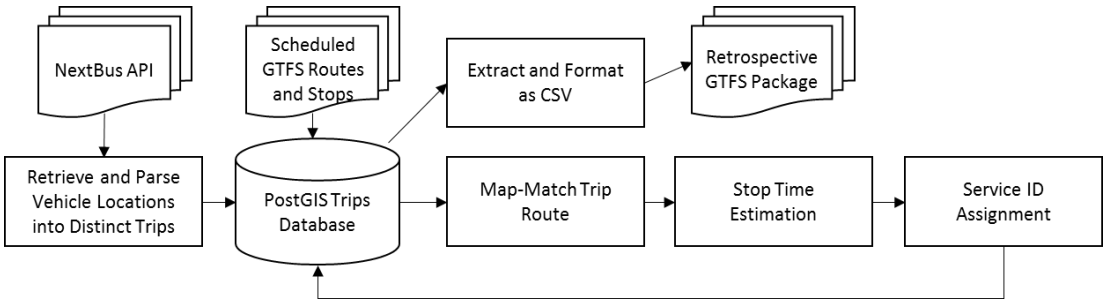


Figure 1. Creation of a Retrospective GTFS Package.

1. Retrieve and parse location data. The NextBus API is a publicly available web service designed to serve real-time transit data primarily to mobile phone or web applications. One of its functions is to report the latest locations for all operating vehicles in the fleet. For each vehicle, the API returns information on the vehicle's ID, route, heading, last known location, and the time it reported that location to the server. Vehicles update their location about every 20 seconds.

Every 10 seconds, a Python script requests all newly updated locations from the API, parses the response, and stores the results in a PostGIS database. Each vehicle is taken to be operating a trip which terminates when the vehicle changes its reported route or heading, or when it fails to update its location for more than 60 seconds. Each trip is thus essentially a GPS trace, with a set of ordered points through time and space. To obtain a better spatial resolution, the points of each trip are map-matched to a road/rail network using the Open Source Routing Machine (OSRM) software and detailed network data from OpenStreetMap. Points or trips that could not be plausibly matched to the street network (~2% of trips) were discarded. Many of these trips were also ambiguous or erroneous upon visual inspection, with much of this due to clear GPS error.

2. Estimate Stop Times. Given the heading and route ID, we find the set of stops being serviced from the NextBus API's routeConfig command. From these, stops within 20 meters of the estimated route of the vehicle are matched to a point on that route and the time for each stop is interpolated linearly from the times associated with the vehicle position reports.

3. Create New GTFS Package. With distinct trips, and stop times in sequence for each trip, we have essentially the same data structure as in the GTFS format. Each observed trip becomes a trip in the retrospective GTFS package, and each interpolated time becomes a stop for that trip. A distinct service ID is generated for each day of observed data. All stops with observed times are included, with locations given by the NextBus API. Since real-time data were not available for the TTC's four subway lines, we appended the scheduled GTFS data for these routes in our realtime GTFS package.

All tools we used were open source, and the code we developed is available on GitHub.

3. Accessibility Case Study

We demonstrate the utility of the retrospective transit network by comparing a simple measure of jobs accessibility using the scheduled and observed transit networks. The cumulative accessibility to jobs score for a location i is shown in (1).

$$A_i = |\boldsymbol{M}|^{-1}\left[\sum_{m\in\boldsymbol{M}}\sum_{j=1}^{J} O_j f(i,j,m,T)\right] \quad (1)$$

The inner summation in (1) is an accessibility score for location i at time m. A_i is therefore the mean accessibility score for location i within the time window $\boldsymbol{M}$. O_j is a count of jobs at location j and $f(i,j,m,T)$ is an indicator function that equals 1 if the transit travel time from i to j at departure time m is less than some threshold, T. In our specific case, we compute accessibility from census Dissemination Areas (DAs) to the estimated number of jobs at Traffic Analysis Zones (TAZ). We measure average accessibility within the morning rush hour (7:00am-9:00am) using a threshold of 45 minutes, a figure close to the mean one-way transit commute time in Toronto. All travel time calculations were performed using OpenTripPlanner which computed centroid to centroid travel times inclusive of walking, waiting, in-vehicle, and transferring times. We calculated accessibility scores based on the official GTFS package released by the TTC, and compared them to scores based on the observed travel times encapsulated in the retrospective GTFS package.

4. Results

Our primary result is displayed in a map of accessibility differences in Figure 2. Negative values denote locations where accessibility scores using the real-time network are lower than those using the scheduled network, and indicate where observed service (i.e. vehicle speeds and headways) culminated in lower than expected levels of accessibility to jobs. Positive values denote the opposite; these are areas where observed accessibility levels are higher than those obtained by using the scheduled network. This could be due to conservative schedules, expecting delays that weren't actually encountered (Wessel 2015). As expected, the two accessibility scores obtain similar values nearer to the city centre and along the subway lines. This is likely due to a higher dependence of accessibility on walking and subway routes from these locations, and therefore lower levels of sensitivity to street-level transit operations. The reader is reminded that real-time subway data were not available so the real-time network actually contains the scheduled travel times for this mode.

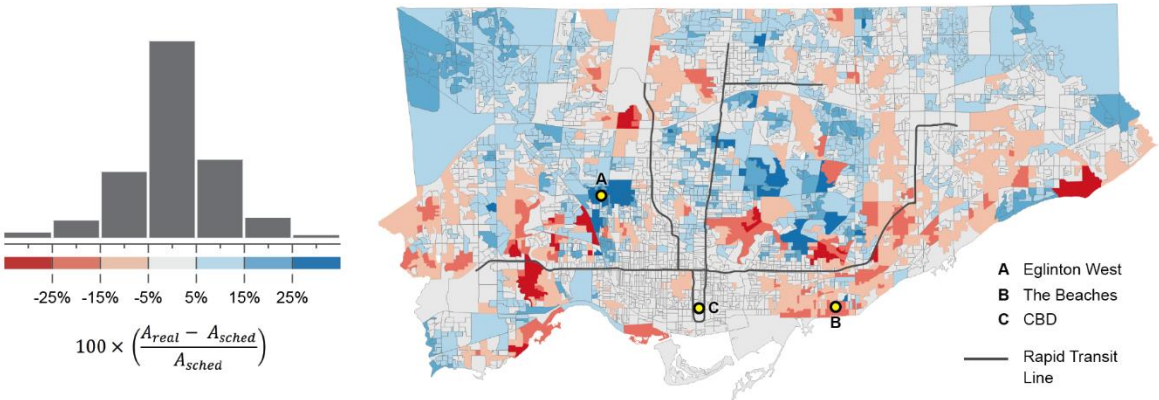


Figure 2. Differences Between Scheduled and Real-Time Access Scores.

Perhaps surprisingly, there are about as many areas exhibiting relative declines in accessibility as there are increases. These differences in accessibility are clustered in certain neighbourhoods and along specific routes where transit operations differ substantially from the published schedule. This is further visualized in Figure 3 with two comparative minute-by-minute travel time plots, each from a residential neighbourhood (The Beaches and Eglinton West) to Toronto's Central Business District (CBD). The left plot is an example of where transit operates with greater headways and results in longer commutes compared to the published schedule. Presumably there were less transit vehicles in operation than scheduled during this period or there was severe vehicle bunching. Conversely, the plot on the right is an example of where travel times from the real-time network are on average less than those from the scheduled network. This could be because the TTC overestimated operational delays when scheduling their service.

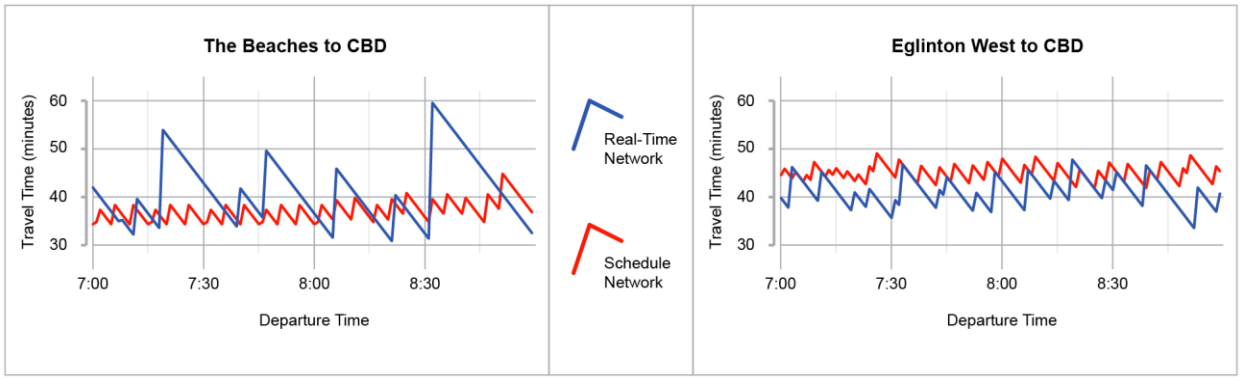


Figure 3. Scheduled and Real-Time Travel Times Resulting in Accessibility Differences.

5. Conclusions

This paper presented a new methodology for constructing a routable retrospective transit network based on freely available data and open-source software. The utility of the tool was demonstrated in a case study of accessibility to jobs, a measure of transit benefit with a wide range of applications. To our knowledge, this is the first time that accessibility on observed transit service has ever been computed and compared to scheduled accessibility metrics. Such results, especially if averaged over longer time periods, could be used to provide more realistic accessibility measures than what is available from GTFS schedule data alone. Future work will extend the analysis to include real-time subway travel times and to explore the causes for, and implications of, any systematic differences between scheduled and observed travel times and accessibility levels.

References

Farber S, Morang MZ and Widener MJ, 2014, Temporal variability in transit-based accessibility to supermarkets. *Applied Geography,* 53:149–159.

Farber S, Ritter B and Fu L, 2016, Space–time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: A social equity perspective. *Travel Behaviour and Society,* 4:40–48.

Fransen K, Neutens T, Farber S, De Maeyer P, Deruyter G and Witlox F, 2015, Identifying public transport gaps using time-dependent accessibility levels. *Journal of Transport Geography,* 48:176–187.

Hadas Y and Ranjitkar P, 2012, Modeling public-transit connectivity with spatial quality-of-transfer measurements. *Journal of Transport Geography,* 22:137–147.

Owen A and Levinson DM, 2015, Modeling the commute mode share of transit using continuous accessibility to jobs. *Transportation Research Part A: Policy and Practice,* 74:110–122.

Wessel N, 2016, *Discovering the Space-Time Dimensions of Schedule Padding and Delay from GTFS and Real-Time Transit Data*. Masters Thesis, University of Cincinnati.

A Multistart Heuristic Approach to Spatial Aggregation Problems

Ningchuan Xiao[1], Peixuan Jiang[2], Myung Jin Kim[3], Anuj Gadhave [4]

[1] Department of Geography, Ohio State University, Columbus, OH 43210
Email: xiao.37@osu.edu
[2] Esri, Redlands, CA, 92373
[3] Gyeonggi Institute of Science and Technology Promotion, Suwon, Gyeonggi, Korea
[4] Department of Computer Science and Engineering, Ohio State University, Columbus, OH 43210

Abstract

In this paper, we present a heuristic method that can be used to search for a diverse set of solutions to spatial aggregation problems. The algorithm is developed using a multistart strategy. Computational experiments are conducted to test the effectiveness of the algorithm.

Spatial data are often aggregated for different purposes. The census data in the United States, for example, are aggregated into levels such as blocks, block groups, and tracts. In another example, political redistricting requires the aggregation of spatial units into districts such that some objectives can be optimized. Though aggregation is a common exercise in the use of spatial data, it has been noted that many of the aggregations are arbitrary and may not provide effective spatial units for applications (Martin, 1998; Cockings and Martin, 2005), which often leads to the modifiable areal unit problem (Openshaw, 1983). Equally important is the multiplicity of spatial aggregations: given an aggregation of spatial units, many equivalent schemes may also exist. For example, there often exist many perfect political redistricting plans when population equality is the only objective (Kim and Xiao, 2016). Subsequently, it is important to explore a diverse set of aggregation schemes in order to fully understand the complexity of aggregation. Researchers have developed a wide range of methods that can be used to solve aggregation problems (Openshaw and Rao, 1995; Xiao, 2008). However these methods are designed for specific purposes and they generally do not aim to explore the complexity of aggregation.

The purpose of this paper is to develop a heuristic method that can be used to find, not one, but a diverse set of high quality solutions to an aggregation problem. This method first uses a search algorithm to find a set of good solutions. These solutions are stored in a pool and another algorithm is developed to improve the solutions in the pool by recombining them into new solutions. This method is heuristic, meaning it cannot guarantee optimal solutions be found, and we test its effectiveness using a set of benchmark problems.

We have tested the multistart algorithm with a wide range of data. Due to the page limit of this abstract, we discuss the first data set of Iowa congressional redistricting for year 2000. The Iowa Constitution dictates that the counties shall not be split for political redistricting purposes, which make it a problem of aggregating 99 spatial units into 5 regions. The official 2000 plan has an objective function value of 0.0080, and the literature has documented a number of redistricting

plans that can be considered as the benchmark for other algorithms: 0.0045 (Xiao, 2008), 0.0066 (Kim, 2011), 0.0011 (Guo and Jin, 2011), and 0.00079 (Kim, 2011).

Figure 1 includes the solutions found in a run with a pool size of 50, where the upper-left map shows the best solution found with an objective function value of 0.00017, which is better than the best solution in the literature, and the lower-right map the worst in the pool (with an objective function value of 0.0037, which is still an improvement over the official plan).

The multistart heuristic method presented in this paper is shown to be effective in generating aggregated regions with an objective of equal total weight (population) among the regions. We stress that the method is designed specifically to explore the multiplicity of spatial aggregation and the method can be used to generate a diverse set of aggregations. Though we only considered spatial contiguity and weight equality in the current form of this method, we believe it is possible to incorporate other constraints and objectives in this framework. Data sets of practical sizes (in hundreds of spatial units) were used to test this method. One of the future directions is to utilize high performance computing techniques to further speed up so that large data sets with thousands of spatial units can be practically processed.

References

Cockings, S. and D. Martin (2005). Zone design for environment and health studies using pre-aggregated data. *Social science & medicine 60*(12), 2729–2742.

Guo, D. and H. Jin (2011). iredistrict: Geovisual analytics for redistricting optimization. *Journal of Visual Languages & Computing 22*(4), 279–289.

Kim, M. and N. Xiao (2016). Contiguity-based optimization models for political redistricting problems. *International Journal of Applied Geospatial Research*, accepted.

Kim, M. J. (2011). *Optimization Approaches to Political Redistricting Problems*. Ph. D. thesis, The Ohio State University, Columbus, OH.

Martin, D. (1998). Optimizing census geography: the separation of collection and output geographies. *International Journal of Geographical Information Systems 12*(7), 673–685.

Openshaw, S. (1983). *The Modifiable Areal Unit Problem*. Norwich: Geo Books.

Openshaw, S. and L. Rao (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning A 27*, 425–446.

Xiao, N. (2008). A unified conceptual framework for geographical optimization using evolutionary algorithms. *Annals of the Association of American Geographers 98*(4), 795–817.

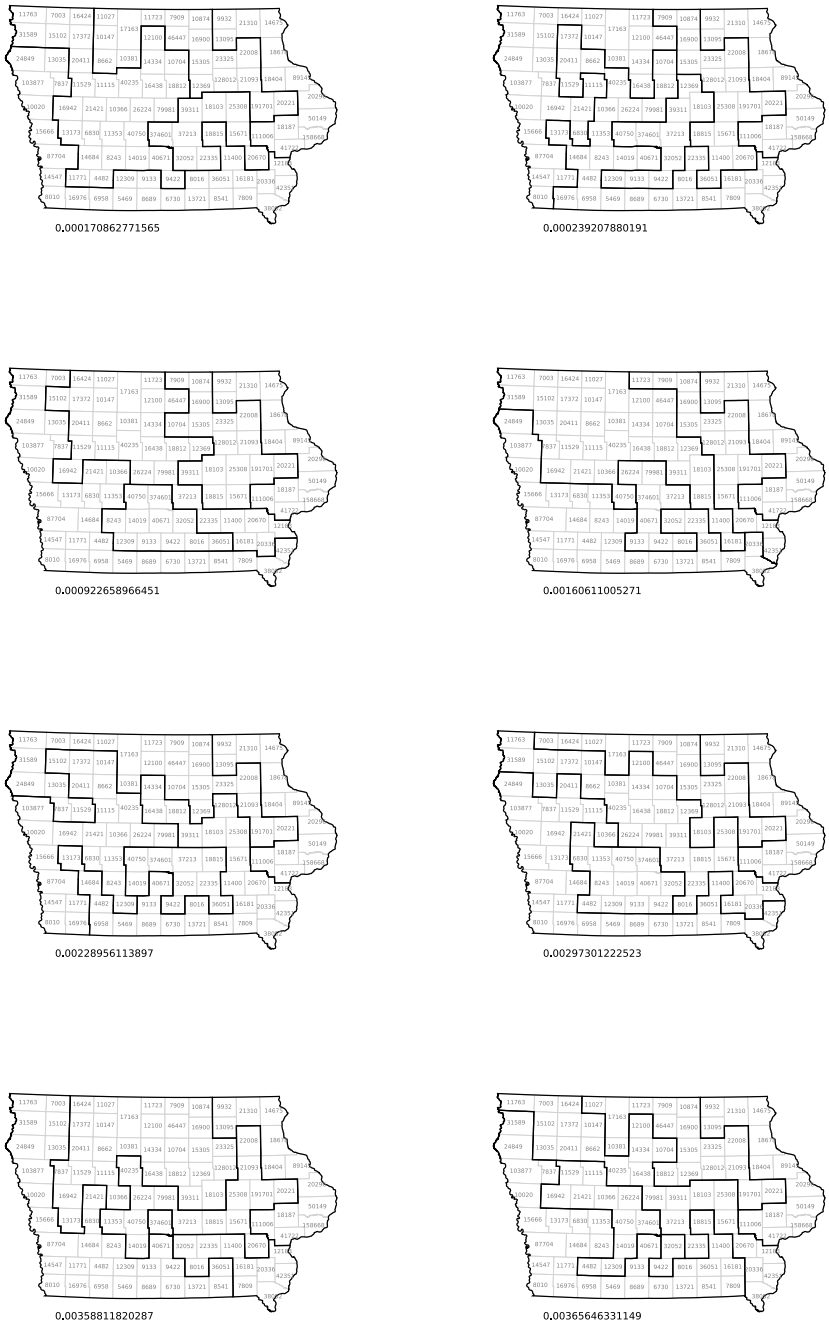


Figure 1: Political redistricting plans for Iowa (using 2000 census data).

A Principal Curve-based method for Geospatial Data Smoothing

Xiliang Liu, Feng Lu, Kang Liu, Peiyuan Qiu, Li Yu, Mingxiao Li

State Key Lab of Resources and Environmental Information system,
Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences
{liuxl;luf;liukang;yul;limx}@lreis.ac.cn

Abstract

We propose a principal curve-based method for geospatial data smoothing. Firstly we test its performance with traditional approaches using floating car data (FCD). Secondly we evaluate its robustness with spatial-temporal dependence using Spearman rank correlation analysis. Final results show that the proposed method not only takes precedence over traditional methods (Mean and Median) in accuracy (about 10%-15% higher in RMSE), but also performs more robust, showing a distinct changing trend of the original data. These findings demonstrate the feasibility of the principal curve-based method in geospatial data smoothing.

Keywords: Principal curves, Data smoothing, Robustness

1. Introduction

Nowadays various GPS-equipped sensors, such as operating vehicles (taxicabs, probe cars, buses, private cars, etc.), mobile phones, wearable devices and so on, have become the mainstream in the research of GIScience and many other location based services (LBS) because of the cost-effectiveness and flexibility compared with other data sources. However, these geospatial data collected from GPS devices cannot be utilized directly owing to: (1) the sampling interval in most cities is low-frequency due to transmission bandwidth, energy consumption and storage pressures, and (2) the spatial-temporal distribution of these GPS-equipped devices among a city or a given region is heterogeneous. To further mine these data, the data sparseness, data missing and noise problem make geospatial data smoothing an unavoidable step.

Previous studies mainly focus on parametric approaches including Kalman filter, particle filter, piecewise linear (PWL) curves, and so on. These methods can effectively deal with data noise problem, but behave unsatisfactory with data sparseness and data missing problems. Traditional Mean and Median are also conducted by using current and near-past records from a historical perspective. However, the prerequisite of Gaussian distribution in most cases cannot be satisfied so that the traditional Mean and Median methods can only be employed in linear systems.

In this paper, we propose a principal curve-based method in geographical data interpolation. We evaluate its performance using floating car data (FCD), and analyze its robustness with Spearman's rank correlation analysis. A series of experiments demonstrate its feasibility for geospatial data smoothing.

2. Methodology

Principal curves give a summarization of the data in terms of a 1-*d* space nonlinearly embedded in the data space (Hastie and Stuetzle 1989). The original definition of a principal curve $f(t) = (f_1(t), ..., f_d(t))$ relies on the self-consistency property of principal components,

following these requirements: (1) f does not intersect itself; (2) f has finite length inside any bounded subset; (3) f is self-consistent.

We design the iterative strategy for principal curves as follows:

Algorithm 1. PrincipalCurveIteration

Input: Time-labelled geographical data list x;
Initial value formula f;

Output: principal curve list $t_f(x)$

1: Initialization.
The initial principal curve $f^{(j)}(t)$ is defined as the first line principal component, here $j=1$;

2: Projection.
$\forall x \in R^d$, calculate the $t_f(x)$ in Equation 2 using Euclidean distance.

3: Expectation.
Based on the self-consistent character, the first principal curve is re-calculated as follows:

$$f^{(j)}(t) = E[X \mid t_{f^{(j)}}(X) = t] \tag{3}$$

4. Adjustion.
If $1 - \frac{\Delta(f^{(j+1)})}{\Delta(f^{(j)})} < \varepsilon$, the iteration is stopped;
Else $j = j+1$, and go to step 2

To evaluate the performance of this proposed method, we employ two means: the smoothing accuracy and the spatial-temporal dependence using root mean square error (RMSE) and Spearman's rank correlation analysis.

3. Experiments

3.1 Smoothing accuracy analysis

We employ intersection delay data (Liu *et al*, 2013) which are derived from Beijing's FCD to analyze the smoothing accuracy. 400 main intersections from all the 14,614 ones are selected to represent the skeleton of Beijing's road network. We record the road ID, the driving directions, the time interval ID and the intersection delays. The whole process is as follows:

(1) For a given intersection, the total turn delay records is expressed as $Td = \{tde_1, td_2, \ldots, td_n\}$.Here n stands for the number of driving directions. For a given driving direction i ($i=1,2,..,n$), the intersection delay records are $td_i = \{td_{i1}, td_{i2}, \ldots, td_{im}\}$, where m stands for the record number in a given time slot. The time slot starts from 1 to 96, the span of time interval is 15 minutes in the original intersection delay dataset;
(2) Test if the intersection delays follow normal distribution in this time span. If so, go to (4), or else go to(3);
(3) Smooth the given intersection delay records based on the proposed principal curve method (Algorithm 1).
(4) Average all the intersection delays as the final turn delay value for this given time slot.

We also employ two classical approaches, Mean and Median. In order to demonstrate the performance for the real dynamic situations, we apply these three methods for the intersection delay smoothing for a given driving direction of a selected intersection in Figure 1.

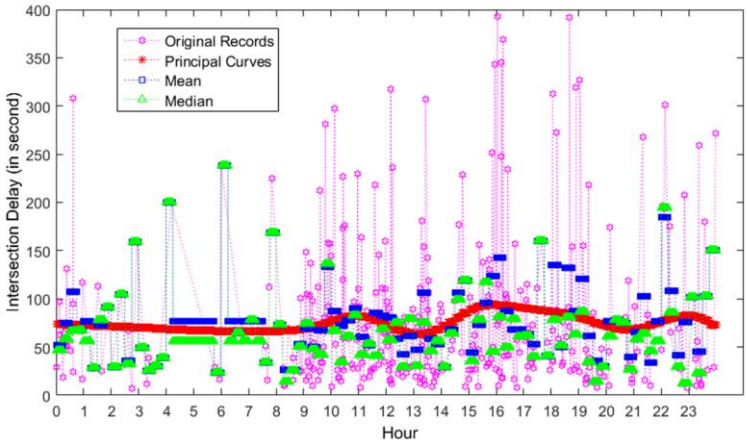


Figure 1. Results comparison (in a whole day)

During all the experiments among these 400 main intersections, the proposed principal curve-based method generally surpasses the traditional Mean and Median methods by 10%-15% in root mean square error (RMSE) for a given FCD file, more higher during the peak hours and for the marginal intersections which have fewer taxicab records.

3.2 Spatial dependence analysis

We utilize road travel time dataset of Beijing's road network ranging from February to June in 2012. In total, the average length of the road segments is 309.7 meters. The sampling interval for this dataset is 5 minutes, with 288 time slots in a day. The main problems of this dataset exist in that the records of some road segments are not complete due to heterogeneous distribution of GPS devices. Furthermore, some records are obviously abnormal during a given time interval.

We put the principal curve-based method, the mean and median into the smoothing of the road travel time dataset with the same parameters in Algorithm 1. In the implementation of Mean and Median, the time window size is set as 5 according to Liu *et al*. (2013) so as to keep the details of the original data and satisfy the smoothing requirements. We perform the Spearman's rank correlation analysis for these three methods and fit the original data with linear regression. For simplicity, we compare the fitted values between three different methods, as shown in Figure 2. In Figure 2, the fitted values of Spearman's rho between each results and the original data all elevate when the road' length increases, and the correlation between the results of principal curves and the original data behaves the strongest compared with the other two traditional ones.

3.3 Temporal dependence analysis

We design the experiment for temporal dependence analysis based on the Shanghai's high-frequency FCD. The average sampling interval is 5 seconds, ranging from 2012.10.1 to 2012.12.31. A typical intersection is selected with 1,509,906 trajectories and 1,006,708,483 GPS observations. For each driving direction, we calculate the intersection delays under different sampling intervals (5s, 10s, 15s, 30s, 60s, 90s, 120s). In order to quantify the temporal dependence among different sampling intervals, we first take the all intersection delay records of a given driving direction as a whole time series list. Then we employ different methods, including principal curves, Mean and Median. After the smoothing, we estimate the Spearman's rank correlation coefficient ρ between different results and the original data of a specific driving directions under different sampling intervals (i.e. 5s, 10s, 15s, 30s, 60s, 90s, and 120s). Finally, we average all the 12 driving directions' Spearman's ρ

under different sampling intervals. Figure 3 illustrates the results of temporal dependence analysis.

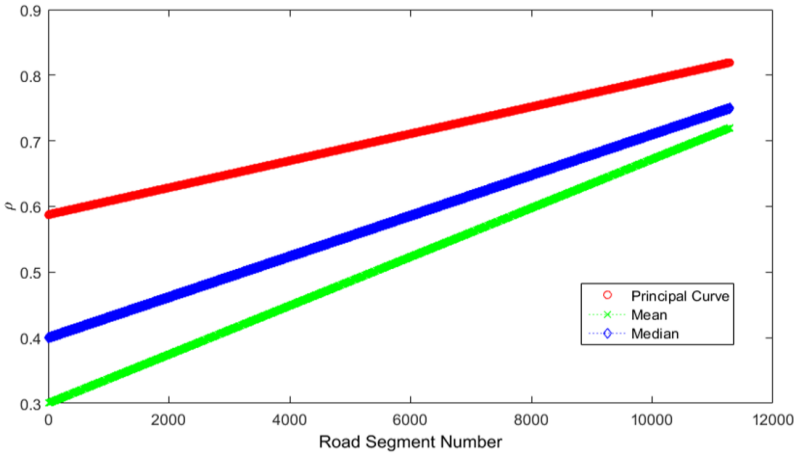


Figure 2. Spatial dependence comparison between fitted values of Spearman's rho

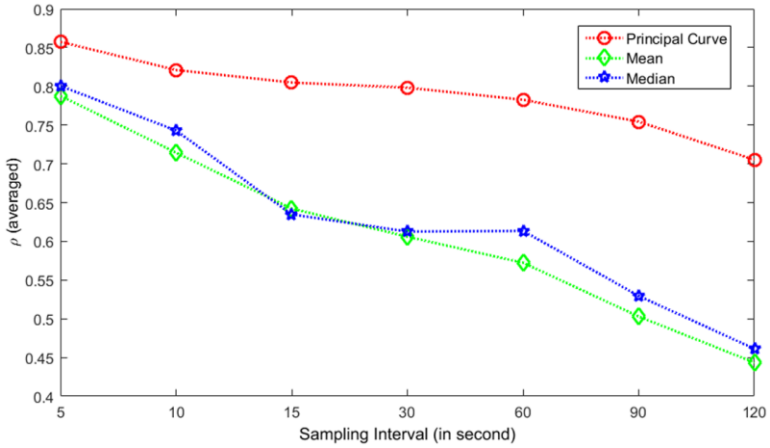


Figure 2. Temporal dependence comparison between fitted values of Spearman's rho

4. Conclusions

We propose a principal curve-based method for geospatial data smoothing. The proposed method not only takes precedence over traditional methods (Mean and Median) in accuracy (about 10%-15% higher in RMSE), but also performs more robust in dealing with data sparseness, data missing and noise problems, showing a promising feasibility in geospatial data smoothing.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No. 41271408), the National Hi-tech Research and Development Program of China (Grant No. 2013AA120305) and China Postdoctoral Science Foundation funded project (Grant No. 2015M581158).

References

Hastie T, Stuetzle W, 1989, Principal curves. *Journal of the American Statistical Association*, 84(406):502-516.

Liu X L, Lu F, Zhang H C, Qiu P Y, 2013, Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. *Frontiers of earth science*, 7(2):206-216.

A Data-Driven Approach for Detecting and Quantifying Modeling Biases in Geo-Ontologies Using a Discrepancy Index

Bo Yan, Krzysztof Janowicz, and Yingjie Hu

STKO Lab, Department of Geography, University of California, Santa Barbara, USA
{boyan,jano,yingjiehu}@geog.ucsb.edu

Abstract

Geo-ontologies play an important role in fostering the publication, retrieval, reuse, and integration of geographic data within and across domains. The status quo of geo-ontology engineering often follows a centralized top-down approach, namely a group of domain experts collaboratively formalizing key concepts and their relationships. On the one hand, such an approach makes use of the invaluable knowledge and experience of subject matter experts and captures their perception of the world. On the other hand, however, it can introduce biases and ontological commitments that do not well correspond to the *data* that will be semantically lifted using these ontologies. In this work, we propose a data-driven method to calculate a *Discrepancy Index* in order to identify and quantify the potential modeling biases in current geo-ontologies. In other words, instead of trying to measure quality, we determine how much the ontology differs from what would be expected when looking at the data alone.

Keywords: geo-ontology; ontology engineering; DBpedia; Linked Data; Discrepancy Index

1 Introduction

Due to the diverse and eclectic nature of geographic information, geographic data usually comes from different sources, in different formats, and are conceptualized from different perspectives. These heterogeneities in terms of provenance and standards create a barrier for integrating data to perform more comprehensive analysis. Geo-ontologies provide a promising way to alleviate this long-standing issue by enabling a flexible integration of geographic information based on semantics, i.e., regardless of representational choices and syntax.

However, the common ways in which geo-ontologies are developed top-down by a team of knowledge engineers and domain experts carry the risk of generating biased or unsuitable geo-ontologies (Hu and Janowicz, 2016). To give a concrete example, in the current version of DBpedia's ontology (DBpedia 2015-10), the class *Canal* is classified as a sibling class of *River*, and both are defined as subclasses of *Stream*. This seems to be a rational classification at first glance since canals are usually channels of water. However, *Stream* is a subclass of *BodyOfWater* and *BodyOfWater* is a subclass of *NaturalPlace*. Due to the transitivity of the *rdfs:subClassOf* relationship, canals become natural places. However, this seems like an odd modeling choice as canals are defined as "an artificial waterway constructed to allow the passage of boats or ships inland or to convey water for irrigation" according to the Oxford dictionary. Words such as "artificial" and "constructed" make canals man-made features rather than natural place. This example indicates that top-down geo-ontologies may suffer from the issues such as modeling biases, oversights, and ontological commitments that do not well represent the real data needs.

Scrutinizing the geo-ontologies and making revisions manually on a regular basis are common solutions to such problems. But such methods are usually labor-intensive and create a gap between the geo-ontology and its corresponding Linked Dataset. In this research, we introduce initial results on a *Discrepancy Index* that helps geo-ontology engineers by detecting and quantifying potential issues using a series of data mining steps.

2 Proposed Method

Our approach consists of two parallel threads. The first thread comes from Linked Datasets that are transformed from unstructured data, such as Wikipedia pages. This thread focuses on the bottom-up

part. The second thread originates from the top-down geo-ontologies which are constructed manually by expert with their domain knowledge.

From a Linked Dataset, we select instances and properties concerning the specific classes in the top-down ontology. These instances and properties then act as input for our data-driven approaches. During the feature extraction, we focus on properties in each class. Properties in a Linked Dataset are analogous to attributes of different place types. The rationale is that similar place types share similar attributes while distinct place types have distinct attributes. For example, the place types *City* and *Town* are similar and they have similar properties such as *populationOf* and *totalAreaOf*. However, *City* and *Mountain* are very different from each other because a mountain can have a peak whereas a city usually does not. Based on this train of thought, we build a feature set that shows the relative frequency of each property in each class.

In pursuance of comparing the results of our data-driven approach, we also consider different variations of the feature set. We take into account four variables in our feature selection. They are *filler*, *specificity*, *literal* and *uniformity* (Table 1). All of them are boolean variables. The variable *filler* decides whether we use {property, object type} pairs or property alone to count the frequency. The variable *specificity* takes into consideration the hierarchical structure of object types. The variable *literal* acknowledges the fact that, in Resource Description Framework (RDF), object and literal have different typing schemes, namely object type and data type. The variable *uniformity* considers the cases in which the literal of the same property has different data types because the original Wikipedia page does not define a uniform data type for each infobox entry. For example, a city may have a property *population*, but the literal value for this property may be of type integer, double or even string. In a nutshell, *filler* and *literal* decide whether we incorporate object type and literal type into our feature extraction; *specificity* and *uniformity* deals with the granularity and accuracy of object type and data type respectively.

Table 1: Definition for four variables

	True	**False**
Filler	Include object types	Do not include object types
Specificity	Include only the most specific object types	Include all object types
Literal	Include literal types	Do not include literal types
Uniformity	Unify literal types	Do not unify literal types

Considering all the variables listed here, we have seven feature sets. This whole process of feature selection can be viewed as a decision making process, visualized in a decision tree shown in Figure 1. We make a boolean decision on one of the four variables described above at each internal node. Each internal node branches into two sub-nodes which are the outcomes of the two decisions based on each of the four variables in our feature selection process. The seven leaf nodes are the final outcomes of the decision tree, which are also the seven feature sets.

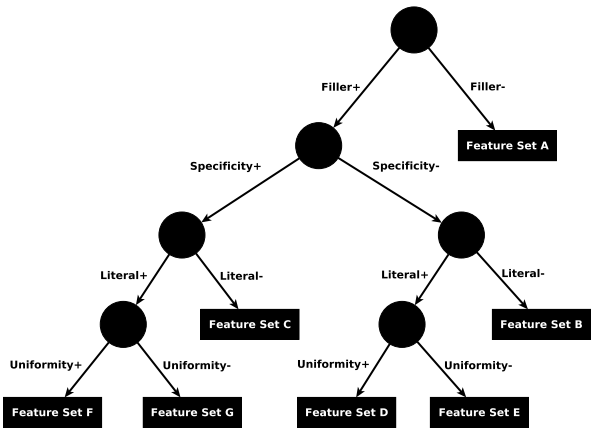


Figure 1: The tree structure of feature sets

In order to avoid the curse of dimensionality, we transform our raw feature space into a lower dimensional space using Multidimensional Scaling. We then use hierarchical clustering to obtain the hierarchies derived from the selected feature sets. In the next step, information content-based semantic similarity measures (Jiang and Conrath, 1997; Sánchez et al., 2011) are implemented to compute the pairwise similarity between each pair of classes in the derived as well as the original ontology hierarchy. The Mantel test (Mantel, 1967) is implemented to select the best representing feature set based on the correlation coefficient and p-value. The result shows that Feature Set E performs the best. In the end, we

use the semantic similarity results from the best feature set to calculate the *Discrepancy Index Matrix* $IndexMatrix = SimMatrix_{original} - SimMatrix_{derived}$. Individual *Discrepancy Index* can be found in this matrix.

The value range for the *Discrepancy Index* is $[-1, 1]$. If $IndexMatrix(i, j) > 0$, it implies that class i and j are less similar in the derived hierarchy; whereas if $IndexMatrix(i, j) < 0$, it tells us that class i and j are more similar in the derived geo-ontology. The value of $|IndexMatrix(i, j)|$ gives us the information about the extent to which the similarity in two hierarchies differ from each other. This *Discrepancy Index* is useful in assisting geo-ontology engineers to refine and further develop the geo-ontology in that it gives guidance on correcting the potential modeling biases or misclassifications.

3 Case Study: Are *Canals Natural Places*?

Among the various different types of geo-ontologies, we selected the DBpedia ontology for our case study. Previously, we discussed that *Canal* should not be classified as the subclass of *NaturalPlace* based on its definition. To further verify this, we can check if the DBpedia Linked Dataset supports our observation. Browsing through the DBpedia page of Panama Canal, we found properties such as *dbo:principalEngineer*, *dbp:dateUse* and *dbp:company* (see Figure 2). Intuitively, it is unreasonable for an instance of *NaturalPlace* to have principal engineers, to have the date of first use, and to be originally owned by a company. We can apply the *Discrepancy Index* to see if our approach can detect such a case.

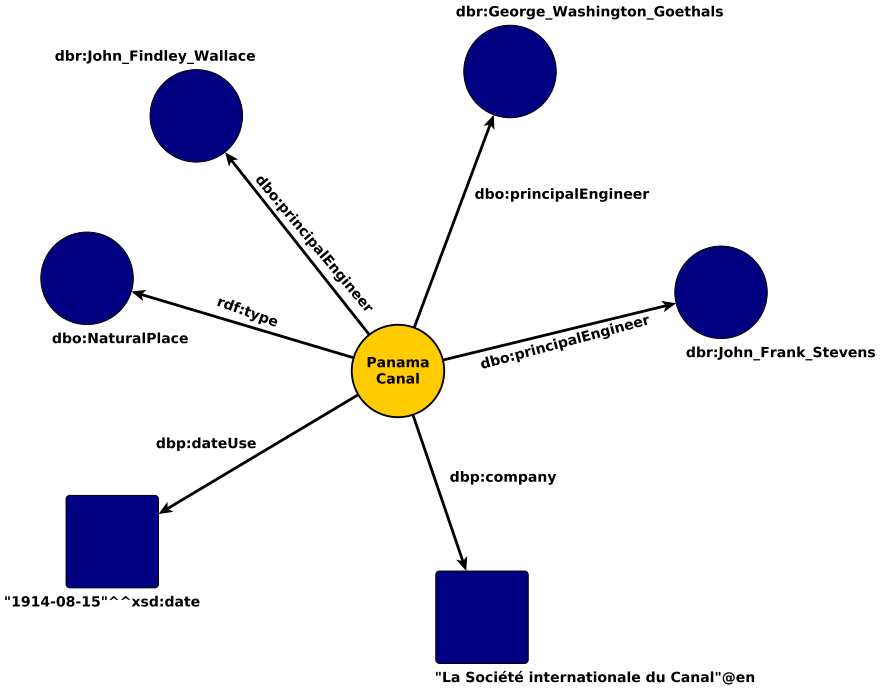


Figure 2: Panama Canal in DBpedia

We approach this by comparing the semantic similarity between *Canal* and its sibling classes, in this case, *River*. We would assume that since *Canal* and *River* are sibling classes, they are semantically similar to each other or at least more similar to each other than to classes in other branches of the place type hierarchy. We first plot the bar graphs for *Canal* and *River*. The bar chart uses information from the feature space after dimensionality reduction. The reasons we use the transformed feature space are two-fold. First, the original feature space contains numerous features which are hard to visualize and plot in a graph. Moreover, the original feature space is very sparse, making it difficult to analyze the results in a plotted graph. Second, some of the features are dependent on each other, making some of the information redundant and unreliable. The transformed feature space only contains 63 features and all of the features are independent from each other. Figure 3 and Figure 4 show the bar chart of these two classes. The x axis represents the statistical features for each class in the same order while the y axis is the value for each feature. From here, we can tell that *Canal* has a distinct pattern from *River*. However, the charts alone cannot quantify the difference between these classes. Moreover, it is also impossible to obtain the same kind of graphs using the original hierarchy in the ontology. Thus, these charts alone cannot hint at the direction of required geo-ontology refinements and the extent of the modeling biases.

Therefore, we can use the proposed *Discrepancy Index* to help us. After looking up the similarity matrices for the original hierarchy and the derived hierarchy, we find that the similarity between *Canal*

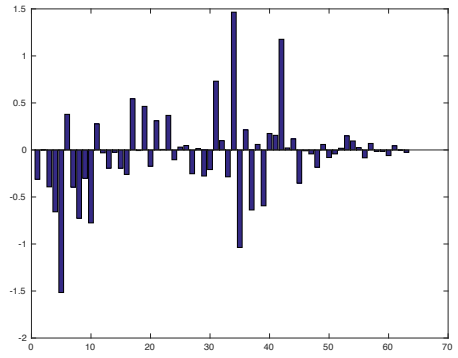


Figure 3: Feature values for *Canal* after MDS

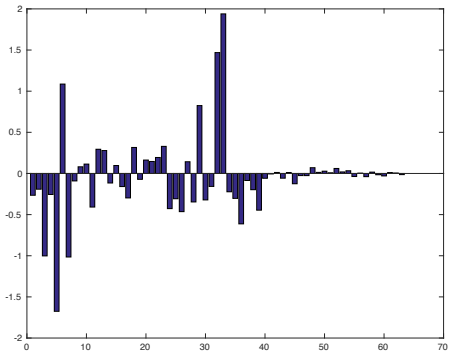


Figure 4: Feature vlaues for *River* after MDS

and *River* is 0.84 and 0.23 respectively. The *Discrepancy Index* for (*Canal*, *River*) is 0.61. This index implies that *Canal* is less similar to *River* in the derived hierarchy and leads to further investigation, which in this case is the result of modeling bias. This result corresponds to our observation in the DBpedia geo-ontology.

4 Conclusion and Future Work

Current geo-ontology engineering procedures often heavily depend on the knowledge of domain experts and a top-down style of engineering. The potential pitfall to this routine is that the resulting geo-ontologies may be biased and not representative of the data that will be semantically lifted using these ontologies. In this initial research we propose a data-driven approach that integrates geo-ontologies and Linked Dataset during the dynamic course of geo-ontology engineering and assist engineers in identifying and quantifying potential geo-ontology modeling bias via a *Discrepancy Index*. The initial case study suggests that the results returned by our method correspond to our observation, hinting at the usefulness of the *Discrepancy Index*.

This work can be extended in several aspects. First, this initial method can be extended into a systematic framework that can be applied to a variety of geo-ontologies and to guide engineers in understanding differences and similarities in their conceptualizations Janowicz et al. (2008) . Second, our experiment so far focuses on one particular ontology and dataset from DBpedia. With a wide range of availability of Linked Data and ontologies on the Web, we can test our approach using different data sources. Moreover, candidate solutions to the bias detected by the data-driven method can be developed in future work.

References

Hu, Y. and Janowicz, K. (2016). Enriching top-down geo-ontologies using bottom-up knowledge mined from linked data. In Onsrud, H. and Kuhn, W., editors, *Advancing Geographic Information Science: The Past and Next Twenty Years*, chapter 13, pages 183–198. GSDI Association Press.

Janowicz, K., Maue, P., Wilkes, M., Schade, S., Scherer, F., Braun, M., Dupke, S., and Kuhn, W. (2008). Similarity as a quality indicator in ontology engineering. In *Formal Ontology in Information Systems*, pages 92–105. IOS Press.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2 Part 1):209–220.

Sánchez, D., Batet, M., and Isern, D. (2011). Ontology-based information content computation. *Knowledge-Based Systems*, 24(2):297–303.

Location Optimization of Fire Stations: Trade-off between Accessibility and Service Coverage

J. Yao[1], X. Zhang[2]

[1]Urban Big Data Centre, University of Glasgow, 7 Lilybank Gardens, Glasgow, G12 8RZ, UK
Email: Jing.Yao@glasgow.ac.uk

[2]Department of Geographical Information Science, Hohai University, 1 Xikang Road, Nanjing, 210098, China
Email: Xiaoxiang@hhu.edu.cn

Abstract

Fire and rescue service is one of the fundamental public services provided by government in order to protect people, properties and environment from fires and other disasters, and thus promote a safe living environment. Efficient deployment of fire stations is necessary and essential if timely response to the emergencies is to be achieved. Spatial optimization approaches have been long employed in public facility location studies. In particular, coverage-based models, such as the location set covering problem (LSCP) and the maximum coverage location problem (MCLP), have been widely adopted to achieve complete or maximum coverage of service demand. This paper extends the LSCP by accounting for both partial coverage and access to the demand areas. The proposed model is applied to the optimization of fire station locations in Nanjing, China. The results can be used to assist future fire station location planning and rescue resource deployment.

1. Introduction

Fire caused by humans or nature can pose hazard to people, properties and environment, and lead to psychological damage, physical injuries (even death) and significant economic losses. Fire prevention and protection is necessary and essential for a safe living environment. The associated fire and rescue service therefore needs to be properly deployed to ensure efficient fire safety management. A fundamental concern in this regard is the spatial configuration of fire stations as it is critical to timely response to emergency calls.

Given the inherent spatial nature, fire station location problems have been well studied using geographical information system (GIS)-coupled location modelling (Chevalier *et al*. 2012; Aktaş *et al*. 2013; Murray 2013). In particular, LSCP (Toregas *et al*. 1971), MCLP (Church and ReVelle 1974) and their extensions have long been employed to evaluate the locational efficiency of existing fire stations as well as seek sites for new fire stations (Chevalier *et al*. 2012; Murray 2013). Common goals of locating fire stations include maximizing the access to provided services, covering as much demand as possible and minimizing total costs of service provision, usually subject to available resources. In practice, two or more objectives are often considered to capture different aspects in relation to fire service delivery.

The aim of this paper is to seek best locations of fire stations with spatial optimization approaches, particularly considering accessibility and service coverage. The proposed model is applied in an empirical study in Nanjing, China, to assist future fire station location planning and rescue resource deployment.

2. Methods

The proposed spatial optimization model is built on LSCP but with extensions from three aspects. One extension is to allow partial service coverage. That is, unlike the original LSCP where a neighbourhood (areal demand) is considered covered by fire service only if it is completely within the service area of one or more fire stations, we allow the demand to be partially covered by multiple fire services to achieve a full coverage. The second extension is to include a preference that the selected locations are closer to the demand areas with higher fire risks. Finally, we account for the impact of existing fire stations. Consider the following parameters:

I, J: set of demand areas and potential fire station locations, respectively;
i, j: index of demand areas and potential fire station locations, respectively;
w_i: estimated fire risk at i;
d_{ij}: distance between i and j;
a_{ij}: proportion of i's area covered by j;
Ω_i: set of fire stations that can provide service to i;
Φ: set of existing fire stations;
p: number of existing fire stations that remain in the final solution;

and the decision variable:

$$Y_j = \begin{cases} 1 & \text{if a fire station is sited at } j \\ 0 & \text{otherwise} \end{cases}$$

The proposed model can be formulated as follows:

$$\textit{Minimize} \quad \sum_{j \in J} Y_j \tag{1}$$

$$\textit{Minimize} \quad \sum_{i \in I} \sum_{j \in \Omega_i} w_i d_{ij} Y_j \tag{2}$$

$$\textit{Subject to} \quad \sum_{j \in \Omega_i} a_{ij} Y_j \geq 1 \quad \forall i \in I \tag{3}$$

$$\sum_{j \in \Phi} Y_j = p \tag{4}$$

$$Y_j \in \{0, 1\} \quad \forall j \in J \tag{5}$$

In the above definition, objective (1) is to minimize the total number of fire stations, and objective (2) is to minimize the total weighted travel distance to the places requesting services, implicitly encouraging siting fire stations near the demand with higher fire risks. Constraints (3) require that each neighbourhood can be fully covered by the total services from the qualified stations, which are defined by Ω_i based on response time or service area. Constraint (4) from Schilling *et al*. (1980) limits the number of existing fire stations that will remain in the solution. Finally constraints (5) ensure that the decision variables only can have values 0 or 1.

It is worth noting that although constraints (3) extend LSCP by allowing the full service coverage to be achieved through the overall partial coverage, the complete service coverage for a neighbourhood might not be necessarily guaranteed due to the overlap among partial coverage. Nevertheless, according to objective (1), the model implicitly will disperse the selected fire stations to minimize the total number while satisfying constraints (3). Tong (2012) showed that MCLP using partial coverage could improve computation efficiency while generating satisfactory results. Similar method was also employed in Murray (2013) for location modelling of fire stations.

3. Preliminary Results

The above model is applied in an empirical study in Nanjing, China, to evaluate the locational efficiency of existing fire services and seek locations for new fire station location. The study area

is located in the south of Yangtze River within Nanjing, China, covering seven main districts of the city. The total area is about 598.1 km^2 (9.1% of total area of Nanjing) and the total population is about 5.06 million (2010) (54.4% of total population of Nanjing), currently served by 19 fire stations.

The study area is discretised into a set of 1 km * 1 km grid cells as this is the finest scale available for the population data. Thus, each grid cell represents a demand area with the fire risk estimated by a combination of population information and the fire incidents during 2002 – 2013. The potential fire station locations are represented by the centroids of those grid cells. To solve the model, the two objectives are combined through a weighted sum. That is, a weight ranging from 0 to 1 is assigned to each objective so that the sum of the two weights equals one, which is commonly used in solving multi-objective optimization problems.

Two scenarios are considered here: one is to assume no existing fire stations, and the other is to keep all the existing 19 fire stations in the final solution. The results are presented by Figure 1. For both scenarios, the two objectives have equal weights, that is, 0.5 and 0.5, respectively.

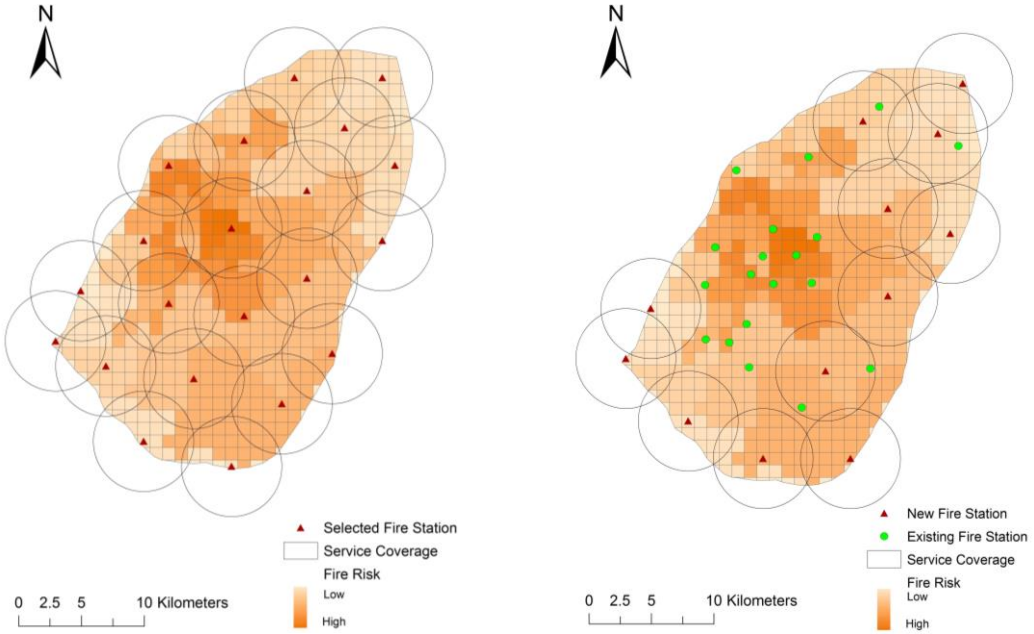


Figure 1. Optimal Solutions:
(a) assuming no existing fire stations; (b) including existing fire stations.

As shown in Figure 1(a), ideally total 21 fire stations are needed to obtain a trade-off between the proximity to high-risk areas and the service coverage. If keeping all the 19 existing fire stations, Figure 1(b) shows that additional 12 new sites are necessary to achieve the same goals. Also, it can be seen that most areas with higher fire risks have been covered by the existing stations around the city center.

4. Summary

Fire and rescue services remain fundamental to the safety of human beings, properties and the physical environment. Efficient fire prevention and protection can greatly reduce the losses of lives and economies. This paper proposed a spatial optimization model to find the best locations

for fire stations, particularly considering the trade-off between accessibility and service coverage. The empirical results demonstrated that spatial optimization can be a powerful tool to assist with the spatial deployment of fire service resources.

References

Aktaş E, Özaydın Ö, Bozkaya B, Ülengin F and Önsel Ş, 2013, Optimizing fire station locations for the Istanbul metropolitan municipality. *Interfaces*, 43(3):240-55.

Chevalier P, Thomas I, Geraets D, Goetghebeur E, Janssens O, Peeters D and Plastria F, 2012, Locating fire stations: an integrated approach for Belgium. *Socio-Economic Planning Sciences*. 46(2):173-82.

Church R and ReVelle C, 1974, The maximal covering location problem. *Papers in Regional Science*, 32(1):101-118.

Murray AT, 2013, Optimising the spatial location of urban fire stations. *Fire Safety Journal*, 62:64-71.

Schilling DA, ReVelle C, Cohon J and Elzinga DJ, 1980, Some models for fire protection locational decisions. *European Journal of Operational Research*, 5(1):1-7.

Tong D, 2012, Regional coverage maximization: a new model to account implicitly for complementary coverage. *Geographical Analysis*, 44(1):1-14.

Toregas C, Swain R, ReVelle C and Bergman L, 1971, The location of emergency service facilities. *Operations Research*, 19(6):1363-1373.

Using GPS-enabled mobile phones to characterize individuals' activity patterns for epidemiology applications

E.-H. Yoo and Y.-S. Eum

University of Buffalo, SUNY, Buffalo, NY, USA
Email: {eunhye, yeum}@buffalo.edu

Abstract

We assessed the potential of global positioning system (GPS)-equipped mobile phones for health-related studies. We demonstrated the use of GPS data as a means of collecting individuals' activity patterns for personal exposure assessments and public health surveillance. The widespread use of mobile phones has enabled investigators to conduct exposure studies and to detect infectious disease at the individual level on a massive scale. However, still substantial uncertainties are present in converting raw GPS data into relevant information. To address these issues, we proposed three algorithms for pre-processing and classification of raw GPS data, and demonstrated their applications to real world data in a case study.

1. Introduction

Exposure models typically impose unrealistic assumptions such that subjects within a neighborhood are equally exposed to air pollution and/or most individuals spend their time at their residences. Similarly, a lack of understanding of human movement, which is an important component of disease transmission, has been considered as an obstacle to develop effective national communicable disease control programs. In exposure modelling, some improvements have been achieved by adopting a microenvironment (ME) approach where individuals' time spent at MEs, such as outdoors, residence, and workplace, was explicitly taken into account. However, collecting the information on individuals' time-activity patterns has been cost-, time-, and labor-intensive with limited reliability and accuracy. Comparably, aggregated data have limited efforts to reconstruct the complex and dynamic nature of real-world contact networks, which plays a critical role of contact network in an outbreak of dangerous infectious disease. The emergence of lightweight, low-cost, and accurate GPS devices has provided a promising tool for objectively assessing the geographic positions of the environmental context in which health-related behaviors take place (Schipperijn *et al.* 2014). GPS technology enabled investigators to capture daily trajectories of individuals with higher temporal resolution at increasing locational precision (Gerharz *et al.* 2013, Dias & Tchepel 2014), although the use of GPS data in health research is not without challenges. As reviewed by Krenn *et al.* (2011), the positional accuracy of GPS data collected in real world is often unacceptable in health studies, especially, over longer study periods. The data quality of GPS traces depends on the amount of data lost from signal drop-outs, loss of device battery power, and poor adherence of participants to following the specific research protocol. Despite the advancement in GPS technology, signal acquisition is still affected by the presence of tall buildings and significant uncertainties associated with the processing and classifying raw data are present.

In this study we focus on the GPS data collected from a mobile phone with and without data connection. Our primary goal is to identify major MEs associated with health-related activities using GPS data. First we developed and applied a "selective resampler" to raw GPS data for pre-processing. Using the processed data, we identify significant places and travel modes using the two automated classification schemes.

2. Study Area and Data

The GPS data are collected from Mobile phone (Moto X with Android 5.1) in a highly urbanized environment, Sydney in Australia, for 9 days in February 2016. We collected data for a week and two randomly selected weekdays from one subject (Figure 1). We used GPS Logger for Android (GPS Logger 2016), which is a lightweight, battery efficient app to log GPS coordinates at 1-minute interval to a file. This app runs in the background of the phone and upload the data to a cloud server every 30-minutes with varying positional accuracy.

Figure 1. Activities identified by MEclassifier for weekday (left) and weekend (right)

3. Methods

3.1 Pre-processing raw GPS data

We cleaned raw GPS data by excluding data points with high positional error (e.g., above 100m) to avoid spurious results. Followed by the elimination of extreme outliers, we applied resampling to non-moving objects, i.e., GPS point associated with static activities. Our algorithm ("selective resampler") is different from the previous work (Dodge *et al.*, 2009), which enables us to avoid making unrealistic assumptions about human movement, such as it being linear, by performing resampling only on static activities. Our selective resampling algorithm requires the specification of two parameters: distance and time thresholds. These two parameters determine if any successive GPS points belong to a static activity with missing data points by comparing their separation distance and time interval to the specified thresholds. That is, resampling will be performed only if the distance and time interval between any two consecutive points is less than the distance threshold and greater than the time threshold, respectively.

3.2 Extracting significant places from GPS traces

We classified GPS traces to extract the information on static MEs using "MEclassifier", which is an extension of the density-based spatio-temporal clustering algorithm, ST-DBSCAN (Birant & Kut 2007). ST-DBSCAN detects non-linearly shaped clusters by separating clusters from noises by taking into account both spatial and temporal attributes. The unique feature of MEclassifier lies in the ability of discerning GPS-inherent error from true noise. We explicitly accounted for temporal continuity of GPS data by re-defining GPS noise, while defining moving activity as the GPS points between clusters. We developed a merging procedure to address acceptable GPS error and to simultaneously avoid situations where a single space-time cluster is divided into two MEs. We achieved this goal by defining two additional parameters to determine if any discrete cluster would be merged into a single cluster if both the minimum spatial distance between their centroids and the time interval are

smaller than the specified merge parameters. Lastly, we contextualized the space-time clusters by matching with the information from a baseline survey, such as home and work address of individuals, and digital parcel data.

3.3 Detecting travel mode

We developed a travel model detection algorithm, named as "TMdetector", by combining statistical methods with machine learning algorithms. Any GPS trace connecting significant places (MEs identified from 3.2) was defined as a moving activity, which was further partitioned into multiple segments based on the mean and variance of speed. More specifically, we used a Pruned Exact Linear Time optimization method (Killick *et al.* 2012) to compute a modified Bayesian Information Criterion as a penalty function (Zhang & Siegmund 2007). The prediction of travel mode was obtained using random forest classifier based on speed, acceleration, and travel distance information.

4. Results

The GPS logger was programmed to record the location of the mobile device every minute, which yields a total of 1,440 GPS points per day. We quantified the amount of GPS data loss during the nine days of the study period, which varies day-to-day depending on the participant's activity patterns (see Table 1). The performance of MEclassifier is sensitive to missing GPS data because density-based clustering algorithms rely on the number and the location of GPS traces. Table 1 shows percentage of missing data associated with static activities, to which the resampling algorithm was applied. We selected two resampling parameter values based on the sensitivity analysis, which indicates that non-trivial amount of data was lost: all activities from 7.30% to 21.60% and static activities from 1.32% to 8.48% (sensitivity analysis results are not reported due to the limited space). After resampling, we re-examined the percentage of missing data for static activities, which shows that improvement was made on the four days— week days (day 4,8) and weekend (day 6, 7)—on which the subject participated on non- routine activities than routine activities of "home-work-home" as shown in Figure 1.

Table 1. Missing data (%) in all activities (All), static activities alone (Static), and static activities after resampling (Re-static)

Day	1	2	3	4	5	6	7	8	9
All	7.30	16.28	10.42	8.26	12.44	17.44	10.22	10.56	21.60
Static	2.62	6.11	2.64	6.81	1.32	7.44	8.48	5.77	6.25
Re-Static	2.62	6.11	2.64	**0.90**	1.32	**0.90**	**0.63**	**1.18**	6.25

We assessed the effects of resampling on classification by comparing the extracted information on MEs from raw GPS data and resampled data on day 4. Classification results were compared to the activity diary for validation, and the results are summarized in Table 2 and and Figure 1(Right). The classification of the raw GPS data failed to reproduce Event 2 while the Resampled data reproduce all the events.

Table 2. The effects of resampling on classification results (Day 4)

			Activity Diary		Before Resampling		After Resampling	
Event	ME-Type	TM-Type	Start	End	Start	End	Start	End
1	Home		00:00	10:20	00:00	10:20	00:00	10:20
		Walk	10:20	10:39	10:20	12:40	10:20	10:36

2	Other Indoor		10:39	12:00			10:36	12:02
		Walk	12:00	12:40			12:02	12:40
3	Home		12:40	14:31	12:40	14:31	12:40	14:31
		Walk	14:35	14:45			14:31	14:43
4	Other Indoor		14:45	16:36	14:43	16:36	14:43	16:36
		Walk	16:36	16:45			16:36	16:50
5	Home		16:45	23:59	16:50	23:59	16:50	23:59

5. Discussion and Conclusion

The importance of pre-processing of raw GPS data was highlighted for the applications in epidemiological studies, particularly when they are collected from mobile phones with potentially irregular data connection. The consequence of using missing GPS data may lead to biased or incorrect inference on the environments in which activities related to health outcomes occurred. Both the pre-processing and classification algorithms are an adjustment of existing methods, whereas the adjustment has critical implications to address research questions in epidemiological studies using GPS data collected from mobile phone. This paper presented a subset of our on-going project where the performance of the proposed algorithms has been tested using data collected from different regions over different periods. In near future we plan to integrate the extracted ME information with the air pollution concentrations to estimate personal exposure to air pollution. In addition, the identified time-activity patterns will be used to provide information on individuals' interactions and travel behaviours, which are the crucial information to capture dispersion process of influenza.

Acknowledgements

This work is funded from NIH/NIGMS), R01GM108731.

References

Birant, D. & Kut, A. (2007), 'ST-DBSCAN: An algorithm for clustering spatial–temporal data', Data & Knowledge Engineering 60(1), 208–221.

Dias, D. & Tchepel, O. (2014), 'Modelling of human exposure to air pollution in the urban environment: a GPS-based approach', Environmental Science and Pollution Research 21(5), 3558–3571.

Dodge, S., Weibel, R. & Forootan, E. (2009), 'Revealing the physics of movement: Comparing the similarity of movement characteristics of different types of moving objects', Computers, Environment and Urban Systems 33(6), 419–434.

Gerharz, L. E., Klemm, O., Broich, A. V. & Pebesma, E. (2013), 'Spatio-temporal modelling of individual exposure to air pollution and its uncertainty', Atmospheric Environment 64, 56–65.

GPS Logger (2016), 'GPS Logger for Android'. URL: https://play.google.com/store/apps/details?id=com.mendhak.gpslogger&hl=en

Killick, Rebecca, Paul Fearnhead, & IA Eckley. 2012. "Optimal Detection of Changepoints with a Linear Computational Cost." *Journal of the American Statistical Association* 107 (500). Taylor & Francis: 1590–98.

Krenn, P. J., Titze, S., Oja, P., Jones, A. & Ogilvie, D. (2011), 'Use of global positioning systems to study physical activity and the environment: a systematic review', American Journal of Preventive Medicine 41(5), 508–515.

Schipperijn, J., Kerr, J., Duncan, S., Madsen, T., Klinker, C. D. & Troelsen, J. (2014), 'Dynamic accuracy of GPS receivers for use in health research: a novel method to assess GPS accuracy in real-world settings', Frontiers in public health 2.

Zhang, Nancy R, & David O Siegmund. 2007. "A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data." *Biometrics* 63 (1). Wiley Online Library: 22–32

Study on the Effects of Human Intention on Spatial Data Quality

Bo Zhao

College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, 101 SW 26th St, Corvallis, OR 97331
Email: zhao2@oregonstate.edu

Abstract

The GIScience community has devoted considerable research to locational uncertainty rather than other types of spatial data quality issues. Today, the flood of spatial big data has brought about new concerns, such as location spoofing, GPS jamming, or AIS (Automatic Identification System) hacking. Yet, the current data quality assessment framework falls short in defining, interpreting and analyzing these critical issues. By examining the reasons for measuring the geographic world, I suggest a modification of the hypothesis of the rational geographer in this paper, and further to analyze the distinctions among mistakes, spoofing and uncertainties, with the goal of placing the identified types of locational inconsistencies into a more holistic theoretical framework for spatial data quality. I call on GIScientists to pay more attention to the role of human intentions and advocate for a human-centric assessment of spatial data generation. Only then can we more effectively handle the emerging quality issues in the era of big data.

1. Introduction

The GIScience community frequently focuses on uncertainty with regard to spatial data quality (Devillers et al., 2010). In the era of big data, the advent of mobile, social and geospatial technologies has created a considerable degree of heterogeneous, real-time and geo-referenced data. A large percentage of such data, especially VGI (Volunteered Geographic Information), geo-tagged social media, or location based service feeds, may be generated by ones' mistakes and/or created deliberately rather than being merely affected by inherent uncertainties. Although geographers have recognized the significance of human intention, the motivations of the data generator are seldom examined, and the recent popularity of spoofings, especially those in the form of location spoofing (Zhao, 2015), GPS jamming (Grant et al., 2009) and AIS hacking (McCauley et al., 2016) are often dismissed by GIScientists. In order to more effectively address this critical issue, I will investigate the role of human intentions in the process of spatial data generation and clearly distinguish among uncertainty, mistakes and spoofing.

In the remainder of this paper, I will review the concepts of error, accuracy, precision, uncertainty, mistakes, and spoofing in the context of geography. Then, I will examine the role of human motivation with regard to spatial data quality, and end with a brief concluding remark.

2. Uncertainty, mistakes, and spoofing

Since human beings are forced to view the world through a fuzzy and distorting lens, the measured data are inevitably generalized, approximated, and subject to uncertainties (Zhang & Goodchild, 2002). In other words, the way we observe the world has invoked an inevitable locational (or positional) inconsistency between any observed object in the geographic world and the data that it produces (value of the object being measured). When referring to such locational inconsistencies, geographers usually consider them to be underlying uncertainties. One older and simpler term to describe uncertainty is *error*. By definition, error is the difference between the measured value and the "true" value of the object being measured. It is represented as an estimation of the range of values within which the true value is likely to be found. There are two types of errors: *systematic*

and *random error*. While the former affects the *accuracy* of a measurement and can be reduced only by refining the method of measurement or technique, the latter affects the *precision* of a measurement and can be improved by repeating those measurements. It should be noted that systematic and random errors are inherent in the measuring process and cannot be totally eliminated. In the past three decades, the GIScience community began to avoid the term "error" by equating uncertainty as how confident about the measured value. For example, if I say the shortest distance between downtown Portland and Corvallis is 85.3 miles, +/- 0.5 miles at a 95% confidence level. This indicates I am 95% sure that the shortest distance is between 84.8 to 85.8 miles.

Notably, the underlying locational uncertainty is measured by one crucial assumption, that the measuring process was conducted by a *rational geographer* who has a basic geographic knowledge and is acting within a certain set of circumstances with ordinary prudence to control spatial data quality. However, in today's data-intensive society, rational geographers are not the only ones who can contribute to the massive amount of spatial data. Most human beings are irrational, imperfect and have a less than professional knowledge of measuring methods. Even the most careful geographer in the finest laboratory is likely to make mistakes and miscalculations. In such circumstances, the measured locational inconsistency can be regarded as a *locational mistake*. Notably, the term "mistake" is different from error. An error, inherent within spatial data, is locationally inconsistent at the uncertainty level; while a mistake, externally generated contrary to the data generator's original intention, is locationally inconsistent beyond the uncertainty level. Additionally, people may intentionally exaggerate locational inconsistencies even beyond the uncertainty level. For example, a social media user at a local night bar purposely put a geo-tag to his place of employment via a location spoofing app (Andev, 2013), with the goal of deceiving the user's followers as if he was working late rather than out drinking. Obviously, the distance between these two locations greatly surpasses the locational uncertainty level. This type of inconsistency is termed *locational spoofing*.

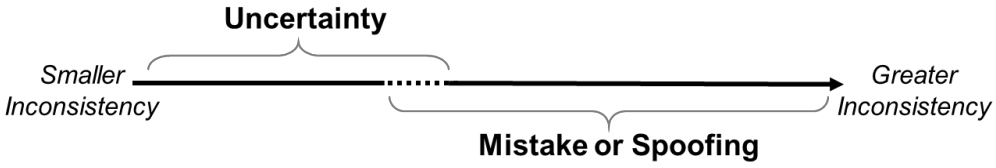


Figure 1. Comparing uncertainty to a mistake or spoofing

Notably, mistakes and spoofing usually bring about more inconsistencies than uncertainties (see figure 1). However, with respect to the degree of locational inconsistency, there is no clear demarcation between a locational mistake and spoofing. For any value that falls in the overlapping region (indicated by the dashed line), the locational inconsistency could be caused by either uncertainty, a mistake or spoofing. Regarding to the values greater than the overlapping region, it is very difficult to distinguish between a mistake and spoofing only by the degree of locational inconsistency *per se*. At this point, it is crucial to examine the intentions of the data generator.

3. The role of human intentions with regard to spatial data

In order to build a holistic framework for spatial data quality assessment, it is imperative to modify the longstanding hypothesis of rational geographer. Indeed, people may have an infinite number of reasons for wanting to generate spatial data. For example, Coleman, Georgiadou, and Labonte

(2009) have discovered eight positive intentions, including, altruism, professional or personal interest, intellectual stimulation, protection or enhancement of a personal investment, social reward, enhanced personal reputation, creative or independent self-expression, and pride of place; as well as three negative aspects: mischief, agenda setting and malicious and/or criminal intent. Furthermore, a legalistic interpretation of human intentions helps to identify hidden motivations in these spoofing cases. In Latin, the standard common law test of criminal liability states, "*Actus reus non facit reum nisi mens sit rea* (it means that the act is not culpable unless the mind is guilty)". In other words, to be guilty of committing a crime, the defendant must have knowingly committed the act (*actus reus*), accompanied by some level of guilt (*mens rea*). According to Model Penal Code[i] (Wechsler, Schwartz, Ploscowe, & Tappan, 1962), there are four levels of guilt for which a suspect is potentially culpable: purposely, knowingly, recklessly, and negligently (in descending order of severity). In addition, if a criminal act is committed out of ignorance (no explicit purposefulness is detected) or by mistake (the mind negates the guilty act), such circumstances can serve as defenses[ii] to withdraw the charge of a crime.

Table 1. Locational inconsistencies categorized by intention

	Intentional+	**Mistaken+**	**Ignorant**	**Mistaken-**	**Intentional-**
Intention	controlling locational inconsistency		no explicit intention	amplifying locational inconsistency	
Measurement	in accord with the intention	contrary to the intention	not specified	contrary to the intention	in accord with the intention
Locational Inconsistency	*uncertainty*	*a mistake*	*a mistake or uncertainty*	*uncertainty*	*spoofing*
Example	Accurate earthquake locations were monitored and then published by USGS geoscientists.	The route to India was prudently measured by Christopher Columbus. However, it led his ship to America.	Locational data was measured through a total station; the operator had little knowledge of how to measure with that equipment.	Accurate AIS data of a fishing ship was sent to the data hub. However, the captain originally wanted to hack/jam the AIS data in order to illegally fish in a prohibited fishing zone.	False locational information of a truck was generated because the driver had jammed all surrounding GPS signals to deceive his boss about the truck's whereabouts.

By the same token, I modify the hypothesis of rational geographer and argue that most measurements (*actus reus*) are the result of five major levels of intention (*mens rea*), including intentional+, mistaken+, ignorant, mistaken-, and intentional- (refer to Table 1). The symbol "+" and "-" denotes two opposite intentions of the data generators: "+" denotes that the intention is to control spatial data quality, or in other words, to control locational inconsistency at the uncertainty level; while "-" denotes the intention to amplify the locational inconsistency of an object. These five levels of intention can be qualitatively determined through participatory observation, questionnaire or online interviews. Specifically, an individual or organization measures an object of the geographic world with an intention to control locational inconsistency, if the measurement aligns with such intention, the measured locational inconsistency will be represented as uncertainty; if contrary, the measured locational inconsistency will be represented as a mistake. If the measurement is not driven by any explicit intention, the measured locational inconsistency may be viewed as either a mistake or uncertainty. Or with an intention to amplify the locational inconsistency, if the measurement aligns with such intention, the measured locational

inconsistency will be represented as spoofing, if contrary, the measured locational inconsistency will be represented as uncertainty (the data generator consciously wants to amplify locational inconsistency in the measurement but fails). Thus, based upon underlying motivations, the proposed framework enables us to distinguish locational uncertainty, mistakes and spoofing. In practice, it also allows researchers to detect different types of locational inconsistencies from a dataset, thereby generating a more accurate subset for reuse. For example, geo-tagged Twitter feed enables researchers to track Oregon residents' opinions on specific topics (i.e., terrorist attacks, presidential campaign). Before drawing any conclusions from a set of geo-tagged tweets sent out from Oregon, researchers must delete those tweets that are falsely attributed to Oregon residents because of spoofing or mistakes, and then estimate the collective opinion using the rest of the tweets. More importantly, this framework transcends a purely technical interpretation by providing a holistic perspective from which to examine spatial data quality. In this sense, this framework encourages GIScientists to devote more effort to discovering the generative mechanism of various locational inconsistencies at the human intention level.

4. Concluding remark

In sum, I investigated the effects of human intention on spatial data. Through a detailed examination of various motivations associated with measurement, I suggested to modify the hypothesis of rational geographer and put spoofing and mistakes align with uncertainty under a greater umbrella topic - spatial data quality. I call on GIScientists to establish a human-centric assessment strategy to better understand this topic. Only then can we more effectively handle the emerging quality issues in the era of big data.

Acknowledgements

For their invaluable assistance with this article, I wish to thank Daniel Sui, Li He, Xining Yang, Peter Chen, and three anonymous reviewers.

References

Andev. (2013). Fake gps - fake location. Retrieved on May 4, 2016 from https://play.google.com/store/apps/details?id=com.fakegps.mock

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. *International Journal of Spatial Data Infrastructures Research, 4*(1), 332-358.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., & Shi, W. (2010). Thirty years of research on spatial data quality: Achievements, failures, and opportunities. *Transactions in GIS, 14*(4), 387-400.

Grant, A., Williams, P., Ward, N., & Basker, S. (2009). GPS jamming and the impact on maritime navigation. *Journal of Navigation, 62*(02), 173-187.

McCauley, D. J., Woods, P., Sullivan, B., Bergman, B., Jablonicky, C., Roan, A., . . . Worm, B. (2016). Ending hide and seek at sea. *science, 351*(6278), 1148-1150.

Wechsler, H., Schwartz, L. B., Ploscowe, M., & Tappan, P. W. (1962). *Model Penal Code*. Philadelphia, PA: ALI (American Law Institute).

Zhang, J., & Goodchild, M. F. (2002). *Uncertainty in geographical information*: CRC press.

Zhao, B. (2015). Detecting location spoofing in social media: Initial investigations of an emerging issue in geospatial big data. Ph.D. Disseration. The Ohio State University.

[i] MPC is highly influential across the United States for clarifying these various levels of culpability.

[ii] A defense can also be a reasonable excuse (e.g., infancy, insanity, involuntary intoxication, etc.) or justification (e.g., duress, necessity, self-defense, consent of the victim, entrapment, etc.). Since these intentions are far off the intention for a locational inconsistency, I did not discuss in details.

Semi-parametric Geographically Weighted Regression (S-GWR): a Case Study on Invasive Plant Species Distribution in Subtropical Nepal

Qunshan Zhao[1], Elizabeth A. Wentz[1], Stewart Fotheringham[1], Scott T. Yabiku[2], Sharon J. Hall[3], Jennifer E. Glick[2], Jie Dai[4], Michele Clark[3], Hannah Heavenrich[3]

[1]Center for Geographical Information Science, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ 85287, USA
Email: {qszhao; wentz; sfotheri}@asu.edu
[2]Department of Sociology and Criminology, Pennsylvania State University, University Park, PA 16802, USA
Email: {sty105; jeg115}@psu.edu
[3]School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA
Email: {sharonjhall; mdclar10; hheavenr}@asu.edu
[4]Department of Geography, San Diego State University, San Diego, CA 92182, USA
Email: jdai@rohan.sdsu.edu

Abstract

Geographically weighted regression (GWR) is a spatial statistical methodology to explore the impact of non-stationarity on the interaction between spatially measured dependent and independent variables. In this paper we use a semi-parametric geographically weighted regression (S-GWR) and demonstrate the effectiveness of the method on a case study on socio-ecological factors on forest vulnerability. The case study is based on community forests in and around the buffer zone of Chitwan National Park, Nepal, a biodiversity hotspot that is being rapidly degraded by exotic invasive plant species. This research integrated heterogeneous data sources such as observational ecological surveys, household interviews, and remotely sensed imagery. These data were utilized to extract and represent invasive plant species coverage, human activity intensity, topographical parameters and vegetation greenness indices. Research findings both demonstrate the S-GWR method and offer possible interventions that could slow the catastrophic spread of invasive plant species in Chitwan, Nepal.

1. Introduction

Geographically weighted regression (GWR) is a spatial analysis method that uses the spatial distribution of dependent and independent variables to specify non-stationarity to quantify the drivers of spatially dependent processes. GWR has been widely applied in application domains such as species distribution modeling in ecology and crime analysis in sociology (Foody 2004; Zhang and Song 2013). This paper presents a semi-parametric geographically weighted regression (S-GWR) method to model the factors that influence the spatial distribution of invasive plants in Nepal. We chose to use a semi-parametric model to include both parametric and non-parametric variables in the model specification. The S-GWR is implemented in a case study on the relationship between socio-ecological factors and invasive plant species in Chitwan, Nepal.

Invasive plant species are considered as a serious global environmental threat to ecosystem structure and function by creating disturbances in ecosystems, reducing native species diversity and abundance, limiting human usage of ecosystems and triggering environmental changes. It is an emergent research topic to prevent invasive plant species spread and alleviate their influences in forest ecosystems. The case study aims to quantify the relationships between invasive plant species coverage and socio-ecological factors in community forests (CFs) in Chitwan, Nepal. We

demonstrate the use of S-GWR to understand the spatial nonstationarity in invasive species distribution modeling (iSDM). With the social factors, the modeling results enable us to better understand how invasive species are influenced by human activities in the CFs, and how to balance the human activities and CF ecosystems by appropriate forest management.

2. Case study in Chitwan, Nepal

In forest ecosystems, community forest (CF) management has been recognized over the past three decades as a successful method for improving forest conservation methods (Klooster and Masera 2000). In Chitwan, Nepal, CFs were established with the intent of allowing citizens to manage forest resources while reducing negative impacts on the surrounding National Park. Even though the CF ecosystem has been extremely successful in the buffer zone of Chitwan National Park since the 1990s, by 2005 exotic plant invasion broke the current ecological equilibrium. Among all the invasive plant species found in Chitwan, *Mikania micrantha*, or "Mile-a-Minute weed", is considered the most harmful invasive species to ecosystem processes and requires an immediate forest management intervention in Nepal (Rai and Scarborough 2015). *M. micrantha* has the potential to catastrophically disrupt this urbanizing socio-ecological ecosystem in inequitable ways depending on CF vulnerability. Thus, the relationships between *M. micrantha* coverage, forest biophysical characteristics and human activities must be studied to design an effective intervention for the exotic plant invasion in the CF ecosystems at Chitwan, Nepal.

2.1 Study area

The study area focuses on the 12 continuous southern riverine CFs in western Chitwan district, which are in or nearby the buffer zone between the cultivated Chitwan valley and the Chitwan National Park at Nepal (see Figure 1).

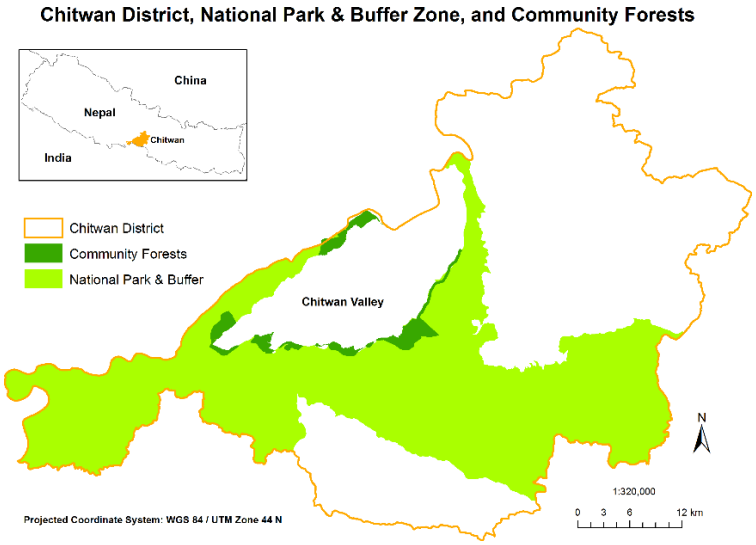


Figure 1. Study area in community forests at Chitwan National Park, Nepal.

2.2 Data sources

This research integrates numerous heterogeneous data sources such as observational ecological survey, household social survey, and remotely sensed imagery to understand the drivers of the *M. micrantha* invasion in this region. Invasive plant species and environmental surveys were conducted from 2013-2015 in the CFs. From the environmental survey, *M. micrantha* coverage,

canopy coverage, fire evidence and dominant vegetation type were observed and recorded. A household social survey collected information on household resource use activity information in the CFs, including the intensity of collecting firewood, fodder, thatch, and medicinal plants. A set of enhanced vegetation index (EVI) images were derived from Landsat imagery from 1988 to 2015 to represent the time series forest greenness conditions. Topographical parameters such as elevation and distance to river were extracted from ASTER DEM and Landsat image.

3. Semi-parametric geographically weighted regression

GWR captures locally varying processes to better understand the drivers of the spatial distribution of the dependent variable. The S-GWR, in contrast, has both geographically varying coefficients and fixed coefficients in the same model (Fotheringham *et al.* 2002; Nakaya *et al.* 2009). This enables S-GWR to model spatial stationarity and spatial nonstationarity in the same framework.

The S-GWR model can be presented as equation (1):

$$y_i = \sum_{j=1}^{k} \alpha_j x_{ij} + \sum_{l=k+1}^{p} \beta_l(u_i, v_i) x_{il} + \varepsilon_i \quad (1)$$

where (u_i, v_i) are coordinate locations for each location i, α_j denotes the global parameter estimates of fixed independent variables, and $\beta_l(u_i, v_i)$ denotes the local parameter estimates on each location i in space.

4. Results and discussion

GWR 4.09 software was used for the S-GWR analysis. The results of S-GWR are shown in Table 1. The S-GWR model contains 1 global variable and 5 local variables. Fire evidence has a negative relationship with *M. micrantha* coverage, which tells us fire is a method to temporarily remove the invasive plants. In the local parameter estimates, recent EVI values (average EVI from 2010-11) all have the positive relationship with *M. micrantha* coverage, which indicates that *M. micrantha* increases the forest greenness. The estimates of canopy coverage are all positive, but the parameter estimates vary across different CFs. Both positive and negative coefficients coexist for the past EVI parameter (average EVI from 1988-89), which represents the local properties for each CF. Although the household resource collection activities (firewood, fodder and thatch) positively influence the *M. micrantha* coverage in the ordinary least squares (OLS) regression analysis results, this variable is not statistically significant in the S-GWR results. This finding shows the weak influences from human activities comparing to biophysical factors. The S-GWR model recognizes an extreme local pattern through the small optimal bandwidth found by S-GWR.

Table 1. S-GWR model results.

Global coefficients					
Variable	Estimate		Standard error		*t*-value
Fire evidence	-6.106		2.615		-2.335
Local coefficients					
Parameter	Minimum	Lower quartile	Mean	Upper quartile	Maximum
Intercept	-62.269	-9.836	12.148	24.641	192.111

Canopy	0.027	0.088	0.144	0.180	0.391
Past EVI	-67.162	-24.450	-8.120	5.085	72.560
Recent EVI	5.583	36.658	58.209	78.728	130.590
Elevation	-1.433	-0.233	-0.156	-0.023	0.536

Mean adjusted R^2 = 0.404; Global AICc = 9325.022; S-GWR AICc = 9248.952; Optimal bandwidth size = 52.

5. Conclusions

This research applies S-GWR to capture the spatial stationarity and nonstationarity to model the factors that influence the spread of invasive plants. By using S-GWR method, spatial stationarity and nonstationarity effects are all incorporated into the same framework and a better model fit is achieved with higher R-squared and lower AICc. Human resource collection activities only exhibit statistically significant influence to *M. micrantha* coverage in OLS model rather than S-GWR model. A further examination needs to be conducted to better understand how human activities influence the spread of *M. micrantha* in Chitwan area.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1211498, Dynamics of Coupled Natural and Human Systems Program, and by the National Institute of Child Health and Human Development under grant 1R21HD073758. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and National Institute of Health. And the authors want to thank DigitalGlobe Foundation and European Space Agency for providing satellite imagery for this research, and the anonymous reviewers for their valuable comments and suggestions to improve the manuscript.

References

Foody G, 2004, Spatial nonstationarity and scale-dependency in the relationship between species richness and environmental determinants for the sub-Saharan endemic avifauna. *Global Ecology and Biogeography*, 13(4), 315–320.

Fotheringham A, Brunsdon C, and Charlton M, 2002, *Geographically weighted regression: the analysis of spatially varying relationships.* Chichester, England ; Hoboken, NJ, USA: Wiley.

Klooster D and Masera O, 2000, Community forest management in Mexico: carbon mitigation and biodiversity conservation through rural development. *Global Environmental Change*, 10(4), 259–272.

Nakaya T, Fotheringham A, Charlton M, and Brunsdon C, 2009, Semiparametric geographically weighted generalised linear modelling in GWR 4.0, 1–5.

Rai R and Scarborough H, 2015, Understanding the Effects of the Invasive Plants on Rural Forest-dependent Communities. *Small-Scale Forestry*, 14(1), 59–72.

Zhang H and Song W, 2013, Addressing issues of spatial spillover effects and non-stationarity in analysis of residential burglary crime. *GeoJournal*, 79(1), 89–102.

Accessing Distributed WFS Data Through A RDF Query Interface

Tian Zhao[1], Chuanrong Zhang[2], and Weidong Li[2]

[1] University of Wisconsin – Milwaukee
[2] University of Connecticut

Abstract. Geospatial data stored in databases and other formats can be accessed through Web Feature Service (WFS). However, it is not convenient to access data in multiple WFS servers since WFS protocol is geared towards single server. In this paper, we propose an algorithm to query and synthesize distributed WFS data through a RDF query interface, where users can specify data requests to multiple WFS servers using a single RDF query. The algorithm translates each RDF query written in SPARQL-like syntax to multiple WFS get-feature requests, and then convert the WFS results to answers to the original query. A lightweight Web-based prototype is implemented based on this approach.

1 Introduction

In this paper, we propose the design of a RDF query interface for distributed WFS data, which accepts queries in SPARQL-like syntax to provide more flexibility and usability than direct WFS queries.

As a motivational example, imagine a scenario where a developer needs to implement a program to display the flooded streets near the high schools of a city. The programmers can define an interface backed by some predefined WFS queries to fetch data from two servers. For each user request, the interface will request data from the servers and then integrate the results. If the servers use different data definitions, a translation step is needed to reconcile and integrate the data. If the developer needs to support another query such as finding the bridges of major highways, the developer has to perform the above steps again.

While the implementation of two queries shares many similarities, it is not apparent how to re-factor the duplicated code, which include the communication with WFS servers and the transformation and integration of the responses. The difficulty is not with building shared library but with composing specific WFS requests and interpreting and integrating the corresponding results, which may be different from each query. In addition, while queries may involve the same intermediate data, it is not straightforward to implement a caching strategy to improve performance. This is especially critical during emergency when peak data requests can overwhelm data servers.

Our design of RDF query interface aims to improve the productivity for rapid prototyping of WFS query applications. The RDF interface automatically

translates user queries formulated in a SPARQL-like syntax to WFS requests sent to multiple servers and then integrates WFS response to answer the original user queries. Using this design, application developers do not need to write code for WFS request and data processing. Instead, they can accomplish the same goal by defining mappings from WFS feature types to RDF definitions and by writing RDF queries.

Note that the data provided by WFS servers may be backed by databases or shapefiles. The cost of converting the data to a uniform format may not be feasible for large or frequently updated data sets.

2 Related Work

In literature, ontology has been used in search tools to help to discover geospatial web services related to certain domain concepts [2]. Tools have also been developed to convert geospatial ontology data to forms that can be accessed via WFS protocol [1]. Given the abundance of data available from geospatial web services and databases, a more interesting direction is to make data from geospatial web services and databases accessible via RDF protocols.

The closest study is the work of Tschirner et al. [3], who proposed a method to convert GML data into ontology data by translating SPARQL queries into WFS requests. Their approach maps a SPARQL query to a WFS request that returns a superset of intended results, transforms the WFS results into ontology data, and then applies the original SPARQL query to obtain the final answer. While this paper shares similar workflow, our approach is different in several ways. Firstly, we do not assume the WFS data is centralized in one server or having a unified definition. This requires the translated WFS requests be separated for each feature type and the final join is done at the client side. We generate multiple OGC filter encoding for each SPARQL query. Secondly, our approach implements a light-weight Web client with roughly 1000 lines of JavaScript without library dependencies for query processing, which is easier for deployment. Lastly, our approach is designed to use a SPARQL-like syntax to bring more convenient query interface to WFS services while Tschirner et al. provide a more complete service of SPARQL endpoint for GML data.

In our prior work [4], we proposed a query rewriting algorithm to translate SPARQL query to WFS requests and database queries using idealized syntax. This paper extends that approach by considering a more realistic subset of SPARQL syntax and by implementing a Web-based prototype that incorporates caching optimization and data rendering.

3 Translate RDF Query to WFS Requests

Typically, a SPARQL query is translated to multiple WFS requests by grouping triples related to the same feature type together so that there is one WFS request per feature type. Any remaining triples are translated to spatial joins to be applied to the results of the WFS requests.

As a concrete example, the below query Q1 is an RDF query in a SPARQL-like syntax for retrieving the streets nearby each high school in New Haven, CT, where a geometry is *nearby* another one if their distance is less than 500 meters (this distance is arbitrarily chosen and can be modified).

```
select  ?s      ?p        where                              (Q1)
        ?s      rdf:type        streets.
        ?p      nh:category      "High School".
        ?s      nearby           ?p
```

In the query, identifiers starting with *?* are variables. The solutions to the variables *?s* and *?p* (which stand for streets and points respectively) between *select* and *where* are the intended results of the query. The lines after *where* are called triples that specify the conditions with which the variable solutions must satisfy. Each triple has the form of *subject predicate object*, which restricts the relation between the subject and object with the predicate. For example, the triple *?s rdf:type streets* says that the solution to ?s must has a type called *streets*. The triple *?p nh:category "High School"* specifies that the category of *?p* is *High School*. Finally, the triple *?s nearby ?p* relates the spatial attributes of *?s* to those of *?p* by the distances between them. Note that users can change the RDF definitions by editing a configuration file.

Pre-processing The pre-processing phase separates the triples into four groups based on the triple predicates. For example, the triples in Query Q1 can be separated as below, where *?p nh:category ?c* and *?c == "High School"* are automatically generated from *?p nh:category "High School"*.

type triple	*?s rdf:type streets.*
property triple	*?p nh:category ?c.*
filter triple	*?c == "High School".*
spatial join triple	*?s nearby ?p*

Query rewriting Based on the separated triples, the rewriting phase includes five steps: (1) identify the set of feature variables that correspond to feature types; (2) find the set of triples related to each feature variable; (3) find the set of feature types for each feature variable; (4) find the set of filter expressions for each pair of feature variable and its type; (5) construct a set of WFS get-feature requests for each feature variable.

For the query Q1, the feature variables are ?s and ?p and their feature types are `new_haven_streets` and `new_haven_places` respectively. The following get-feature requests can be generated.

`getFeature`(*new_haven_streets*)
`getFeature`(*new_haven_places*, CATEGORY *PropertyIsEqualTo* 'High School')

Post-processing After receiving the responses from the get-feature requests, we perform spatial joins on the retrieved geospatial features if necessary. The results of the spatial join are filtered by the types of the features identified from the

selection variables of the SPARQL query. The features retrieved from the get-feature requests are cached, which greatly improves performance by avoiding network and server overhead.

Implementation Figure 1 is a screen shot of our prototype, which is available at `boyang.cs.uwm.edu:8080/newHaven/bing.html`. It parses each SPARQL query in text form and generates a set of get-feature requests, which are sent as AJAX calls to WFS servers and the responses are joined before being displayed on a map. There are some pre-defined queries though they are not hard-wired in any way and users can revise the query in the textbox. The query interface actions are logged below the textbox so that user can track the query progress.

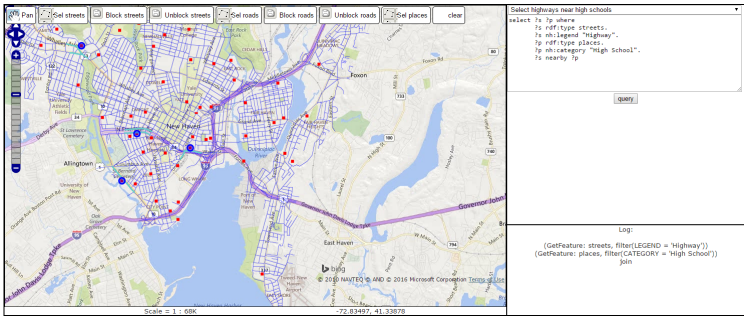


Fig. 1. Highways near high schools

4 Conclusion

In this paper, we present an algorithm to convert RDF queries to WFS requests so that users can query distributed WFS features as if they were RDF instances. The algorithm avoids the cost of converting features to RDF objects while retaining the benefits of RDF queries. As future work, we will extend the algorithm to include static checking capability to detect semantic errors before runtime.

References

1. J. Jones, W. Kuhn, C. Keler, and S. Scheider. Making the web of data available via web feature services. In *AGILE 2014*, 2014.
2. W. Li, C. Yang, D. Nebert, R. Raskinc, P. Houser, H. Wu, and Li Z. A semantic-based web service discovery and chaining for building an arctic spatial data infrastructure. *Computers & Geosciences*, 37:1752–1762, 2011.
3. Sven Tschirner, Ansgar Scherp, and Steffen Staab. Semantic access to inspire how to publish and query advanced gml data. In *Workshop in Conjunction of 10th International Semantic Web Conference*, 2011.
4. T. Zhao, C. Zhang, M. Wei, and Z.-R Peng. Ontology-based geospatial data query and integration. In *Lecture Notes in Computer Science LNCS5266: Geographic Information Science, 5266*, pages 370–392, 2008.

Predicting Influenza Dynamics using a Deep Learning Approach

Shiran Zhong and Ling Bian

University at Buffalo, the State University of New York, 105 Wilkson Quad, Buffalo, NY 14261
Email: shiranzh@buffalo.edu, lbian@buffalo.edu

Abstract

Disease transmission is a complex spatio-temporal process. A great number of approaches have been developed to predict influenza epidemics. Few of them have focused on the temporal dynamics of individual infected locations. Location networks, where locations are nodes and disease flows between them are links, provide a promising approach for such dynamic analyses, but also present challenges. In this study, we employ a deep learning approach to capture the dynamics of disease flows in location networks. We also analyze how the attributes of locations have an impact on the prediction accuracy via a sensitivity analysis.

1. Introduction

Disease transmission is a complex spatio-temporal process (Ferguson et al., 2005; Charaudeau et al. 2014). Among the prevailing approaches to predicting the dynamics of influenza epidemics, few have focused on the transmission at a location-specific scale. Location networks, where locations are nodes and disease flows between them are links, provide a promising basis for such dynamic analyses (Zhong and Bian 2016), but also present challenges as each location might have peaks and troughs of different magnitudes throughout the epidemic (Bian et al. 2012). Conventional approaches are not adequate to capture such complex, dynamic patterns.

The objectives of this study are two-fold. We explore the use of Deep Convolutional Networks (DCN), a deep learning approach, to capture and predict disease flow dynamics represented by the presence of links between locations. We also analyze how the attributes of locations impact the prediction accuracy via a sensitivity analysis.

2. Data and Study Area

The dataset consists of 73 daily disease flow networks of an urban area. In these networks, all 1,026 location nodes remain the same across 73 days. Links can be present or absent depending on the occurrence of disease flow between locations on any particular day. The dataset was obtained from the China Information System for Diseases Control and Prevention.

3. Methodology

To achieve the objectives stated above, the methodology is divided into three parts: the first part describes the principle of DCN; the second part introduces the training and testing processes involved in the DCN; and the last part focuses on the sensitivity analysis.

3.1. Principle of DCN

In contrast to classic neural networks which require good features for supervised training, DCN is a training process where good features could be automatically learned from input data (LeCun et al. 2015). The workflow of DCN in this study (Figure 1) contains three processes: 1) the convolution process, where a convolution filter is applied on the input data (in a format of matrix) in order to amplify the feature signal and suppress the noise; 2) the pooling process, which extracts features that represent location characteristics in the past few days; and 3) the training and testing process, where the learned features are used as input to predict the presence

and absence of links between locations as output (Busseti et al. 2012; LeCun et al. 2015). Predictions are made from past observations and validated with current observations, The process is repeated as the time window moves forward. Of the three processes, training and testing process is further explained below.

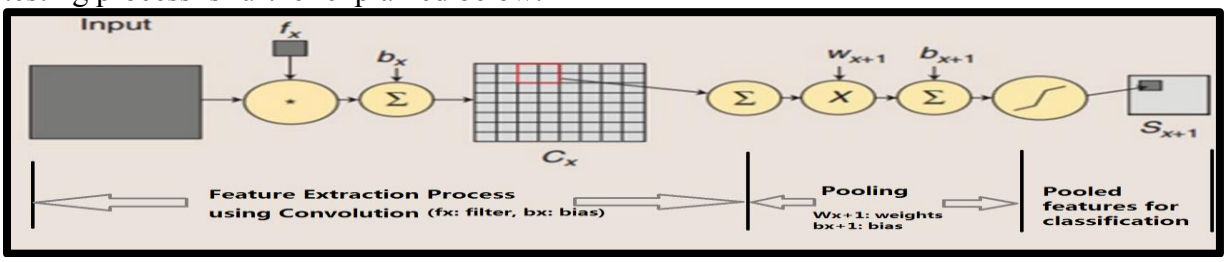


Figure 1. Schematic view of deep convolutional network workflow.

3.2. Training and testing process

The features are represented by 13 attributes of locations in three categories. The first category of attributes describes the epidemiological behavior of locations, such as number of cases at locations, and the previous presence of disease flow between locations. The second category describes the characteristics of location nodes in the networks. These include path length between location pairs, degree, betweenness, closeness, clustering coefficient, eccentricity, bridge, radiality, stress and topological coefficient. The third category takes spatial information into consideration by counting the number of cases at locations' nearest neighbors.

In the training process, we use **DCN** to build an optimized weight matrix $\boldsymbol{W}$, which represents how the presence of link on a certain day is associated with its connecting nodes' attributes in the past few days. The weight matrix W is trained according to Equation 1:

$$\text{Output } (Y_{ijm}) \Leftarrow \text{Input } (X_{ijm-1}, X_{ijm-2}, \ldots X_{ijm-n}) * \boldsymbol{W} \qquad \boldsymbol{(1)},$$

where $\boldsymbol{Y_{ijm}}$ represents the presence/absence of the link between Locations $\boldsymbol{i}$ and $\boldsymbol{j}$ observed on Day $\boldsymbol{m}$; $\boldsymbol{X_{ijm}}$ indicates the input attributes on Locations $\boldsymbol{i}$ and $\boldsymbol{j}$ observed on Day $\boldsymbol{m}$. $\boldsymbol{X_{ijm-1}, X_{ijm-2}, \ldots X_{ijm-n}}$ represent attributes of Locations $\boldsymbol{i}$ and $\boldsymbol{j}$ observed one day, two days, up to n days prior to Day m, respectively. $\boldsymbol{n}$ is the temporal lag, which is initially set as five days. The 73 daily networks are divided into two parts, a training set and a testing set. The training set contains the daily network of the first 50 days, and the remaining 23 days are used as the testing set.

3.3. Sensitivity Analysis

We perform a sensitivity analysis to evaluate the impact of the 13 attributes and the training parameters, i.e. the temporal lag and the length of training/testing sets, on the prediction. The impact of the 13 attributes is evaluated by removing them in two approaches: 1) a conventional approach in which one attribute is removed at a time, yielding 13 scenarios, and 2) an alternative approach in which the exhaustive combination of multiple attributes are removed, yielding 2^13=8,192 scenarios. Regarding the impact of training parameters, the temporal lag $\boldsymbol{n}$ is varied from 3-10 days and the training set is lengthened to 50 to 59 days, while the testing set is shortened accordingly.

Four criteria are utilized to evaluate the accuracy of predicted presence and absence of links: 1) Overall Accuracy (OA): the sum of correctly predicted presence and absence divided by the sum of observed presence and absence over the testing period, 2) Precision of Presence (PP): the correctly predicted presence divided by the total number of predicted presence over the testing period, 3) Precision of Absence (PA): the correctly predicted absence divided by the total number of predicted absence over the testing period, and 4) F1 Score: the balance between PP

and PA as shown in Equation 3(Powers 2011). All accuracy measurements are standardized from 0 to 1.

$$F1\ score = \frac{(2 * PP * PA)}{(PP + PA)} \quad (3)$$

For comparison purposes, the classic neural networks analysis (FNN) is also applied to the same 73 daily networks to predict the present/absence of links, using the same set of attributes, temporal lag, training and testing length division, the sensitivity analysis, and the four accuracy criteria.

4. Results and Discussion

Figure 2 illustrates the prediction results using DCN (a) and classic neural networks (b), with respect to the varying temporal lag and training and testing length division. For the DCN results, OA and PA are above 92%. PP and F1 Score are above 80% when the temporal lag is within six days. As the observed peaks and troughs of the epidemic last approximately 12-14 days, the 6-day half cycle corresponds to the rising slope of the epidemic peak before it declines. The same principle is applicable to the troughs. The training and testing length division does not have a noticeable impact on the prediction results. In addition, the DCN produces results with considerably higher accuracy than that generated by classic neural networks in terms of all four criteria (Figure b).

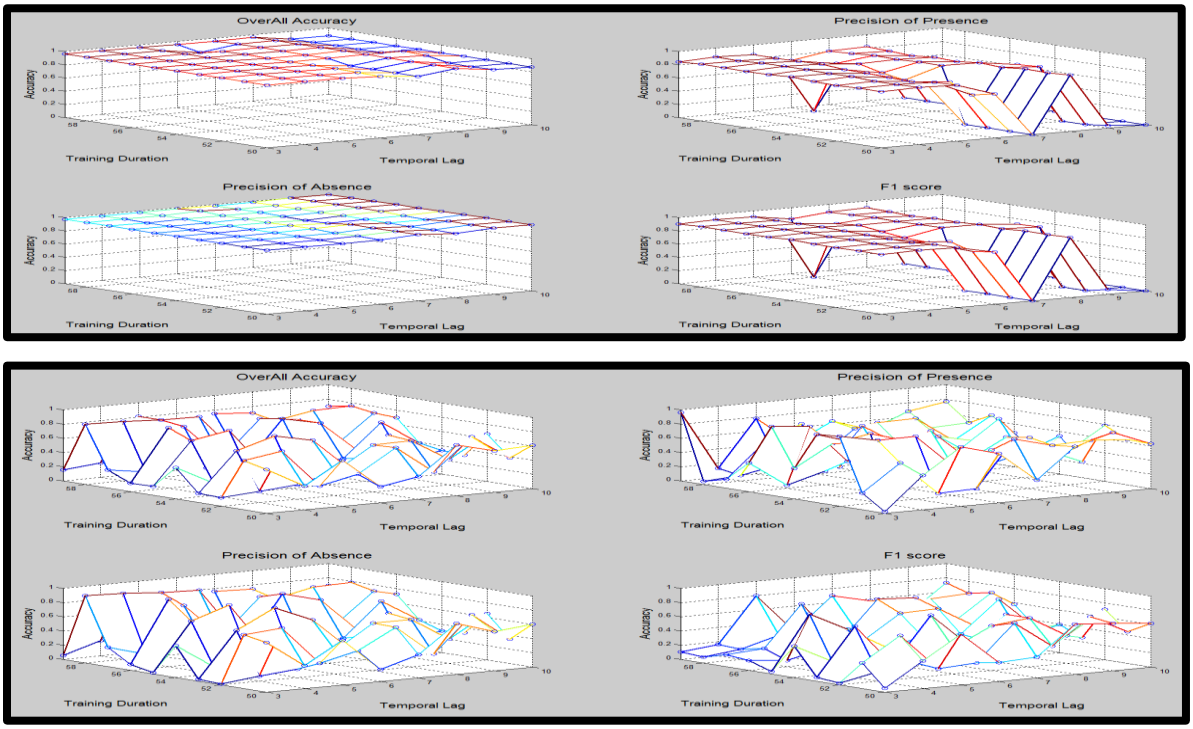


Figure 2. Prediction results using DCN (a) and classic neural networks analysis (b). Each sub figure corresponds to one of the four evaluation criteria: OA, upper left; PP, upper right; PA, lower left, and F1 score, lower right (x-axis: temporal lag; y-axis: length of training period; vertical axis: prediction accuracy).

Figure 3 illustrates the results from the sensitivity analysis regarding the impact of attributes on prediction results, with the F1 score reported as a balanced evaluation. Among the

13 scenarios yielded by removing one attribute at a time, The prediction is most sensitive to the removal of network path length. This attribute measures the length of disease flow pathways.

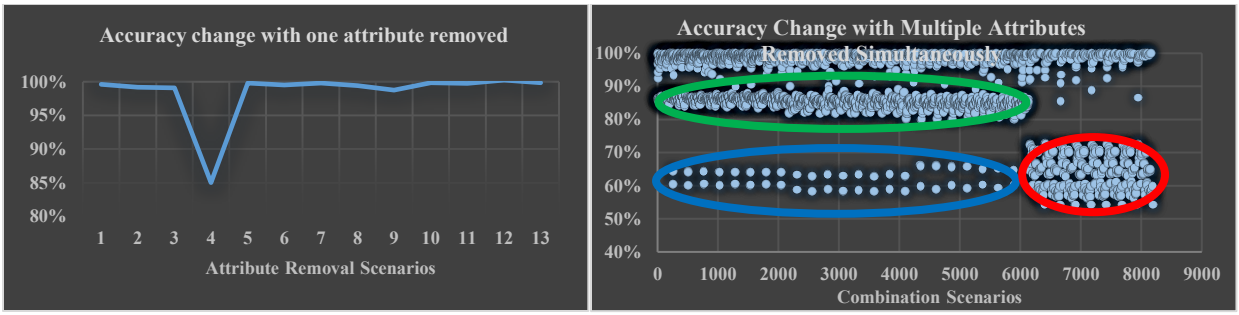


Figure 3. Sensitivity analysis results with respect to the 13 attributes: a) 13 scenarios with one attribute removed at a time and b) 8,192 scenarios with multiple attributes removed at a time (x-axis: scenario ID; y-axis: prediction accuracy).

Prediction results from the 8,192 scenarios fall into three groups (Figure 3b). The green circle highlights the scenarios whose accuracy is reduced by 17.75% on average, when two attributes, the previous presence of disease flow between locations and the path length, are removed. The blue circle highlights scenarios whose accuracy has decreased by 39.5% on average, when two attributes, degree and closeness, are removed. The red circle highlights scenarios whose accuracy has decreased by 40.87% on average, with two attributes, neighbors' cases and eccentricity, are removed. These six attributes represent the source and pathways of disease flow. The prediction of links is sensitive to these attributes. These findings help develop location-oriented intervention strategies to mitigate the spread of disease, e.g. quarantine at location pairs of short path length.

Acknowledgements

Research reported in this publication was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM108731. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Reference

Bian, L., Huang, Y., Mao, L., Lim, E., Lee, G., Yang, Y., ... & Wilson, D. (2012). Modeling individual vulnerability to communicable diseases: A framework and design. *Annals of the Association of American Geographers*,*102*(5), 1016-1025.

Busseti, E., Osband, I., & Wong, S. (2012). *Deep learning for time series modeling*. Technical report, Stanford University.

Charaudeau, S., Pakdaman, K., & Boëlle, P. Y. (2014). Commuter Mobility and the Spread of Infectious Diseases: Application to Influenza in France. *PloS one*, *9*(1), e83002.

Ferguson, N. M., Cummings, D. A., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic. *Nature*,*442*(7101), 448-452.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*,*521*(7553), 436-444.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Zhong, S., & Bian, L. (2016). A Location-Centric Network Approach to Analyzing Epidemic Dynamics. *Annals of the American Association of Geographers*, 1-9.

Which Kobani? A Case Study on the Role of Spatial Statistics and Semantics for Coreference Resolution Across Gazetteers

Rui Zhu, Krzysztof Janowicz, Bo Yan, and Yingjie Hu

STKO Lab, Department of Geography, University of California, Santa Barbara, USA
{ruizhu,jano,boyan,yingjiehu}@geog.ucsb.edu

Abstract

Identifying the same places across different gazetteers is a key prerequisite for spatial data conflation and interlinkage. Conventional approaches mostly rely on combining spatial distance with string matching and structural similarity measures, while ignoring relations among places and the semantics of place *types*. In this work, we propose to use spatial statistics to mine *semantic signatures* for place types and use these signatures for coreference resolution, i.e., to determine whether records form different gazetteers refer to the same place. We implement 27 statistical features for computing these signatures and apply them to the type and entity levels to determine the corresponding places between two gazetteers, which are GeoNames and DBpedia. The city of Kobani, Syria, is used as a running example to demonstrate the feasibility of our approach. The experimental results show that the proposed signatures have the potential to improve the performance of coreference resolution.

Keywords: Spatial statistics, coreference resolution, gazetteers, semantic signatures

1 Introduction and Motivation

Coreference resolution across gazetteers is an important prerequisite for spatial data conflation and interlinkage. Conventional approaches, such as coordinate matching, string matching, and feature type matching, often focus on the footprints, names, and types of places, as well as the combination of these three properties (Sehgal et al., 2006; Shvaiko and Euzenat, 2013). However, such approaches have their limitations. Today, most gazetteers still rely on centroids for representing geographic features (even for feature types such as counties, rivers, or oceans). These centroids differ significantly across datasets, often by more than 100km. Furthermore, it is difficult to select a place type agnostic distance threshold as initial search radius. Polygon and polyline based matching, e.g., using Hausdorff distance, comes with its own limitations, scale and the resulting generalization being key problems. For string matching, such as using Levenshtein distance, the same place may have substantially different toponyms (e.g., Ayn al-Arab in TGN and Kobani in DBpedia) while different places may share common names. In addition, simply relying on direct feature type matching is likely to fail since different gazetteers employ incompatible typing schemata/ontologies. In conjunction, these problems often lead to either false negative or false positive matches.

In previous work, we proposed using *spatial signatures*, which are derived from spatial statistics, to understand the semantics of places types bottom-up (Zhu et al., 2016). In this work, we apply these signatures to coreference resolution. The used spatial statistics are selected from three perspectives; a detailed list is shown in Table 1:

- **Spatial point pattern analysis**. Point coordinates are used to quantitatively measure the spatial point patterns of place types (such as *populated place*). Kernel density estimation, Ripley's K, and standard deviational ellipse analysis are conducted and corresponding statistics are obtained for representing the signatures. Furthermore, we computed these statistics from both local and global aspects.
- **Spatial autocorrelation analysis**. In order to capture the interaction between places, we converted the point patterns into raster maps where each pixel represents the intensity of points. Spatial correlation statistics, such as Moran's I and semivariograms, are subsequently used to improve the signatures.
- **Spatial interactions with other geographic features**. In contrast to the first two perspectives, this group of statistical features is derived by integrating other geographic features. These

geographic features are further separated into internal features, in which the target feature's neighbors are considered, and external features, in which external data sources such as populations and road networks are incorporated.

A feature type's statistics (e.g., *Mountain*) are calculated considering all the points across the continuous US that belong to said feature type. To tackle the scalability issue of conducting spatial point analysis, like the Ripley's K, a spatial sampling technique was introduced (Zhu et al., 2016). Furthermore, in order to analyze the spatial autocorrelation of one feature type's intensity, the points are converted to a raster map whose cell sizes are consistent across all feature types. These statistics together (see Table 1), work as the spatial signature (here feature vectors) for the target feature type. Note that these 27 statistics are regarded as equally weighted in our work but more sophisticated models can be investigated in the future work.

Table 1: A summary of the 27 statistical features used to derive place type signatures.

<table>
<tr><th colspan="2">Spatial Point Patterns</th><th>Spatial Autocorrelations</th><th colspan="2">Spatial Interactions with Other Geographic Features</th></tr>
<tr><td rowspan="11">Local</td><td>Intensity</td><td rowspan="4">Global Moran' I</td><td rowspan="4">Internal</td><td rowspan="2">Count of distinct nearest feature types</td></tr>
<tr><td>Mean distance to nearest neighbor</td></tr>
<tr><td>Variance distance to nearest neighbor</td><td rowspan="2">Entropy of nearest feature types</td></tr>
<tr><td>Kernel density (bandwidth)</td></tr>
<tr><td>Kernel density (range)</td><td rowspan="3">Semivariogram value (at first distance lag)</td><td rowspan="8">External</td><td rowspan="2">Population value (min)</td></tr>
<tr><td>Ripley's K (range)</td></tr>
<tr><td>Ripley's K (mean deviation)</td><td>Population value (max)</td></tr>
<tr><td>Standard deviation ellipse (rotation)</td><td rowspan="3">Semivariogram value (at median distance lag)</td><td>Population value (mean)</td></tr>
<tr><td>Standard deviation ellipse (std dev along x-axis)</td><td>Population value (std dev)</td></tr>
<tr><td>Standard deviation ellipse (std dev along y-axis)</td><td>Shortest distance to road (min)</td></tr>
<tr><td rowspan="3">Global</td><td>Intensity</td><td rowspan="3">Semivariogram value (at last distance lag)</td><td>Shortest distance to road (max)</td></tr>
<tr><td>Kernel density (bandwidth)</td><td>Shortest distance to road (mean)</td></tr>
<tr><td>Kernel density (range)</td><td>Shortest distance to road (std dev)</td></tr>
</table>

2 Method and Case Studies

We have computed a unique semantic signature for each place type defined in the used gazetteers based on the statistical features outlined above. Next, these signatures are used within two case studies to show how they can be applied to improve coreference resolution. In the first case, we will use the signatures as an additional matching score to complement spatial distance as well string and type label similarity used in the existing literature. In the second case we also consider the signatures of neighboring places.

To illustrate our method, Kobani, Syria, is used as an example. Kobani is a city near the border between Syria and Turkey. It is a typical example for the complexities arising when multiple parties such as the local population, news outlets around the world, government agencies from different states, and so forth, refer to a place by different names such as Aarab Peunar, Kubani, Kobane and 'Ayn al' Arab, to name but a few. Different and overlapping sets of toponyms are stored in gazetteers making straightforwards string similarity matching challenging. To further complicate issues, different places may have similar or even the same names while being in relative vicinity, such as a river's centroid that shares its name with a populated place. As we outlined above, the introduction of feature types into the matching process is often of limited use as there is no commonly agreed geographic feature type ontology and the individual ontologies and vocabularies used by gazetteers are under-specific to a degree where one has to rely on the types labels (and therefore string matching).

Figure 1 illustrates search results for Kobani in two gazetteers, GeoNames and DBpedia Places. Since the records of DBpedia are extracted from Wikipedia articles, only significant places are included. Therefore, the result for Kobani in DBpedia leads users directly to the city Kobani in Syria. However, since GeoNames is a more comprehensive gazetteer that attempts to collect all geographic features in the world, the search result for Kobani includes multiple records. Intuitively, and compared to types such as *stream* and *ruins*, the *seat of second-order administrative division* feature types in GeoNames is more likely to correspond to the *populated place* feature type in DBpedia despite having a very low string similarity between their labels. This make the matching task easy for humans but difficult for an automated machine-based matching.

2.1 Place Type Signatures as Additional Matching Characteristic

So far, we established the argument that toponym and centroid distance based matching on its own is not always sufficient and that place types (present in any modern digital gazettes) should be used in addition.

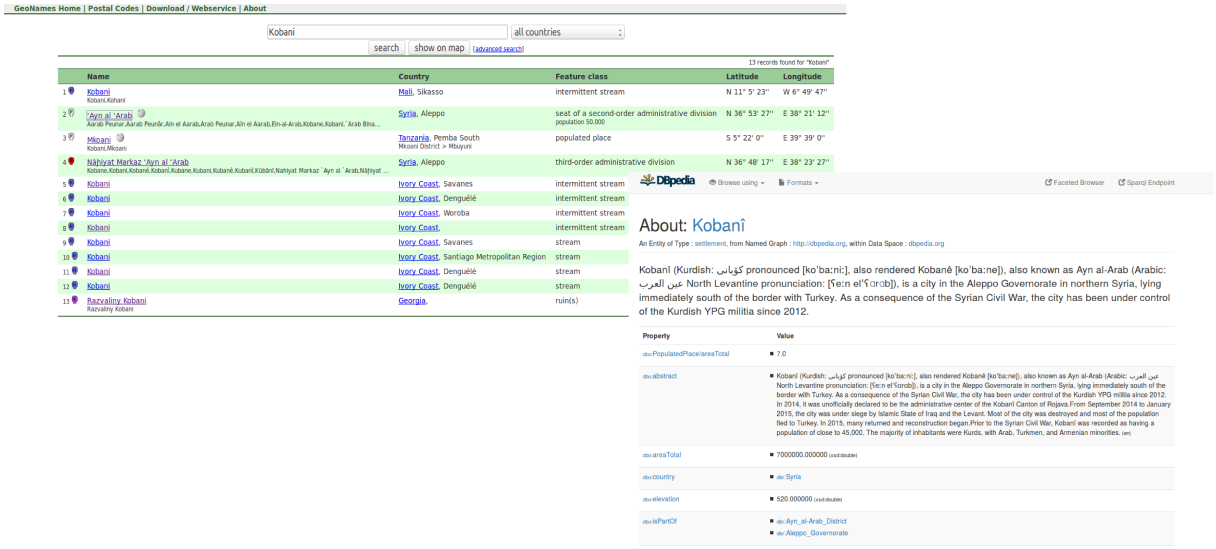


Figure 1: Results of searching for 'Kobani' in GeoNames (left) and DBpedia (right).

However, we also pointed out that given today's gazetteer ontologies/vocabularies the comparison of place types often boils down to string matching or simple structural measures as the used ontologies rely on a lightweight axiomatization (if any) and thus are not readily alignable. Here we propose to use the mined place type signatures as an additional matching characteristic that communicates the semantics of place types beyond labels alone. To give an intuitive example, we will make use of the fact that the spatial distribution of places of type *populated place* differs substantially from those computed for *river* but not from those computed for *seat of second-order administrative division* irrespective of the fact that those places are in different gazetteers and that the type labels show no similarity (Zhu et al., 2016).

As illustrated in Table 2, there are three place types associated with candidates for Kobani in GeoNames and one place type (*populated place*) in DBpedia. Computing the Euclidean distance between these three GeoNames place signatures (essentially feature vectors comprised of the 27 statistics listed in Table 1) and the *populated place* signature from DBpedia shows that *geonames: seat of second-order administrative division* and *dbpedia: populated place* should indeed be matched. This matching score would then be combined with the other classical matchers such as toponyms, spatial (centroid) distance, and so forth; thereby improving coreference resolution.

Table 2: Dissimilarities between the *populated place* signature for DBpedia and three example place type signatures in GeoNames.

Dissimilarity (Euclidean distance) with DBpedia: *Populated Place*	
GeoNames: *seat of second-order administrative division*	7.22
GeoNames: *stream/intermittent stream*	8.96
GeoNames: *populated place*	9.22

It should be noted that the place type *populated place* in GeoNames shares the same name with the DBpedia type. However, there is a substantial dissimilarity between their signatures. Such an observation indicates that despite a high string similarity, two place types might still have a different underlying semantics and we were able to show exactly this in previous work (Zhu et al., 2016). This further confirms that using string matching in isolation, either for the feature names or types, is not necessarily sufficient for coreference resolution.

2.2 Place Type Signatures of Neighboring Places

One drawback of the approach outlined thus far is that if multiple candidates shared the same place type, the signatures are incapable of providing any further distinctions. Therefore, we propose to include the signatures of neighboring places as well. To the best of our knowledge, neighboring places have not been proposed as part of any existing matching framework.

For our running example, we queried the 9 nearest neighbors for each candidate place and recorded their place types. This can be done directly using the precomputed *nearbyFeature* RDF predicate in GeoNames. Next, the averaged signatures (i.e., the averaged feature vector of the 27 statistics listed in Table 1) of these 9 place types are calculated for characterizing the neighborhood of the specific candidates. Lastly, candidates that have the smallest Euclidean distance in terms of their averaged

neighboring signatures are regarded as corresponding places. Since neighbors differ from candidate to candidate, their averaged neighboring signatures also differ despite potentially having the same place type. For instance, one populated place of a given name will have places of different types (e.g., a river and an island) nearby while another place of the same type (and similar names) will have places of other types (e.g., a mountain) nearby.

We tested this neighborhood based approach using our Kobani running example. Table 3 lists the place types of nearest neighbors for three example candidates in GeoNames and the Kobani from DBpedia. As can be seen, 'Ayn al 'Arab in GeoNames and Kobani in DBpedia both have relatively more diverse neighbors compared to the other two candidates. Table 4 shows the dissimilarity values which lead to the same conclusion (and correct matching) made in the direct matching case above and further supports our proposed approach.

Table 3: List of place types of the nearest neighbors for example candidates in GeoNames and DBpedia.

	'Ayn al' Arab (GeoNames: seat of a second-order administrative division)	Kobani (GeoNames: stream)	Mkoani (GeoNames: populated place)	Kobani (DBpedia: populated place)
Feature types of 9 nearest neighbors	section of populated place	populated place	populated place	settlement
	office building	stream	populated place	settlement
	school	stream	populated place	village
	square	stream	third-order administrative division	settlement
	prison	stream	third-order administrative division	tunnel
	section of populated place	stream	third-order administrative division	settlement
	section of populated place	stream	populated place	village
	market	stream	populated place	dam
	square	populated place	populated place	populated place

Table 4: Dissimilarities between *Kobani*'s neighboring signatures in DBpedia and the three example places' neighboring signatures in GeoNames.

Dissimilarity (Euclidean distance) with Kobani (DBpedia)	
'Ayn al' Arab (GeoNames)	4.23
Kobani (GeoNames)	10.57
Mkoani (GeoNames)	6.98

3 Conclusions and Future work

In this work, we presented an initial case study that demonstrates how semantic signatures mined from spatial statistics can reveal additional information about the semantics of place types on top of relying on type labels alone. Our work shows how spatial statistics and ontology engineering and alignment can go hand in hand to provide additional characteristics for tasks such as coreference resolution which play an increasingly important role as drivers of record linkage and conflation. In essence, we make use of the fact that different *types* of places can be told apart by the results of various spatial statistics performed over their instances, i.e., particular places. This, in turn, enables us to regard the resulting place type specific signatures as feature vectors and compute their dissimilarity using Euclidean distance (or other measures), thereby gaining an additional matcher on top of the string, spatial distance, and structural matchers used in the literature. Finally, we also go beyond existing work by taking neighboring places into account to improve the matching, instead of comparing 1:1 matches in isolation. In the future, we will apply the presented work to more (Linked Data) gazetteers and all their geographic features.

References

V. Sehgal, L. Getoor, and P. D. Viechnicki. Entity resolution in geospatial data integration. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 83–90. ACM, 2006.

P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):158–176, 2013.

R. Zhu, Y. Hu, K. Janowicz, and G. McKenzie. Spatial signatures for geographic feature types: Examining gazetteer ontologies using spatial statistics. *Transactions in GIS*, 2016.

Building Consistent Multi-temporal Population Data at Fine Resolution through Spatially Refined Areal Interpolation

H. Zoraghein[1], S. Leyk[1]

[1]Department of Geography, University of Colorado Boulder, CO, 80309, USA
Email: {hamidreza.zoraghein; stefan.leyk}@colorado.edu

Abstract

Demographic data are aggregated over areal units to protect privacy and are often inconsistent over time. Areal interpolation methods are used to estimate population in one census year within the units of another year to construct temporally consistent small census units. This research enhances these methods by using three advanced spatial refinement approaches, tested in Mecklenburg County, North Carolina to estimate population in 2000 within census tracts from the 2010 census. The results demonstrate the effectiveness of spatial refinement in reducing estimation errors, systematically. The proposed methods can be used to analyze micro-scale spatio-temporal demographic processes with minimum estimation error.

1. Introduction

Spatial analysis on demographic data aggregated over incompatible boundaries represents a challenge, particularly when the data were collected over historical inconsistent units. To understand micro-scale spatio-temporal demographic processes, data need to be collected over temporally consistent fine-resolution census geographies such as census tracts. However, in reality, their boundaries change over time due to population fluctuations, especially in rapidly growing areas.

Areal interpolation transfers the variable of interest from source zones to target zones and is used in temporal demographic applications (Gregory 2002; Schroeder 2007). In such applications, source populations in one census year (enumerated in source zones) are estimated within enumeration units from the target census completed in another year (target zones).

If the underlying assumptions of areal interpolation methods are not met, accuracy can be very low. Therefore, recent studies have developed spatially refined interpolation techniques with the objective of decreasing population estimation errors (Buttenfield *et al.* 2015; Ruther *et al.* 2015).

Areal interpolation methods are based on population density and area calculations. It can be expected that if these methods are constrained to spatially refined inhabited sub-areas of source and target zones, area and population density estimates will be more precise and realistic. Commonly, the spatially refined sub-areas are delineated using ancillary variables presumably related to population distribution in a dasymetric mapping approach (e.g., Mennis 2003).

This research extends the previous efforts, leveraging three advanced spatial refinement strategies to estimate total population enumerated in census tracts in 2000 within census tract boundaries used in the 2010 census.

2. Study Area and Data

The study area spans Mecklenburg County, North Carolina. It includes both urban areas of Charlotte at its center and large rural areas at its margins and has a history of rapid population growth.

Primary datasets include census tracts and census blocks from the 2000 and 2010 decennial censuses. Residential parcels, NLCD 2001 and 2011 and TIGER/Line data for road networks are used as ancillary datasets for spatial refinement.

3. Methods

3.1 Unrefined Areal Interpolation

Target Density Weighting (TDW) as a versatile areal interpolation method is included in this research and assumes the ratios of population densities of atoms (intersections of source and target zones) to source zones remain the same over time (Schroeder 2007).

3.2 First Spatial Refinement

The first strategy applies TDW to only refined sub-areas of source and target zones delineated by residential parcels as the ancillary variable (Zoraghein *et al.* 2016). The built-year attribute that records when the main structure of a parcel was built is used to match parcels with the census year.

3.3 Second Spatial Refinement

In addition to the geometric footprints of residential parcels, the second refinement uses their housing type to cap or amplify population density within different residential zones. For example, the population density of parcels of type apartments is higher than parcels with single-family residences, and this inherent diversity is addressed in this strategy.

Expectation Maximization (EM) is an iterative statistical optimization technique (Dempster *et al.* 1977), used for the second strategy. All the residential parcels of the same type (e.g., condominium) form control zones, and population density for each zone is estimated through EM. Some control zones include parcels with high variability in area. Thus, assuming one population density value for these zones is unrealistic. Therefore, "Enhanced EM" (EEM) is applied to address this issue.

EEM first identifies the three control zones that represent the highest variability in parcel area measures and the three control zones with the highest number of parcels. Each of these control zones are divided into four homogeneous control sub-zones using a quantile classification scheme for parcel area. For example, instead of using only one single-family residential control zone, four sub-zones of that type are included in EEM. The remaining steps are the same as EM.

3.4. Third Spatial Refinement

The third strategy is not confined to only residential parcels. It leverages additional complementary ancillary variables such as NLCD developed classes (21, 22, and 23) and road network buffer zones (100m buffer distance). The NLCD class selection follows Ruther *et al.* (2015) for delineating refined areas.

This methodology refines initial residential parcels as follows: if a parcel contains instances of the developed NLCD classes, only those instances are used for spatial refinement. However, if no developed land exists, the intersection area of the parcel with road buffers is used to spatially refine the parcel.

This refinement specifically targets rural settings, where large residential parcels overestimate residential areas while NLCD underestimates developed land as a well-known limitation of such databases (e.g., Leyk *et al.* 2014).

3.5. Validation

To derive ground-truth population values for each target zone, block population values in 2000 are aggregated to the target zone boundaries. Accuracy metrics such as Mean Absolute Error (MAE), Median Absolute Error, Root Mean Square Error (RMSE) and 90% percentile of absolute error are calculated based on measured and estimated tract values, and compared across the methods.

4. Results

Table 1 summarizes the results of all three refinement levels.

Table 1. Accuracy metrics of unrefined and refined methods.

Method	MAE	Median Absolute Error	RMSE	90th Percentile Error	Refinement Level
TDW	330	138	531	931	Unrefined
Refined TDW	235	99	379	672	First
Modified Refined TDW	178	75	283	503	Third
EM	447	262	702	1352	Second
Modified EM	236	136	390	611	Third
EEM	192	101	334	498	Second
Modified EEM	152	66	274	382	Third

Figure 1 shows the error maps. Moreover, Table 2 includes the mean normalized absolute errors of the third refinement methods divided by the mean normalized absolute errors of either first or second refinement methods for both total and rural target tracts.

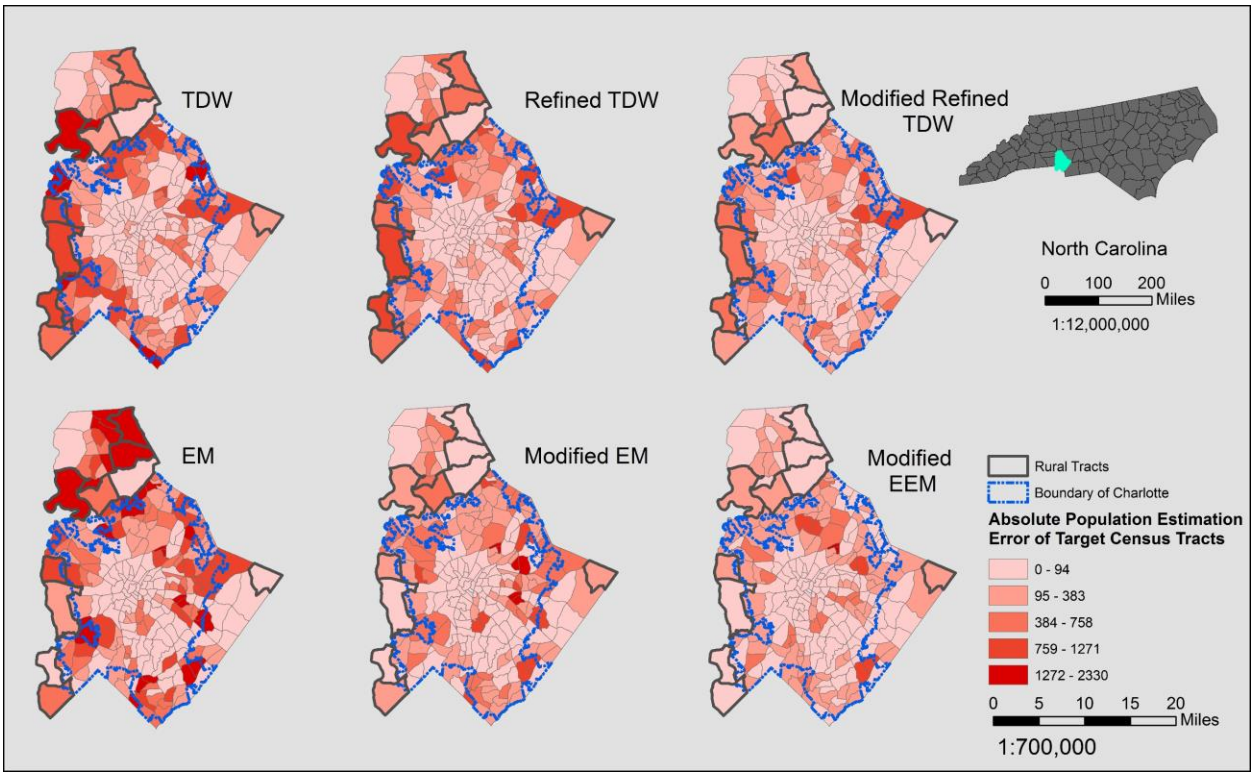


Figure 1. Absolute error maps of the methods.

Table 2. Comparison of the third refinement with first/second refinement.

Method	Total Tracts	Rural Tracts
ModRefAW/RefAW	0.82	0.28
ModRefTDW/RefTDW	1.13	0.53
ModRefPM/RefPM	0.87	0.29
ModEM/EM	0.88	0.22
ModEEM/EEM	1.25	0.38

As a coarse approximation for rural tracts, the number of rural households within each target tract is divided by its total count of households. Each tract with a proportion greater than 0.1 (10%) is considered rural.

5. Discussion and Future Research

Both Table 1 and Figure 1 demonstrate that spatial refinements reduce the error metrics, consistently. Refined TDW is more accurate than the unrefined method, and the third refinement is more accurate than the first. The pattern is similar in Areal Weighting (AW) and Pycnophylactic Modeling (PM) although not included in this paper. The third refinement outperforms EM and EEM as the second refinement methods. The most accurate method is Modified EEM.

As expected, the third refinement results in significant improvements for rural target tracts across all the methods even when it results in less accuracy for total target tracts (Table 2). A value lower than 1 indicates the mean normalized absolute error is lower for the third spatial refinement method than either the first or second spatial refinement approaches.

Future research will focus on data-driven optimization approaches for determining road buffer distance and expand the analyses to longer time periods and different study areas.

References

Buttenfield BP, Ruther M and Leyk S, 2015, Exploring the impact of dasymetric refinement on spatiotemporal small area estimates. *Cartography and Geographic Information Science*, 42(5):449–459.

Dempster A, Laird N and Rubin D, 1977, Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Gregory IN, 2002, The accuracy of areal interpolation techniques: standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems*, 26(4):293–314.

Leyk S, Ruther M, Buttenfield BP, Nagle NN and Stum AK, 2014, Modeling residential developed land in rural areas: A size-restricted approach using parcel data. *Applied Geography*, 47:33–45.

Mennis J, 2003, Generating Surface Models of Population Using Dasymetric Mapping. *The Professional Geographer*, 55(1): 31–42.

Ruther M, Leyk S and Buttenfield BP, 2015, Comparing the Effects of an NLCD-derived Dasymetric Refinement on Estimation Accuracies for Multiple Areal Interpolation Methods. *GIScience & Remote Sensing*, 52(2):158–178.

Schroeder JP, 2007, Target density weighting interpolation and uncertainty evaluation for temporal analysis of census data. *Geographical Analysis*, 39(3):311–335.

Zoraghein H, Leyk S, Ruther M and Buttenfield BP, 2016, Exploiting temporal information in parcel data to refine small area population estimates. *Computers, Environment and Urban Systems*, 58:19–28.

PLATINUM SPONSOR

RESEARCH NETWORK SPONSORS