AN INTEGRATED ARCHIVE OF THE LIFESTYLES OF FAMILIES[1]

Silvestro Montrone, Antonella Massari, Paola Perchinunno, Stefania Girone

1. Introduction

More and more often statistical data are collected with the preconceived objective of combining the information obtained by different investigations, whether sample surveys or censuses, both to extend the extent of knowledge and to guarantee a greater quality of data. The growing potential of computer science makes possible the acquisition, organization and creation of bigger and bigger data archives of ever greater quality.

The methodologies with a statistical basis utilized for the integration of data from a number of sources are: the techniques of Record Linkage (or exact matching) and of Statistical Matching (or synthetical Matching). The techniques of Record Linkage use specific algorithms that are adapted to identify pairs of records (related to the same statistical unit) in two different databases. The purpose of the methodology is therefore to integrate and match information which, though contained in different archives, can be attributed to the same statistical unit that is not identifiable by means of a single code that does not contain errors. The objective of the techniques of Statistical Matching is to identify records relating to similar units and to estimate the unified distribution of a number of variables observed in different data archives and to merge informative records. Both matching techniques make it possible to obtain integrated archives, adopting reasonable and statistically verifiable assumptions.

The present work will describe a model of data integration through a methodology of Statistical Matching (hot deck distance) for the integration of two surveys (EuSilc-Istat and Lifestyle Survey-University of Bari). The construction of an integrated database on the basis of these two surveys may be useful for the study of consumer behavior in relation to specific groups of commodities, in order to analyze the decisions taken by families with regard to saving, to examine economic and social inequality, and to study the impact of public policies by means of

[1] This contribution is the result of joint reflections by the authors, with the following contributions attributed to: S. Montrone (chapter 1,5), to A. Massari (chapter 4), to P. Perchinunno (chapter 2), to S. Girone (chapter 3).

simulations. The coexistence of multiple and differentiated objectives triggers the need to obtain a very general and versatile integrated file, which provides ongoing detailed information on the different types of spending, on levels of saving, on the distribution of incomes, on the occupational conditions of the members of the family unit, etc.

2. Techniques in data integration

2.1 Introduction

The statistical matching techniques are being used in order to link information that come from two or more datasets whose units contain similarities regarding a set of variables previously defined. The basic assumption is that the sources which should be integrated contain information on a series of common variables as well as information on different variables that are never been jointly observed and it would be interesting to relate them to each other.

The idea of using the integration techniques on data that come from multiple sources in order to combine and enrich the information gathered in several surveys is not a recent one: the first methodological examples date back to the '60s. The first important applied examples are in Okner (1972 e 1974) and Ruggles N. N. and Ruggles R. (1974). In recent years, there has been a renewed interest for these techniques after a broader recognition of their application potential, in particular in the area of national statistical institutes' activities (Schoier, Torelli, Zacchigna, Egidi, Sabbadini, 2006).

The methodologies, with statistical basis, that are used to integrate the data deriving from multiple sources can be grouped in two types:

1. *record linkage* or exact matching;
2. *statistical matching* or synthetic matching.

The techniques of exact matching (*record linkage*) aim at identifying the pairs of records, related to the *same* statistical unit, belonging to different datasets by using specific algorithms.

The techniques of statistical matching aim at identifying the *similar units*, estimating the joint distribution of multiple observed variables in different datasets and fusing the information records. Both matching techniques, the exact and the statistical one, provide integrated archives by adopting rational and statistically controllable assumptions (D'Orazio, Di Zio, Scanu, 2002).

2.2 The statistical matching

In the case of statistical matching the initial assumption is to have *two different*

archives that contain information records on two groups of units selected from the same population. Therefore, the starting point of the issue is the existence of two archives:

1. containing information on common variables (the socio-demographic type of variables) as well as on different variables that are never been jointly observed;
2. the sources' units to be integrated are separated, or rather the sample surveys are independent among them (Montrone, Perchinunno, 2005).

Hence, in the statistical matching the individual records deriving from two or more sources are linked, on the basis of their *similarities*, through a set of characteristics measured in each source. We consider here two datasets containing two files: the file A and the file B. In order to make the statistical matching of these files, it is necessary that the common information regarding the units is available in each file. Let X_A be the set of variables measured in the file A, and let X_B be the set of variables measured in the file B, it is assumed that these two sets of variables can be transformed in one set with common characteristics. We can indicate the individual's characteristics measured in the both datasets as the vector $\boldsymbol{X} = (X_1, \ldots, X_P)$. The remaining variables in each file, that are not overlapping, are indicated as $\boldsymbol{Y} = (Y_1, \ldots, Y_Q)$ in the file A and as $\boldsymbol{Z} = (Z_1, \ldots, Z_R)$ in the file B.

The common variables *X*, defined from now on as the *integration variables*, are used to identify the units to be linked, while the non-common variables (*Y* e *Z*), defined as the *descriptive variables*, represent the information that is the object of the matching procedures. Therefore, it determines a situation in which the information that is simultaneously gathered on the same units for *Y* and *Z* is missing.

The aim of the statistical matching is to create a file, the *file C* (*the synthetic file*) in which each record contains all variables *X, Y and Z.* For each unit in the file A, a similar unit in the file B is identified, whereas the similarity is evaluated in terms of a function of variables *X*. The variables *Z* in the file B (defined as *the donor*) are then attributed to the matching record in the file A, creating thus a *record with full data (X, Y, Z).*

Considering the two archives A and B (which in this study case are the archives of family consumption and income), we will match the common variables in order to link the descriptive variables.

Some of the methods used for statistical matching are the following:

- *hot deck random*;
- *hot deck rank*;
- *hot deck distance*.

The method applied in this work is the *hot deck distance* which consists of matching the record *a* from A with the record b^* from B that is the "closest"

considering the variety of common variables. Hence, we will have that:

$$d_{a,b^*} = \left| x_a^{\mathrm{A}} - x_{b^*}^{\mathrm{B}} \right| = \min_{1 \le b \le n_{\mathrm{B}}} \left| x_a^{\mathrm{A}} - x_b^{\mathrm{B}} \right| \quad (1)$$

Each unit of the base dataset (i.e. archive defined as the receiving one) is operatively associated to another unit of the second dataset (i.e. the archive defined as the donor) by applying the *distance function* $d(x_i, x_j)$ which is computed on the integration variables and assumes much lower values as the individuals are more similar among themselves *(nearest neighbor match)*.

3. Construction of the integrated Eu-Silc and Lifestyles archive

3.1 Data sources

This report evaluates the possibility of utilizing two different sample sources to construct an integrated database with information relating to the economic situation of families. The two surveys used are *Eu-Silc* (ISTAT) and *Lifestyles* (University of Bari "A. Moro").

The *Eu-Silc survey* conducted by Istat collects detailed information on incomes and on family expenditure for the purchase of goods and services for final consumption, on family typology (in terms of composition and characteristics of the family head), on living conditions and spending habits, and relevant information about individual incomes, on the employment of savings and on family wealth.

On the other hand, the *Lifestyles survey* collects information on income, on spending behavior, and on the recourse to loans by families with children.

It is therefore interesting to verify the possibility of creating integration between the two archives, though always bearing in mind that the available data in both cases concern different samples from the same population: in addition, neither of the surveys covers a sample such as to allow the construction of a single database containing information relating to both income and consumption.

3.2 Harmonization of information from two archives

The first phase that is necessary to the accomplishment of statistical matching is the so called harmonization phase of the data from the two archives.

This is an essential process to verify the real comparability of the surveys.

The separate steps involved in this harmonization process can be summarized as follows:

- Analysis of the input sources. The input sources from the separate archives are analyzed in terms of the principle data acquired completeness and validity.
- Selection of stratification variables that characterize the families (number of family members, working condition).
- Selection of common variables (sex, age groups, marital status, tax-and-bills-payment difficulties, school and clothing subsidies, possession of dishwashers, washing machines and refrigerator).
- Selection of integration variables from the Eu-Silc (ability to make ends meet each month; possibility to spend a one-week vacation out of home; saving ability and loan application) or from the Lifestyle (perception on the progress of the family's economic situation; potentially required monthly income to live without economic problems).
- Harmonization of the common variables. The selected common variables are standardized in terms of classification. The codes of one of the two sources are then inverted.

At the end of this procedure the first part of the two harmonized archives will contain the information that is common to both, and the second part the data of each source.

3.3 Integrated archive results

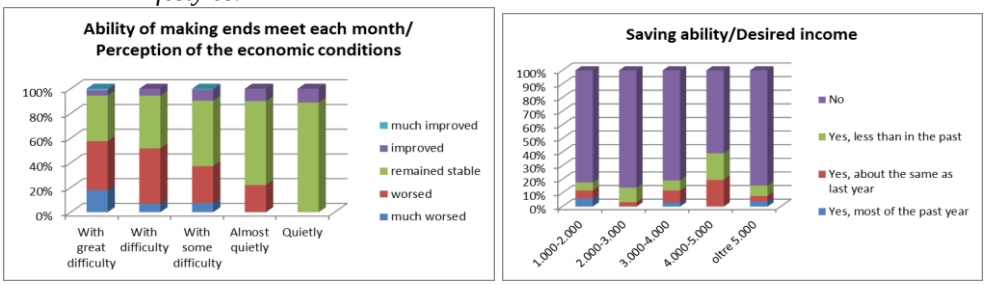
The study case concerns two simulations where the definition of the threshold value for the *distance function*, in the first case is set to be very restrictive, equal to zero (i.e. considering 5 matching variables only the matches that had a distance equal to zero were selected) while in the second case the value is set to be equal to 0,1 (i.e. considering 10 matching variables only those matches with a distance less or equal to one were selected).

The final result of elaborations was the integrated archive that derives from a combination of two surveys and contains *311 records in the first case and 365 matched records in the second one*.

The cross examination of results between the integration variables, that are the variables observed only in the Eu-Silc or only in the Life Styles, appears to be particularly interesting. In particular, comparing the data related to the "perception of the economic conditions" from the Life Styles with the one related to the "ability of making ends meet each month" from the Eu-Silc, it follows that there is a certain percentage of families which, although it makes ends meet each month without difficulties, perceives that its economic condition is deteriorated over the past year.

In addition, comparing the data on the "saving ability" Eu-Silc with the "desired income" Life Style, it follows that the saving ability is common only to those families that have a desired income above 4.000 euro per month.

Figure 1 – *Results from the combination of matching variables between the Eu-Silc and Lifestyles.*



4. Construction of hardship profiles

4.1 The cluster analysis

The next step in the integrated archive analysis was to consider a clustering procedure that will identify some poverty profiles, not previously defined, to which each family with socio-economic attitudes derived from the matching of two archives can be assigned. The cluster analysis proves to be very convenient since it provides "quite different" clusters (i.e. heterogeneous) among themselves where each one of them is composed of units (families) with a high grade of "natural association"

The variety of approaches to the cluster analysis have though a common necessity of defining one dissimilarity or distance matrix between the n observation couples, that represents a point from which every algorithm is being generated. The cluster analysis technique chosen is the one defined as the Two-Step. It is an extension of distance measures used by Banfield and Raftery (1993) based on the model and introduced for data with continuous attributes. The Two-Step algorithm has two advantages: it deals with mixed type of variables and automatically determines the optimal number of clusters, although it allows us to set a desired number of clusters.

The Two-Step procedure, very efficient for large datasets, is a scalar cluster analysis algorithm and is able to treat simultaneously continuous and categorical variables or attributes. It is being solved in two steps: in the first step, defined pre-clustering, the records are pre-classified in many small sub-cluster; in the second step the sub-clusters (generated in the first step) are grouped in the number of clusters that optimizes the BIC (Bayesion Information Criterion) defined as:

$$BIC_K = -2l_k + r_k \log n \qquad (2)$$

where r_k is the number of independent parameters and:

$$l_k = \sum_{v=1}^{k} \xi_v \tag{3}$$

is the function of *log-likelihood*, for the step with *k* clusters, which can be interpreted as the dispersion within the cluster. It also represents the entropy within the *k* clusters in the case in which only the categorical variables are considered.

4.2 Identified profiles

The cluster analysis allowed us to identify several profiles of families derived from the integrated archive Eu-Silc with the Life Style.

By running the cluster analysis on the integrated archive, 5 clusters to which different profiles of surveyed families are associated have emerged. The most important variable in the definition of the profiles resulted to be the perception of the economic situation progress. In particular:

- Cluster 1: families with more than 5 members, which claim that their situation has remained stable over the time or has improved, have no kind of debts nor difficulties paying taxes, bills, school subsidies;
- Cluster 2: families that perceive a deterioration of their economic situation over the past year, have difficulties paying taxes, bills, school subsidies and have other kinds of debts;
- Cluster 3: families with less than 5 members, which claim that their situation has remained stable over the time or has improved, have no kind of debts nor difficulties paying taxes, bills, school subsidies;
- Cluster 4: families that perceive a deterioration of their economic situation over the past year, but have no kind of debts nor difficulties paying taxes, bills, school subsidies;
- Cluster 5: families that perceive a deterioration of their economic situation over the past year, have difficulties paying taxes, bills, school subsidies, but do not have other kinds of debts.

5. Concluding remarks

Through our analysis we have attempted to quantify the influence of income and of family typology (number of members) in order to understand how the lifestyles of the families evolve. The estimates of the risk of poverty based on "objective" indicators, such as income or recourse to going into debt, are completely independent of the state of awareness of those directly involved. It is, however, also useful to observe the "subjective" perception of Italian people in relation to their standard of living and to the recurring causes of economic and social hardship.

It is hoped that the variations regarding the new family profiles that emerge in general from analyses carried out with different criteria provide important suggestions not only for describing and understanding the phenomenon of economic hardship better, but also to obtain indications for the social policies to contrast poverty.

References

BANFIELD J.D. RAFTERY A.E. 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics, 49, p. 803-821.OKNER B.A. (1972), "Constructing a new data base from existing micro data sets: the 1966 merge file", Annals of Economic and Social Measurement, 1.

D'ORAZIO M., DI ZIO M., SCANU M. 2002. Statistical Matching and Official Statistics, Quaderni di Ricerca ISTAT, 1.

MONTRONE S., PERCHINUNNO P. 2005. La stima della povertà basata su modelli, Annali del Dipartimento di Scienze Statistiche, Facoltà di Economia, Università degli Studi di Bari, n. 4.

RUGGLES N. N., RUGGLES R., 1974. A strategy for merging and matching micro data sets, Annals of Economic and Social Measurement, 3.

SCHOIER G., TORELLI N., ZACCHIGNA A., EGIDI V., SABBADINI L.L. 2006. L'abbinamento statistico di dati dal sistema di indagini multiscopo: prime proposte e evidenze empiriche, in "Metodi statistici per l'integrazione di dati da fonti diverse", a cura di Liseo, Montanari, Torelli, Ed. Franco Angeli.

SUMMARY

The present work will describe a model of data integration through a methodology of Statistical Matching (hot deck distance) for the integration of two surveys (EuSilc-Istat and Lifestyle Survey-University of Bari). The construction of an integrated database on the basis of these two surveys may be useful for the study of consumer behavior in relation to specific groups of commodities, in order to analyze the decisions taken by families with regard to saving, to examine economic and social inequality, and to study the impact of public policies by means of simulations.

Silvestro MONTRONE, Università degli Studi di Bari silvestro.montrone@uniba.it
Antonella MASSARI, Università degli Studi di Bari antonella.massari@uniba.it
Paola PERCHINUNNO, Università degli Studi di Bari
paola.perchinunno@uniba.it
Stefania GIRONE, Università degli Studi di Bari stefaniagirone@libero.it