GENERALIZED SOFTWARE TOOLS FOR CROP AREA ESTIMATES AND YIELD FORECAST

Roberto Benedetti[a], Remo Catenaro[a] and Federica Piersimoni[b]

[a] "G. d'Annunzio" University of Chieti-Pescara, Italy
Department of Business, Statistical, Technological and Environmental Sciences
[b] Istat, Agricoltural Statistical Service, Via A. Ravà 150, Rome, Italy

Commission VIII, WG VIII/10

KEY WORDS: informative system, sampling design, SAS, Visual Basic*,* data editing, optimal allocation, data dissemination.

ABSTRACT:

The procedure that leads to the estimates of the variables of interest, such as land use and crop yield and their sampling standard deviations, is rather tedious and complex, till to make necessary for a statistician to have a stable and generalized computational systems available. The SAS is often the ideal instrument to face with these needs, because it permits the handling of data effectively and provides all the necessary functions to manage easily surveys with thousands of micro-data.
This paper focus on the use of this system in different *steps* of the survey: sample design, data editing and estimation. The information produced is, however, available for one user only, the manager of the survey.
Our idea is to reduce the time needed to process the collected data and to reduce the inefficiencies which unfortunately characterize the management and dissemination of aggregated information, without loosing the stability of the SAS and at the same time increasing the informative content of the survey. For this reason we developed a user oriented Visual Basic (VB) software, whose aim is the production of *reports* based on the data coming from the SAS central system, capable of meeting the user needs of private and public agencies involved in this sector. If SAS is excellent for large scale data processing and has a very rich function set, VB 2005 helped us develop a high quality graphical interface that make possible to personalize estimations. Particularly, in this type of survey, we may be asked to provide estimates not only at the national level, but also at provincial and local levels, in other terms we may be interested at the *estimation for domains*: VB software support the definition of a priori domains.
We feel that through the use of this two integrated systems we could help researchers and institutions working on remote sensing in agriculture in growing their interest in this kind of large scale surveys.

1. INTRODUCTION

Information is an instrument of knowledge, and learning new information means raising the level of knowledge. Information becomes "statistical" if it is the result of a survey whose aim is to respond to well-defined cognitive needs.
An informative system is a system that organizes and administers competently and efficiently the information which is necessary to pursue one or more aims.
A statistical informative system aims at rationalizing the production and the spreading of statistical information through the use of some categories of data which, as already stated, is gathered through a survey. This system is characterized by the treatment of aggregates "macro-data" and through particular attention given to meta-data and to meta-information.
Therefore the task of the statistical informative system is:

- Data collection;
- Keeping collected data by archiving it;
- Data processing by transforming it into information;
- Dissemination of information.

To do this, the informative system can make use of technological data processing. Thus, the aim of our work is the creation of a modern informative system, which can develop efficiently, through the integrated use of statistical methods for the analysis of frames of extensive dimensions (data mining), the valuable information collected and elaborated in the field of agricultural surveys.
The system we are presenting foresees the integrated use of two software appliances (figure 1): the former developed by SAS, one of the most used systems of administering information, which deals with data-filing, sample design, data editing and estimation; whereas, the latter, developed in Visual Basic (VB), will help to administer the publication of the results and use the technique of domain estimation.

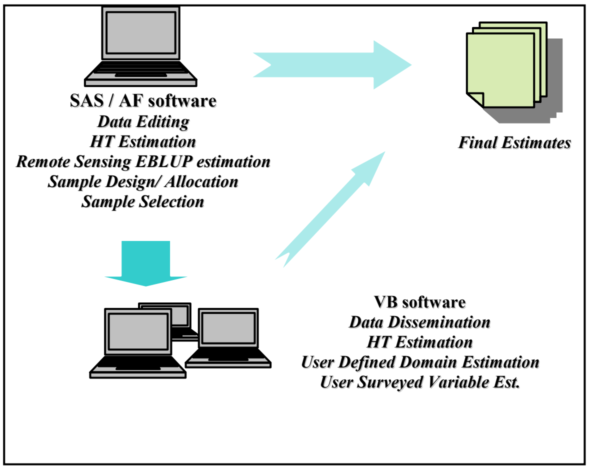

Figure1: Integrated use of two software developed by SAS and Visual Basic 2005.

In section 2 we will describe the way in which we collect data, in section 3 the problem of sampling design (Bethel, 1989), using the procedure of multivariate allocation and the technique of the *permanent random numbers* (Ohlsson, 1995) for the data mining are considered. In section 4 we will describe the problem of identification and correction of errors (data editing). In section 5 the way in which we estimate the variables of interest will be described and in the last one the phase of data dissemination.

2. DATA COLLECTION

In this section, we will describe the way in which we have carried out a physical integration, in a single relational database, of all the information coming from the periodic surveying done on Italian national territory in the field of agriculture.

In particular, for each survey included in the system, two types of information are produced: data and meta-data.

The relational database, which we have carried out, contains an frame of points, identified univocally by a numeric code.

The frame is associated or combined with one or more classifications. In the present project the data is registered and classified by: region, province and strata. Besides the identified codes of the units, the frame contains information about their geographical location and other information that can be used for sampling as well as producing estimations of certain basic characteristics as simple aggregations and tabulations.

The survey in question is the sampling type and for this reason the database contains, besides the frame, also the sample on which the surveys were carried out.

The operations on the database are carried out through the Import Section of the SAS software.

3. SAMPLING DESIGN

For this survey we used the technique of stratified sampling and the procedure of multivariate allocation, whose development (Bethel, 1989) requires a generalization of the classical formulas of calculation of the optimal size of Neyman (Neyman, 1934; Cochran, 1977).

The stratified random selection without replacement of the units is make through the use of the well known technique of the *permanent random numbers* (Ohlsson, 1995) in which, for every unit i of the frame of N dimension is associated, independently by the others, a meaning *pseudo-random* (*pseudo* because it is generated by a *computer*) by a rectangular variable r_i. The stratified simple random sample of a survey is obtained sorting the units for strata and *random number* so that we obtain a random permutation of the units and we can select for every strata h the first units (or the last) n_h.

This technique, particularly easy to develop, assure that, after sorting the frame, every selection of adjacent units represents a random sample with inclusion probability equal to n_h/N_h.

The sample design is carried out through the Sample-Bethel section of the SAS software (figure 2).

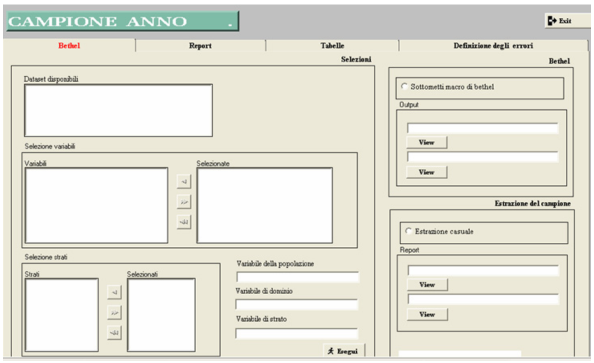

Figure 2: Section Sample-Bethel

4. DATA EDITING

For every survey of sampling or census, the problem of identification and correction of errors present in the information taken comes up.

Although there is no real agreement on the necessity of intervening automatically on the possible errors or on reconstructing the missing information, it is necessary to make the problem of coherence and completeness as automatic is possible and this means a reduction in the time necessary for the elaboration of said data as well as an improvement in the quality of the final information produced.

This last point takes on a remarkable importance if it is considered that the units in which the errors are found rarely constitute a random sub-sample of the population referred to.

Very often even a restricted number of errors can produce a wide impact on the estimation when the said number is localized in the most important units. In such a case, therefore, it is important to make sure that the errors in the units in question are found and eliminated in the most accurate way possible.

Before going ahead it is necessary to state that, according to us, the main problem is not so much in the particular system of correction adopted as in the identification of the errors.

Once the errors have been localized, the fields taken into account are transformed into missing data and, only at this stage can the phase of more or less sophisticated reconstruction of the missing information be activated. As in many surveys the partial missing information is practically inexistent, it is opportune to point out that the problem of missing data to be estimated is often generated after the identification of the errors.

The SAS section that allows data editing is structured in such a way as to allow access to six modules:

The first two are destined to find *outliers* through observation of the variable in question in two different periods of time, using *scatter* and *histograms* (figure 3);

The third allows the crossing of several variables in a set period and the finding of *outliers*;

The fourth module applies to the operations of *macro-editing* based on a manual or automatic type of correction;

The fifth allows the definition of rules concerning the compatibility plan.

The sixth applies to the correction of data and is made up of a table, which can be modified, where doubtful fields, found in preceding modules, are shown.

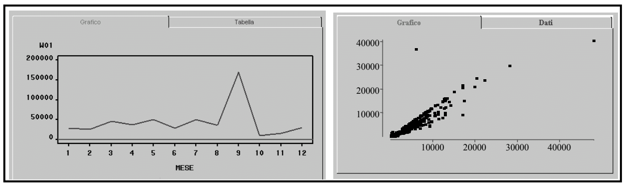

Figure 3: An example for data editing

5. ESTIMATION

Let us consider $s=\{1, \ldots, i, \ldots, n\}$ a sample consisting of n points, randomly selected by a population Ω, through the sample design which assigns to s the probability $p(s)$ to be selected. Analogically at the population, the sample s is divided in $s_1, s_2, \ldots, s_h, \ldots, s_H$. Let us consider $y = \{y_1, y_2, \ldots, y_j, \ldots, y_J\}$ a vector of the J variables of interest and $\pi_i = \sum_{i \subset s} p(s)$ the inclusion probability in the sample for the generic unit $i \in \Omega$.

Suppose that an estimate is needed for the population total of every interest variable *J*.
If we have no good auxiliary data or missing data, an unbiased estimator of the total $Y_j = \sum_{i\subset\Omega} \mathrm{y}_{j,i}$ for the *j* variable in the population is represented, as known, by the Horvitz-Thompson estimator (Horvitz e Thompson, 1952):

$$\hat{Y}_{j,HT} = \sum_{i\subset s} d_i y_{j,i} = \sum_{i\subset s} \frac{N_h}{n_h} y_{j,i}$$

with the variance:

$$\hat{V}(\hat{Y}_{j,HT}) = \sum_{i\subset s} \sum_{z\subset s} \frac{\pi_{iz} - \pi_i \pi_z}{\pi_{iz}} d_i y_{j,i} d_z y_{j,z} =$$

$$= \sum_{h=1}^{H} d_h (N_h - n_h) S^2_{y_j h}$$

where $S^2_{y_j h}$ is the sample variance of y_j in the strata *h* in which the unit *i* is included.
π_i is the first order inclusion probability and $d_i = 1/\pi_i$.
π_{iz} denote the probability that both of the elements *i* and *z* will be included.

The estimates of the variables land use and crop yield at regional and provincial level apart from the national one, take us employing an instrument which considers these needs. The *SURVEYMEANS* procedure allows carrying out sub-populations or domains estimations (Cochran, 1977; Rao, 2003) apart from the population. In particular, for the estimates of the land use, we use *SURVEYMEANS* for totals, instead, for yield forecast we use *SURVEYMEANS* for means.
SURVEYMEANS uses the Taylor series expansion theory to estimate the variance of the mean $\hat{\bar{Y}}$ (SAS manual). The procedure computes the estimated mean as

$$\hat{\bar{Y}} = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \omega_{hij} y_{hij} \right) / \omega ...$$

$$\omega ... = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \omega_{hij}$$

and the estimated variance of $\hat{\bar{Y}}$ as

$$\hat{V}\left(\hat{\bar{Y}}\right) = \sum_{h=1}^{H} \hat{V}\left(\hat{\bar{Y}}\right)$$

where if n_h>1,

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi}\cdot - \bar{e}_h\cdot\cdot)^2$$

$$e_{hi}\cdot = \left(\sum_{j=1}^{m_{hi}} \omega_{hij} \left(y_{hij} - \hat{\bar{Y}} \right) \right) / \omega ...$$

$$e_h\cdot\cdot = \left(\sum_{j=1}^{n_h} e_{hi}\cdot \right) / n_h$$

and if n_h=1,

$$\hat{V}\left(\hat{\bar{Y}}\right) = \begin{cases} \text{missing if } n_h = 1 \text{ for } h = 1,2,...,H \\ 0 \quad \text{if } n_h = 1 \text{ for some } 1 < h < H \end{cases}$$

The estimation of the total is the weighted sum over the sample:

$$\hat{Y} = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \omega_{hij} y_{hij} \right)$$

and the estimated variance of $\hat{Y}$ is: (5.2)

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{H} \hat{V}(\hat{Y})$$

where if n_h>1, (5.3)

$$\hat{V}\left(\hat{\bar{Y}}\right) = \frac{n_h (1 - f_h)}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi}\cdot - \bar{y}_h\cdot\cdot)^2$$

$$y_{hi}\cdot = \sum_{j=1}^{m_{hi}} \omega_{hij} y_{hij}$$

$$\bar{y}_h\cdot\cdot = \left(\sum_{i=1}^{n_h} y_{hi}\cdot \right) / n_h$$

and if n_h=1,

$$\hat{V}\left(\hat{\bar{Y}}\right) = \begin{cases} \text{missing if } n_h = 1 \text{ for } h = 1,2,...,H \\ 0 \quad \text{if } n_h = 1 \text{ for some } 1 < h < H \end{cases}$$

where h = 1, 2, ... , H is the stratum number, with a total of H strata
i = 1, 2, ... , n_h is the cluster number within stratum h, with a total of n_h clusters
j = 1, 2, ... , m_{hi} is the unit number within cluster i of stratum h, with a total of m_{hi} units
$n = \sum_{h=1}^{H} \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
w_{hij} denotes the sampling weight for observation j in cluster i of stratum h
$\mathbf{y}_{hij} = (y_{hij}^{(1)}, y_{hij}^{(2)}, \dots , y_{hij}^{(P)})$ are the observed values of the analysis variables for observation j in cluster i of stratum h, including both the values of numerical variables and the values of indicator variables for levels of categorical variables.

The standard error of the mean is the square root of the estimated variance.
We can show that, not considering the clusters, the results arising by the *SURVEYMEANS* are those achieved by the formulas proposed by Horvitz e Thompson in the 1952.
After the phase of sample selection and the data-editing with the SAS, it generate a database composed by the samples for the estimates of the variables land use and crop yield. This database will be imported in the VB software, which, allows to execute the *SURVEYMEANS.* In other words, it allows to personalize the estimates, for domains: Province and Region or directly defined by the user.

6. DATA DISSEMINATION

Protection of the data contained in the database is very important. Safety, in fact, is at the basis of any project that foresees investments on data of various kinds and in particular on data considered by the law to be “sensitive”. Regarding surveys in agriculture, there is no contact with data strictly sensitive, but in our case we have preferred to deal with basic data as such. In particular, the database is protected in such a way that access can come about exclusively through software interface. In addition, access is reserved only for aggregated data or a database compiled directly by the user. Besides access to the system can come about only through the insertion of a user name and a password.

In the phase of developing the system we pre-established the attainment of certain well-defined objectives. In particular, if the SAS software whose to have bettered the organization of agricultural surveys and lead to a reduction of calculation times, the VB software should have augmented the informative capacity of such surveys.

The inspiring principle which accompanied us during the phase of the development of the software in Visual Basic 2005 was that of the simplicity of use thanks to a strong intuitive interface.

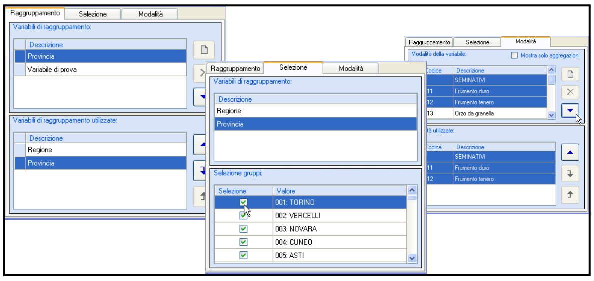

Figure 4: Settings for estimation

The result is a software that in a few clicks allows the carrying out and the personalising of crop yield estimation and land use at different levels of aggregation.

References

Bethel, J. (1989), “Sample Allocation in Multivariate Surveys,” Survey Methodology, 15, 47-57.

Cochran W.G. (1977) *Sampling Techniques*, Wiley, New York.

Horvitz D.G., Thompson D.J. (1952), A generalization of sampling without replacement from a finite universe. Journal of American Statistical Association, 47, 663-685.

Neyman J. (1934) On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of the Royal Statistical Society*, 97, 558-625.

Ohlsson E. (1995) Coordination of samples using permanent random numbers, in: Business Survey Methods, Cox B.G. et al. (Eds.), Wiley, New York, 153-169.

Rao J.N.K., (2003) *Small Area Estimation*, Wiley, New York.

Särndal C-E, Swensson B., Wretman J. (1992) Model Assisted Survey Sampling, Springer Verlag, New York.

SAS® OnlineDoc, 9.1.3: PDF Files, SAS Institute, Cary North Carolina.

Visual Basic 2005: Microsoft Developer Network (MSDN) Library, Microsoft Corporation 1987-2004.