Using Remote Sensing for Agricultural Statistics

Elisabetta Carfagna
University of Bologna, Department of Statistics
via Belle Arti 41, 40126
Bologna, Italy
Carfagna@stat.unibo.it

1. Possible ways of using remote sensing data for agricultural statistics

Remote sensing plays an important role as auxiliary variable in the production of agricultural statistics, when area frame or multiple frames sample designs are used. It can be used at the design level as well as at the estimator level. At the design level, the most typical use of remote sensing data is in the area frame construction and stratification. Remote sensing data can also be used to optimise the sample design where a previous ground survey is not performed; in fact, spatial characteristics (e.g. correlograms) of variables of interest can be estimated on the basis of photo-interpretation of remote sensing images. At the estimator level, remote sensing data are generally used as auxiliary variables in the regression estimator, but some attempts of using an estimator based on confusion matrixes have been done. Satellite data most often used are Landsat-TM and SPOT-XS.

Generally, a supervised classification of remote sensing data should not be used as a direct tool to estimate the area covered by a crop in a region. In fact, the estimator based on the ratio between the number of pixels classified into the specific crop and the total number of classified pixels is known to be strongly biased. Similar problems arise when photo-interpretation of satellite images is used for crop area estimation, since photo-interpretation errors tend to be systematic and, generally, there is no compensation between commission and omission errors for the different crops (Carfagna and Gallego, 1999).

2. The use of remote sensing images at the design level

For agricultural estimates, an efficient and low cost stratification is based on percentages of agriculture, often derived from photo-interpretation of remote sensing images. It is named land use stratification and creates strata with different rates of permanent and annual crops; but doesn't generally discriminate between annual crops. Sometimes, a distinction between areas with predominance of summer crops or winter crops is made. Remote sensing image photo-interpretation is often used to exclude some areas from surveyed area and no segments are allocated in presumed non agricultural area. In some cases, when detailed information is used and combined with other data such as altitude, this is a good practice to reduce the cost of the ground survey. However, the efficiency of the stratification is underestimated.

When an area frame is based on square segments without permanent physical boundaries, the area frame construction doesn't require the use of remote sensing data. Anyway, image photo-interpretation can be used to produce land cover maps that allow subdividing the population of segments into non-overlapping sub-regions in such a way that the variability of crop area per segment within each stratum is low. Practically, a square grid is overlaid on the land cover map. Then, the attribution of each segment to a stratum is made according to a specific criterion. The criterion of majority is generally used. In essence, the stratification is adapted to the sampling grid. Each square of the sampling grid is simply attributed to the stratum with the highest share, but part of the information in the stratification is lost.

Another way of using a photo-interpretation of remote sensing data, to produce agricultural statistics, is creating an index of agriculture intensity. When an area frame with square segments is adopted, this index should be calculated for each cell of a regular grid overlaid on the land cover map. Then a sample of segments can be selected with probability proportional to the values of the index. If, for reasons of economy, the information necessary to calculate the index can be collected only on a sample of the cells of the regular grid, a two-phase sampling can be set up. The second phase is a subsampling with a probability proportional to the index. An experiment carried out on some data collected in Spain in 1992 has shown that proportional probability sampling is very efficient when the ancillary information is highly accurate, but can be disastrous if a moderately high amount of errors appears in the ancillary information (Gallego et al. 1998).

Another important way of getting estimates with better precision and the same sample size is taking into account the positive spatial autocorrelation of agricultural variables, for example through the DUST (Dependent area Units Sequential Technique). This sampling technique (Arbia, 1993) modifies the simple random sampling selection probabilities once a first set of segments has been sampled. Selection probabilities of segments in the population increase with their distance from sampled segments. This sampling technique needs an input: the spatial autocorrelation between contiguous segments in the population. In fact, the higher the autocorrelation at short distances, the smaller will be the selection probabilities assigned by DUST to segments contiguous to selected ones. It is very difficult to estimate the autocorrelation at very short distances, even if a previous survey has been performed, since contiguous segments in the sample are generally so few that the estimate is unreliable. Remote sensing data offer a possible solution to the problem, since the spatial autocorrelation for contiguous segments can be estimated for variables derived from remote sensing data, if these variables are good proxies of agricultural variables. The precision of estimates of an area survey is heavily influenced also by other aspects of the area sample design, e.g. the target segment size and the number of sampling stages. If a previous survey has been carried out, its data can be used to improve the sample design by calculating the correlograms for each observed variable. In fact, an analysis of correlograms gives many suggestions for the optimum segment size and for the number of stages to be adopted to maximise the precision of estimates under a fixed budget and a given cost function (Carfagna 1998). In a similar way, when data have not been collected by a previous sample survey, photo-interpretation of remote sensing data can be used to calculate correlograms. Caution is necessary: if the correspondence between photo-interpreted classes and the variables of interest is not perfect. Usually, the classes of the photo-interpretation are aggregations of the variables of interest; thus the sample design is optimised for these aggregations. Besides, photo-interpreted data can be used as previous survey data only if the difference between their scale and the scale of data acquisition in the sample survey to be carried out is not too big.

3. The use of remote sensing data at the estimator level

The regression estimator is the most widespread way of using remote sensing data as auxiliary variable to improve the precision of estimates at the estimator level (Bellow, 1994). In some areas, experiments have been done to test the performance of a calibration estimator based on confusion matrixes. If the precision to be reached for the estimates is fixed, then the use of remote sensing data as auxiliary variables in the estimator can reduce the amount of ground data to be collected. On the contrary, if the sample size is fixed, the precision of estimates is improved. Auxiliary variables generally involved are thematic maps created by the classification of satellite images. The linear regression correction of a sample survey estimate is a technique to estimate the mean of a variable Y (i.e. area of a crop c) known for the n units of a sample, by using an auxiliary variable X (number of pixels classified into crop c, determined by classification of satellite data) known for the N elements of the whole population and correlated with Y. The area estimates are obtained separately for each crop. When the angular coefficient of the regression line (b) has a previously fixed value (e.g. b is estimated from another population or from a different sample) the regression estimator is unbiased and no assumption is required about the relation between Y and X in the finite population. If b is a least squares estimate, the regression estimator has a bias of order $1/n$. However, for large samples, it has usually a lower variance than the regression estimator with pre-assigned b, unless there is a good a priori knowledge of the reasonable value of *b*. The regression estimator with pre-assigned *b* is safer if unbiasedness is priority, distributions are highly skewed or *X* and *Y* have a non-linear link.

Satellite image classification is most widely carried out by the maximum likelihood classifier (discriminant analysis). It generally presents relevant problems and subjectivity. In fact, when the maximum likelihood classifier is used with uniform prior probabilities, large classes tend to be underestimated and small classes tend to be overestimated. On the contrary, if some information is available on the approximate proportion of the different land cover types, a proportional prior probability may be used; but in this case large classes tend to be overestimated and small classes are underestimated or even disappear. The bias appears also where the theoretical conditions on multivariate Gaussian distribution are true. Intermediate priors should be used that approximately give the correct number of pixels classified in each class. However finding these priors assumes the previous knowledge of the targeted number of pixels in each class (Gallego and Carfagna, 1998).

The crucial point in the use of the regression estimator is the cost effectiveness of the method. The ratio between the variance of the ground survey area estimate and the variance after this estimate has been corrected with the help of satellite images is named the relative efficiency of remote sensing. In the case of simple regression, if the sample is large enough, the efficiency is approximately $1/(1-\rho^2)$, where ρ is the linear correlation between *Y* and *X*. Relative efficiency of remote sensing gives a criterion for an economical evaluation of the procedure. It will be economical if the relative efficiency reaches a threshold, defined by the value for which the cost of remote sensing images acquisition and analysis equals the cost of increasing the size of the ground survey to get the same precision of estimates. Thus the threshold is influenced by variations in the cost of ground survey. Some studies consider that remote sensing starts to be efficient when the relative efficiency reaches a threshold of 1.5 for TM data and 2 for SPOT data; the difference is due to higher price of SPOT (Taylor et al, 1997). These thresholds were estimated in 1988; in successive years, the cost structure changed quite surprisingly, in the sense that ground survey costs decreased more than remote sensing costs and this resulted into much higher efficiency thresholds for cost effectiveness. Particularly, for SPOT-XS sensor, real threshold values revealed to be extremely high, due to high image cost.

Threshold values for TM sensor in Taylor et al, 1997 are very similar to the ones calculated by ITA Consortium for the operational project carried out every year in Italy; thus, they could be considered realistic only for an optimised process. The relative efficiency, across all main crops, in the Italian project was 1.83 in 1997 and 2.58 in 1998; thus, the regression correction of ground survey estimates is cost-effective in the Italian project, given its own cost structure. Other studies arrived at different conclusions.

The threshold values are considerably reduced if the same satellite images are used for more than one year; but good relative efficiencies can be achieved only in areas where the spatial distribution of crops is very stable (Gallego et al., 1998). Anyway, since the only problem concerning the use of the regression estimator with remote sensing data is the cost, we should consider that the cost-quality ratio of computers and satellite images is decreasing. Then, it must be taken into account that remote sensing provides more than an improvement of the statistical precision, since an information on the location of the crops, or other land uses, is also given.

A different approach, named direct calibration estimator based on a confusion matrix, was used in Belgium in 1992 (Gallego 1994) and in an area in the "département" Indre & Loire, in France (Brun et al.1992). In both cases, a SPOT-XS image was used. Relative efficiencies of the direct calibration estimator were compared with the relative efficiencies of the regression estimator. The regression estimator looked a bit more efficient. The use of confusion matrices for improving ground sample estimates is justified by the fact that, if the confusion matrix is based on a random or systematic sample of test segments, then it is an unbiased estimate of the confusion matrix in the population. However a very significant distortion may appear if the confusion matrix is calculated on the same pixels used for training the classifier. Such a matrix tends to give over optimistic estimates of classification accuracy. The evaluation of the classification can be biased also if training and test pixels are selected in the same segments, due to spatial autocorrelation (Gallego et al., 1998), when the number of test pixels is not much higher than the number of training pixels. In France, the ground survey was based on the TER-UTI system. The land use was observed only on 36 points of a regular grid superimposed on 36 segments of 1800 m × 1800 m. Thus, it was necessary to locate the points on the SPOT image. Then, a square block of 1, 9 or 25 pure pixels was created around each point, depending on the size of the field on which the point was located. No pixel was attributed to points falling on a border between fields. These pixels were used as training pixels to create the spectral signatures to be used in the image classification process. In this experiment, test pixels were a subset of training pixels; thus the accuracy of the classification, and consequently the relative efficiency of the method, were probably considerably overestimated

4. Crop yield estimation with remote sensing

Sometimes, remote sensing data have been used, to estimate the production of crops, exploiting the link between yields and vegetation indexes, using TM ~~(Bellow, 1994)~~ or AVHRR data. The latter kind of data, remotely sensed by NOAA satellites, have much lower ground resolution and much higher image acquisition frequency than TM data; thus they can be useful mainly for monitoring the agricultural seasonal behaviour (Benedetti and Palma, 1993).

REFERENCES

Arbia G., (1993). The use of GIS in spatial statistical surveys. Inter. Stat. Review, vol 63, n. 2, pp 339-359.

Bellow M.E. (1994). Application of satellite data to crop area estimation at the county level. Research report no. STB-94-02, NASS, UDA. Washington D.C.

Benedetti R., Palma P. (1993). Toward crop yield estimate and forecast by Remote Sensing: the use of NOAA/NDVI data. Proceedings of 49° ISI Session, Book 4, 309-318. Firenze.

Brun C., Delancé J., Lèo O., Porchier J. C. (1992). Pilot use of the TER-UTI data in agricultural statistics procedure using remote sensing. The application of remote sensing to agricultural statistics, Office for Official Publications of the E.C. Luxembourg.

Carfagna E. (1998). Area frame sample designs: a comparison with the MARS project. Proceedings of Agricultural Statistics 2000. International Statistical Institut. Voorburg.

Carfagna E., Gallego F.J. (1999). Thematic maps and statistics. Proceedings of Land Cover and Land Use Information Systems for European Union Policy Needs. Luxembourg, 21 - 23 January 1998. In press.

Gallego F.J., (1994). Using a confusion matrix for area estimation with remote sensing. Proceedings of Il telerilevamento spaziale per l'ambiente e il territorio in Italia:dalla ricerca al servizi, Rome, 1-4 May 1994.

Gallego F.J., Carfagna E. (1998). Some possible approaches to rapid crop area change estimation in a large region. Problemi statistici nel telerilevamento, L. Lionetti ed., Rubbettino, Soveria Mannelli.

Gallego F.J., Carfagna E., Fuenette I. (1998). Geographic Sampling Strategies and Remote Sensing. Report to EUROSTAT, F2, Agricultural Products and Fisheries.

Taylor J., Sannier C., Delincé J, Gallego F.J. (1997). Regional Crop Inventories in Europe Assisted by Remote Sensing: 1988-1993. Synthesis Report. Office for Publications of the EC. Luxembourg.

RÉSUMÉ

La façon la plus connue d’utiliser la télédétection pour les statistiques agricoles est de s’en servir pour construire une base d’échantillonnage, surtout pour la stratification. La télédétection peut aussi fournir des variables auxiliaires pour améliorer la précision des estimations, d’habitude à l’aide de l’estimateur de régression, mais d’autres méthodes basées sur des matrices de confusion peuvent être appliquées. On présente aussi l’utilisation de la télédétection pour optimiser une base d’échantillonnage en absence d’enquête préalable.. Les caractéristiques spatiales (corrélogrammes par exemple) de variables d’intérêt peuvent être estimées sur base d’images photo-interprétées.