



7 th WORKSHOP ON LABOUR FORCE SURVEY METHODOLOGY

DATA PROCESSING AND DATA QUALITY

Madrid, Spain, 10 – 11 May 2012

**D. Data processing: Effective IT strategies for data processing
(including LFS files for Eurostat) and data storage**

**D2 – An IT framework for a quick evaluation of accuracy of
Italian LFS**

Alessandro Martini – Italy

An IT framework for a quick evaluation of accuracy of Italian LFS

**Cinzia Graziani cingraziani@istat.it, Silvia Loriga siloriga@istat.it,
Alessandro Martini alemartini@istat.it, Andrea Spizzichino spizzich@istat.it**

Italian National Institute of Statistics (LFS Division)

Introduction

In this paper is described the development and implementation of a generalized instrument to improve dimensions of statistical quality such as accessibility, usability, accuracy and interpretability for sample surveys.

In this context Business Intelligence technology (data-warehouse, Olap, Data mining, web application) can be very useful to make easier the process to get information and knowledge from data so it can be quite easy to improve accessibility and usability.

Application metadata, which usually allow navigation through the available information, have been integrated with a set of methodological metadata. This provides a generalized metadata driven tool to support users with different degree of statistical literacy in the evaluation of sample survey estimates. The final goal is the control of critical dimensions of statistical quality:

- Accuracy: defined as the closeness between the estimated value and the (unknown) true value, to assess the reliability of the estimates;
- Interpretability: reflects the easiness for users to understand and properly use and analyze the data or information, for example to evaluate differences among estimates referred to different times or sub populations.

Background

The analysis of the results of a sample survey should always be accompanied by an assessment of the accuracy of the estimates, a measure of the dispersion of the estimates around the true value of the parameter in the population. Furthermore it is an important obligation for national statistical institutes to estimate sampling errors and then to disseminate and represent them to the users in a transparent and clearly readable way.

On the other hand, protocols for LFS micro data communication are very common, in this way many researchers can access to individual (detailed) data and produce estimates on their own in the complete absence of any guidance ensuring exploration consistency except methodological notes and documentation. Thus a main problem arise: the reliability and consistency of the analysis outputs, according to the specifications and the methodological limitations related to the underlying sampling design.

Methods

In sample surveys, specific methods are often used to improve accuracy and to *control* the costs of survey data collection: samples are often clustered geographically, or based on multiple sampling stages so the form of estimators,

usually calibration estimators, is no longer linear. These methods add further complexity to the analysis, which must be accounted in order to correctly use sample estimates. To assess the accuracy of an estimator is appropriate to refer to MSE to take into account its variation and unbiasedness. Calibration estimator is uncorrect but, with increasing size of the sample, the estimator converges asymptotically to the estimator of generalized regression (Deville and Särndal, 1992), It follows that for sufficiently large samples (such as those on labor force survey), we can assume that the estimator has approximately the same properties of the generalized regression estimator (accuracy, consistency) and that have the same sample variance.

An exact computation of the estimate variance is easy only in case of simpler sampling designs.

In all other cases the estimation is quite difficult and requires high-demand procedures in terms of computational complexity.

Moreover, questionnaires are also very complex and many estimations usually are produced so the publication of estimations variances would be very difficult to bring up and to interpret for users. For these reasons regression models may be used to produce synthetic evaluations of sampling errors. The hypothesis is the existence of a relation between relative sampling error $\varepsilon({}_d\hat{Y})$ and the estimation \hat{Y}_d , in particular for qualitative variables a model specification which shows a good fit is:

$$\log \hat{\varepsilon}^2({}_d\hat{Y}) = a + b \log({}_d\hat{Y}) \quad d = 1, \dots, D \quad (1)$$

where d is the general domain of interest.

Models (1) are fitted for each domain of interest on a wide set of estimates, taking care to choose heterogeneous levels for them.

Table 1. Sampling error regression model, NUTS II level, IT-LFS 2011 - Estimated values for R^2

| Nuts II Area | R^2 | | | | |
|----------------|--------|--------|--------|--------|------|
| | 2011Q1 | 2011Q2 | 2011Q3 | 2011Q4 | 2011 |
| Piemonte | 95,1 | 94,4 | 95,1 | 94,5 | 95,5 |
| Valle d'Aosta | 93,6 | 93,2 | 94,0 | 94,1 | 94,8 |
| Lombardia | 94,4 | 96,0 | 94,5 | 95,5 | 96,3 |
| Trentino Alto | | | | | |
| Adige | 95,0 | 94,9 | 94,8 | 95,0 | 95,6 |
| Veneto | 93,8 | 94,2 | 95,2 | 94,0 | 95,6 |
| Friuli Venezia | | | | | |
| Giulia | 92,1 | 93,5 | 93,1 | 94,3 | 95,1 |
| Liguria | 92,1 | 93,2 | 94,8 | 94,2 | 95,7 |
| Emilia Romagna | 92,4 | 93,8 | 93,5 | 92,5 | 94,7 |
| Toscana | 92,5 | 93,6 | 93,3 | 93,4 | 94,7 |
| Umbria | 93,8 | 93,8 | 94,4 | 94,6 | 95,2 |
| Marche | 94,5 | 92,8 | 94,9 | 95,2 | 96,2 |
| Lazio | 94,6 | 92,2 | 92,5 | 95,9 | 96,2 |
| Abruzzo | 92,5 | 94,3 | 93,8 | 92,4 | 94,9 |

| | | | | | |
|--------------|-------------|-------------|-------------|-------------|-------------|
| Molise | 94,1 | 94,2 | 92,9 | 94,5 | 96,0 |
| Campania | 95,5 | 96,1 | 96,5 | 95,6 | 97,2 |
| Puglia | 96,2 | 94,8 | 94,2 | 93,9 | 96,3 |
| Basilicata | 95,0 | 95,8 | 96,2 | 95,5 | 96,7 |
| Calabria | 94,2 | 94,8 | 95,0 | 93,7 | 95,2 |
| Sicilia | 94,8 | 94,7 | 95,9 | 95,3 | 96,5 |
| Sardegna | 92,0 | 93,5 | 89,7 | 92,5 | 95,2 |
| Italy | 96,8 | 96,8 | 97,0 | 97,1 | 97,3 |

Deriving a and b parameters by this way over each domain is quite easy to evaluate for a given estimation \hat{Y}_d , in the d^{th} domain, the relative sampling error by applying the following formula:

$$\hat{\varepsilon}(\hat{Y}) = \sqrt{\exp(a + b \log(\hat{Y}))} \quad (2)$$

The estimation of relative sampling errors makes it possible to define a confidence interval which, with a given probability \square , is likely to include the actual value Y_d .

$$(\hat{Y} - z_{1-\alpha/2} * \hat{Y} * \hat{\varepsilon}(\hat{Y}); \hat{Y} + z_{1-\alpha/2} * \hat{Y} * \hat{\varepsilon}(\hat{Y})) \quad (3)$$

This method is still valid to estimate a generic population proportion P_d defined as:

$$\hat{P}_d = \frac{\hat{Y}_d}{N_d} \quad (4)$$

Where N_d refers to a known total of the population, used as a constraint in the calibration.

If the estimate is a ratio where both numerator and denominator are estimates, an approximation is necessary :

$$\hat{R}_d = \frac{\hat{Y}_d}{\hat{D}_d} \quad (5)$$

An approximate evaluation of relative sample error for \hat{R}_d , under the hypothesis of uncorrelation between \hat{R}_d and \hat{D}_d , can be defined as:

$$\hat{\varepsilon}(\hat{R}) = \sqrt{\hat{\varepsilon}^2(\hat{N}_d) - \hat{\varepsilon}^2(\hat{D}_d)} \quad (6)$$

The SAS prototype and first results

The first prototype, based on SAS system, widely uses metadata to manage available information and makes the user able to browse it. Moreover SAS System is also the computing environment used to estimate variances in complex sample surveys conducted by Istat.

“Application metadata” have been integrated with a set of “methodological metadata” that refers to the information needed to run variance estimation procedures:

- stratification;
- sample design;
- estimation domains;
- non response adjustments;
- calibration constraints;

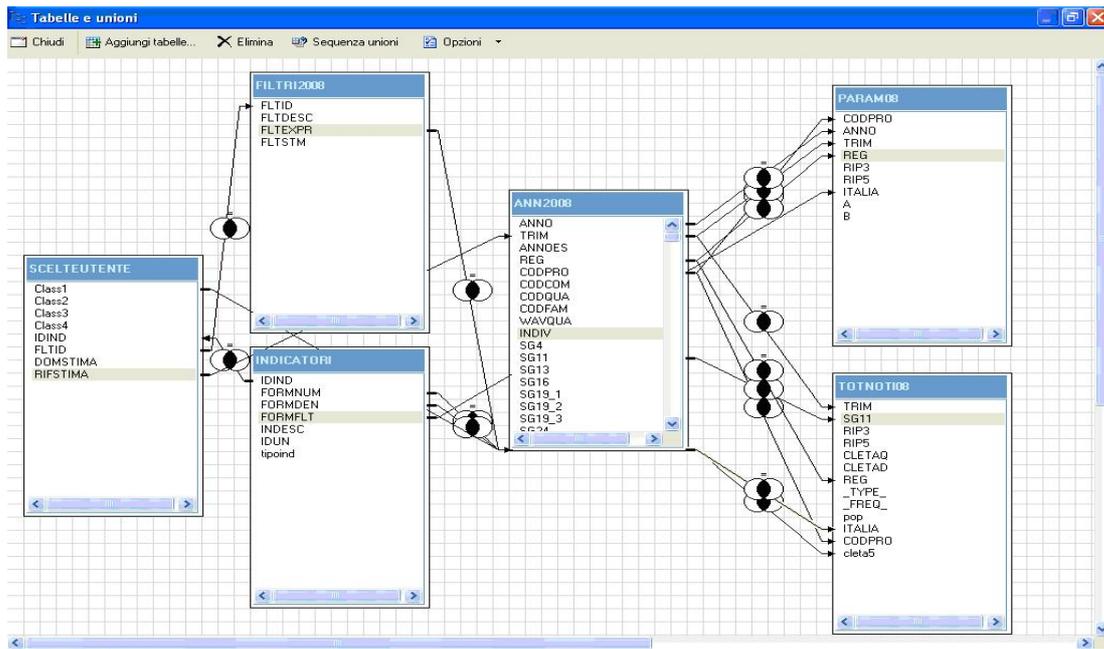
and their final output:

- weights;

- estimated parameters for regression models on relative sampling errors.

The set of metadata also includes a general formalization for ratio estimates that allows to calculate separately numerator and denominator relative sampling errors. We also define a classification for ratios in order to distinguish between those having estimates (5) or known population total as denominator (4). This classification allows to apply the correct method for evaluating relative sample error, using the formula (6) or (2), respectively.

Figure 1. Methodological Metadata Structure

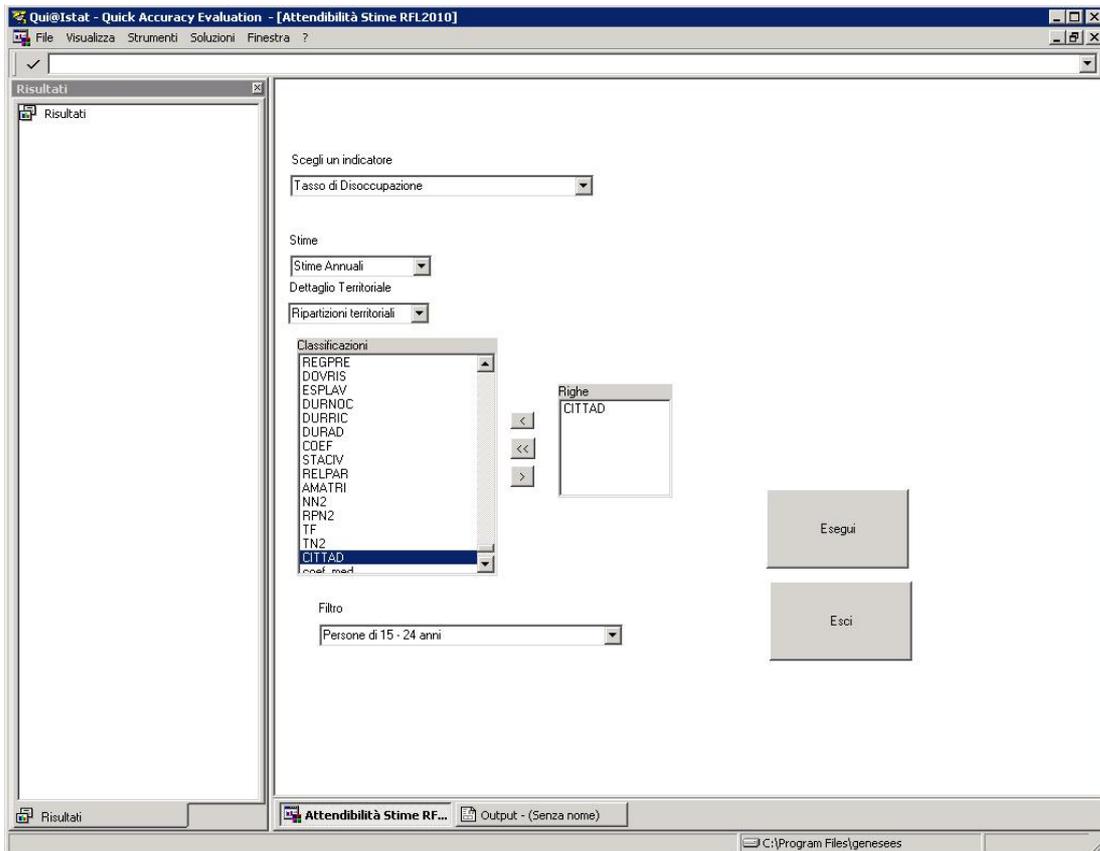


The method choice is not unique since the denominator can become an estimate even for a proportion of a total. An example is given by using Italian Labour Force Survey for the estimation of the activity rate by region and age classes, in this case the denominator consists of a known total considered in calibration procedure. However, if the user applies a filter, specifying the analysis for married individuals, the denominator becomes an estimate (the amount of married individuals) and the formula (6) is required instead of (2). A constant reference to the methodological metadata during processing allows even in this case the correct application of the methods described above. Estimates, once they have been calculated, are compared with known population totals and the correct formula to evaluate the relative sample error can be applied.

Procedures have been written in SAS macro language and shared in a server through SAS/AF forms.

These forms provide graphical interfaces for the users and allow them to define the parameters of the analysis, while each selection is automatically translated in SAS or SQL instructions. Many client applications can connect to the server and call these procedures. Working groups deemed them really useful in estimates validation phase, making estimation comparison easy and effective.

Figure 2. Selection of parameters



The time needed for data processing is quite short, since to carry out the estimations accessing microdata survey datasets is the most complex step of the algorithm. All subsequent steps, performed by SAS data step or SQL instructions, are extremely fast since performed by manipulating and transforming estimations and integrating them with metadata. Finally output is obtained using PROC tabulate and it reports estimations, their confidence interval and an evaluation of estimations accuracy. These outcomes are obtained applying the methods quoted above and synthesized through a graphical presentation. Procedures have been developed for Italian Labour Force Survey and Multipurpose Social Surveys to support researchers in evaluating accuracy and to improve interpretability of data.

Figure 3. An example of the output of the procedure

Tasso di Disoccupazione
 Persone di 15 - 24 anni
 Anno 2010

| RIP5 | Tasso di Disoccupazione | | | | | | | |
|----------------|-------------------------|-------|-----------|---------------|-----------|-------|-----------|---------------|
| | Italiani | | | | Stranieri | | | |
| | Lim. Inf. | Stima | Lim. Sup. | Attendibilità | Lim. Inf. | Stima | Lim. Sup. | Attendibilità |
| rip5 | | | | | | | | |
| 1 - Nord ovest | 19.2 | 20.8 | 22.5 | ████████ | 21.7 | 26.1 | 30.5 | ████████ |
| 2 - Nord est | 15.9 | 17.6 | 19.3 | ████████ | 22.6 | 27.7 | 32.8 | ████████ |
| 3 - Centro | 24.2 | 26.4 | 28.7 | ████████ | 18.2 | 23.3 | 28.3 | ██████ |
| 4 - Sud | 36.7 | 38.6 | 40.4 | ████████ | 14.2 | 22.2 | 30.2 | ██████ |
| 5 - Isole | 38.9 | 41.3 | 43.6 | ████████ | 10.4 | 22.6 | 34.8 | ██████* |

The development in a Business Intelligence framework

The next goal should be to extend metadata information in a business intelligence framework in order to efficiently manage OLAP processing and to enable roll-up and drill-down operations.

Business Intelligence (BI) refers to computer-based techniques used in spotting, digging-out, and analyzing business data to gain an advantage on market competitors.

However, our attention is focused on BI technological framework (data warehouse, Olap, Data Mining, web application) with the objective of improving several dimensions of data quality.

The improvement of accessibility to a common set of validated statistical microdata could be one, the easier to achieve, of our scopes. The capability of self-service reporting might significantly extend the audience of users with different degree of statistical literacy.

They could navigate and produce analysis through self-reporting data by simply using a web browser, avoiding duplicated and/or inconsistent loading and transformation data procedures.

Applying standard data warehouse tools can not offer sufficient guarantees of respecting specific dimensions of data quality. In particular they do not take into account any evaluation of estimates' accuracy and outcomes coherence with sampling design and objectives.

Integration of metadata concerning sample design and variance estimation outcomes can be useful to override these critical aspects of analyzing sample survey microdata using data warehouse tools.

Developing in this technological framework the capabilities described above, that can take into account estimation accuracy, could improve this important aspect of sample survey data quality and their interpretability.

References

Isfol - Centra M., Falorsi P.D. (a cura di), Strategie di campionamento per il monitoraggio e la valutazione delle politiche, Roma, Isfol - Temi e strumenti, 2007

Istat - De Francisci S, Sindoni G., Tininini L, DaWinci/MD: un sistema per data warehouse statistici sul web

Roma, Istituto Nazionale di Statistica, Contributi n. 14 , 2005

http://www.istat.it/dati/pubbsci/contributi/Contributi/contr_2005/2005_14.pdf

Deville, J. C., Särndal, C. E., Calibration Estimators in Survey Sampling, Journal of the American Statistical Association, vol. 87, 1992

<http://www.jstor.org/pss/2290268>

Isfol - Di Giammatteo Michele, L'indagine campionaria ISFOL-PLUS: contenuti metodologici e implementazione, Studi Isfol, 2009/3

http://www.isfol.it/Studi_Isfol/Dettaglio_Studi/index.scm?codi_nota=371&codi_percorso=51

Istat - Gazzelloni S. (a cura di), La rilevazione sulle forze di lavoro:Contenuti, metodologie, organizzazione, Roma, Istituto Nazionale di Statistica, Metodi e norme n. 32 , 2006

http://www.istat.it/dati/catalogo/ricerca.php?tipo=n&ciclo=0&stringa=&collane%5B%5D=14&num_collana=32&anni%5B%5D=2006

Istat - Pagliuca D. (a cura di), GENESEES V. 3.0 Manuale utente e aspetti metodologici, Istituto Nazionale di Statistica, Tecniche e Strumenti 3, 2005

http://www.istat.it/strumenti/metodi/software/produzione_stime/genesees/index.html

Woodruff R. S., A Simple Method for Approximating the Variance of a Complicated Estimate

Journal of the American Statistical Association, Vol. 66, n. 334, 1971.

<http://www.jstor.org/pss/2283947>