

# Integration of Administrative Registers and Statistical Archives.

## The Case of the Eurostat Structure of Earning Survey in Italy.

Stefania Cardinaleschi, Vincenzo Spinelli

ISTAT, Italy's National Institute of Statistics,

via Tuscolana 1788, 00173 - Rome, Italy

[cardinal@istat.it](mailto:cardinal@istat.it), [vispinel@istat.it](mailto:vispinel@istat.it) )

### Abstract

The *Structure of Earnings Survey* (SES) is a series of four-yearly earnings surveys to be conducted under the Council Regulation 530/1999 and latter Regulation. In this work we describe the statistical and methodological steps we made to define the samples of employers and employees for SES 2006 in the public sector for two economic activities according a 2-digit NACE Rev. 1.1. code, i.e., R80 (*Education*) and R85 (*Health*). The focus of this work is on the data integration problems and the solutions we adopted in this context. The project, we described in this work, has two important points worth being underlined: successful appliance of Probabilistic Record Linkage techniques for the integration of both Administrative sources and Statistical ones, and supplying Official Statistics for sectors R80 and R85 without using ad hoc surveys. The joint use of administrative and statistical sources allow us (a) to get the best definitions of the universe for the sectors R80 and R85, and (b) the determination of all the mandatory variables as requested by SES for both public and private sectors.

**Keywords:** Integration of multiple data sources, Data Editing, Use of registers.

## 1. Introduction

The *Structure of Earnings Survey* (SES) 2006 is the second of a series of four-yearly earnings surveys to be conducted under the Council Regulation 530/1999 and latter Regulation, see Eurostat (2007) and Giaccone (2007). The objective of this survey is to provide accurate and harmonized data on earnings in EU Member States. Statistics on earnings and labor costs describe the price dimension of labor input, making available information that support EU level policies in the fields of employment as well as macro-economics economic and monetary policy.

Since the beginning of this survey, Italy provides data for private control enterprises in all the NACE sectors, but Agriculture, by ad hoc electronic survey form. Only for SES 2006 we provided data for public control enterprises but only for R80 (*Education*) and R85 (*Health*) sectors.

In this work we describe the statistical and methodological steps we made to define the samples of employers and employees belonging to in the public sector domain and included in two economic sector identified at 2-digit level of NACE Rev. 1.1 classification, i.e., R80 (Education) and R85(Health). In this context, samples are defined as multi-stage samples, such that the primary units (first stage) are the so-called local units, i.e. second level organization units of the employees, and the secondary units (second stage) are the employers. The variables to be measured for

each units are classified as mandatory or optional by Eurostat<sup>1</sup>. There is not any single archive where all the mandatory variables are gathered together and we needed to collect, harmonize and integrate more than one archive to get the outcome required.

Furthermore, the archives we considered are both administrative registers and statistical archives, and this implies that (a) the archives are really heterogeneous (e.g. for size, structure, and reference period), and (b) the integration can not be straightforward because the statistical units and/or the key variables are not always the same in all the available se sources. Indeed, there are two complementary aspects to consider: a good quality administrative source has considerable benefits as *a sample frame* for many surveys, and, at the same time, a survey *may integrate* data drawn from different administrative sources.

Finally, we observe that the higher the level of integration of the administrative archives the greater is their quality and statistical interest. The focus of this paper is on the integration problems and the solutions we defined and implemented in this context.

## 2. Legal framework and objectives of the survey

In the framework of the European System on Earnings and Labor Cost statistics, SES is a four-yearly survey whose objective is to provide accurate and harmonized data on earnings in the EU Member States for policy-making and research purposes. It is covered by several legal acts, the main one being the Council Regulation 530/992, followed by the implementation of Commission Regulations 1738/2005 (amending Commission Regulation 1916/2000) on definition and transmission of information and Commission Regulation 698/2006 on quality evaluation. More recently SES has been included into Commission Regulation 831/02 on micro data release for scientific purposes.

The objective of SES is to monitor the structure and distribution of earnings.

The survey provides information on earnings taking into account job-related factors, such as working hours, principal economic activity, size of the enterprise and location, and individual characteristics of employees such as gender, age, occupation, education, length of stay in service and others. In order to improve the performance of the entire survey process in terms of quality, timeliness and enterprise statistical burden reduction, in 1999 Istat coordinated a Commission of Experts to study the information value of the SES (2002); the output of the Commission of Experts consisted of a tailor-made questionnaire containing information on flexibility of the labor market and collective pay agreements. SES refers to year 2006 for annual variables and to October 2006 for monthly data in line with Commission Regulation 1916/2000.

The Italian Structure of Earnings Survey has a two stage sampling design: a sample of employees in a sample of enterprises. The target enterprise population has been divided into two strata: small-medium size enterprises (from 10 to 249 employees) and large size enterprises (with more than 250 employees). As far as small-medium enterprises are concerned, the first stage units consisted of a stratified sample of enterprises from 10 to 249 employees, while the second stage units consisted of simple random sampling of employees. For large enterprises, a total survey was carried out at the first stage, followed by a simple random sampling of

---

<sup>1</sup> See Annex 1.

employees at second stage. Finally, it worth noticing that the number of employers to be considered per each employee is linked to the dimension of the same employee, and this number ranges from 10 to 200.

### 3. Data Context

All the data sets we have considered can be classified as *Administrative registers* (AR) or *Statistical archives* (SA). The use of administrative data is being pursued increasingly by Official Statistics throughout the world, not least because of continuing budgetary pressures to find less expensive ways of collecting data, but also as a way to improve the statistical data production by considering the potential for business registers to be used as data sources in their own right, see Rotondi and Spinelli (2008). In this project we have mainly consider three administrative sources: the *770 Form tax register* (770 Form) of the "Agenzia delle Entrate", see Spinelli (2007), for activities included in sectors R80 and R85: the data sets maintained by *Cineca*, a non profit Consortium ([www.cineca.it](http://www.cineca.it)), for university employment, the Payroll data sets by of the "Ministero dell'Istruzione, dell'Università e della Ricerca", see [www.pubblica.istruzione.it](http://www.pubblica.istruzione.it).

The last two data sets were considered only as for the R80 sector. The definition of the universe of units included in the R80 domain is mainly based on these three sources, while for the R85 sector we used the so-called S13 list i.e. the list of units included by Istat in the S13 domain that is the "General government" institutional sector in *European System of National and Regional Accounts*, see ESA (1995).

Furthermore, all these data sets were integrated with those ones coming from two important surveys in Italy: the official *Labor Force Survey*, see LFS, and *EU-SILC Panel Survey*, see EU-SILC. From one side, LFS contains information on all the variables concerning the labor market status of 300 thousands Italian families and about 800 thousands individuals. Up to 2005 the survey was conducted quarterly (January, April, July and October) and the available microdata start in October 1992. The sample is a rotating panel of households: each sampled household remains in the sample for two quarters, than exits for two quarters and enters again for two final one before leaving the sample.

This structure allows reconstructing labor market flows on a quarterly and annual basis, but only the annual longitudinal structure is made available in the microdata. On the other side, EU-SILC is a panel survey, carried out in different EU member states, and providing every year cross-sectional and longitudinal data on income, poverty, social exclusion and living conditions.

Information is collected about both households and household members at the same time. Moreover, households and household members are followed and surveyed yearly, during four years. Given those characteristics, LFS and EU-SILC can be used to integrate the previous mentioned administrative sources both for the R80 sector and for R85 one.

### 4. Data Linkage

The integration of different and not homogeneous data sets implies that we need to solve a *Record linkage* problem (RL). Record linkage refers to the task of finding entries that refer to the same entity in two or more files. Record linkage is an

appropriate technique when the target is merging two or more data sets that do not have a unique database key in common. In this framework, records are linked on the basis of common data. Records from the two or more sources that are believed refer to the same individual are matched in such a way that they may then be treated as a single record for that individual.

A match in a deterministic linkage is made when a sufficient number of identifiers agree exactly between two records. In probabilistic record linkage, the comparison or matching algorithm yields for each record pair, a probability or “weight” which indicates the likelihood that record pairs relate to the same entity, see Winkler (2001).

Record linkage is a useful tool when performing data mining tasks, where the data originated from different sources or different organizations. Most commonly, performing RL on datasets involves joining records of persons based on name, when no "*National identification number*" or a similar code is recorded in the data. In our context the two administrative archives, i.e. 770 Form Register and Cineca one, are defined on key defined by two variables: the *Italian fiscal code* for employees and the *VAT code* for employers. The Italian fiscal code, officially known as Italy's *Codice Fiscale*, is the tax code in Italy; similar to the *Social Security Number* (SSN) in the United States. The tax code is an alphanumeric code of 16 characters, and it serves to identify, unambiguously for tax purposes, individuals resident in Italy. This code is defined on the surname, given name, sex, place of birth, county of birth, and date of birth. The VAT code is made of 11 numeric codes, the first seven of which identify the tax payer by a progressive number, the following three numbers identify the tax office and the last one acting as a control code.

In turn, the two statistical archives, i.e. LFS and EU-SILC, are defined on some internal key variables and, for this reason, we defined and implemented a RL algorithm. In literature, there are several approaches to record linkage. The most straightforward is a *rules-based approach*, in which reasonable rules are developed and then refined as common exceptions are found. The advantage to this approach is that it is possible to get a good level of accuracy without needing a lot of labeled data to train or test the rules on. The disadvantage is that to obtain very high accuracy, more and more exceptions and special cases would need to be handled, and eventually the list of rules gets too complex to be managed by hand. A very popular approach has been the *Probabilistic Record Linkage* (PRL).

In this approach, a large set of pairs of records are human-labeled as being matching or differing pairs. Then statistics are calculated from the agreement of fields on matching and differing records to determine weights on each field. During the process, the agreement or disagreement weight for each field is added to get a combined score that represents the probability that the records refer to the same entity. Often there is one threshold above which a pair is considered a match, and another threshold below which it is considered not to be a match. Between the two thresholds a pair is considered to be "*possibly a match*", and dealt with accordingly (e.g., reviewed by the operator and judged to be linked, or not linked, depending on the application). In recent years, a variety of machine learning techniques have been used in record linkage. It has been recognized that PRL is equivalent to the "*Naive Bayes*" algorithm in the field of machine learning, and suffers from the same assumption of the independence of its features, which is typically not true, see Rish (2001). Higher accuracy can often be achieved by using various other machine learning techniques, including a single-layer Perceptron, see Rojas (1996). Regardless of whether rule-based, PRL or machine learning techniques are used, normalization of the data is very important. Names are often spelled differently in different sources (e.g., "*Giovanni*

*Carlo Rossi*", "*Giovanni C. Rossi*", "*Giancarlo Rossi*", etc.), dates can be recorded various ways ("*2/1/73*", "*1973.1.2*", "*Jan 2, 1973*"), and places can be recorded differently. By normalizing these labels into a common format and using comparison techniques that handle additional variation, much more consistency can be achieved, resulting in higher accuracy in any record linkage technique.

In our algorithm we mainly used two steps, i.e. (a) normalization and (b) PRL approach, based on fitness or score function that has been defined as follows:

```

Function match(r1,r2)
{
    best = 0.0;
    n = number of distinct sequences of (normalized) values in r1;
    for( i=1 ; i≤n ; i++ ){
        k1 = normalization(r1,i);
        w = fitness(k1,r2.key);
        if(w>best) w = best;
    }
    return(best);
}

```

The function *match()* has two input parameters: *r1* is a record without fiscal code while *r2* has it, i.e. *r2.key*. The result of the function "*normalization(r,i)*" is the *i*-th feasible fiscal code we can get from record *r*. From these hypotheses, the function *match()* gives us the best accuracy we can get if the pair (*r1,r2*) is matched. On the top of this function, our PRL algorithm tries to match an administrative archive and a statistical one as best as possible.

The algorithm we defined can be seen as a main loop of two basic steps: (a) automatic assignment of a combined weight per each matched pair, and (b) manual decision of the accepted pairs, i.e. we set the best threshold value by iteration. Our assumption was that pairs for which a one-to-one relationship was obtained, and a best-link was found with the highest combined weight would be considered as univocally matched pairs and should then provide information in order to decide about pairs in which such a relationship could not be established. For example, we observed that for the unequivocally matched pairs a clear and expected relationship between differences sex, place of birth, county of birth, and date of birth could be assessed.

As a result, such a relationship was used to help solving the remaining pairs for which a one-to-one relationship could not be found. Indeed, we reduced the number of non-uniquely matched records.

## 5. Conclusions

The project, we partially described in this work, has two important aspects being underlined: (a) successfully applying of PRL techniques for the integration of Administrative sources and Statistical ones, and (b) compiling wage and labor cost statistics for sectors R80 and R85 without using an additional ad hoc surveys devoted to this task.

The activity developed under the first point clarified the direction of future works in this project: the procedure, we shortly described, needs to be generalized and

integrated with standard software libraries for PRL techniques; in other words, we will transform this procedure from the actual prototype to a standard tool into the statistical production line.

The second point refers to a more general context, i.e., the “*reduction of the response burden*”. In the “*Statistical Programme of the European Commission for the Year 2008*” the Commission adopted a Communication to the European Parliament and the Council on this *reduction, simplification and priority-setting in the field of Community statistics*. A combination of administrative data for extrapolation supported by regular benchmark surveys or low-frequency surveys would be a strategy for reducing the response burden and costs. Moreover, existing administrative sources will be preferred in this project wherever possible, and facilitating the use of administrative data will reduce the burden on respondents.

Finally the joint use of administrative and statistical sources allow us (a) to get the best definitions of the universe for the sectors R80 and R85, and (b) the compilation of all the mandatory variables for Eurostat SES survey, as shown in ANNEX I.

## References

- Cardinaleschi S. (2002). Dipendenti, ore lavorate e retribuzioni nelle imprese dell'industria e dei servizi – Anno 2002. Statistica in Breve- Structure of Earnings Survey 2002.
- Cardinaleschi S. (2006). Rilevazione sulla struttura delle retribuzioni – Anno 2006. Statistica in Breve- Structure of Earnings Survey 2006. (to be published)
- Chronos data base by Eurostat.  
[http://www.esds.ac.uk/international/support/user\\_guides/eurostat/cronos.asp](http://www.esds.ac.uk/international/support/user_guides/eurostat/cronos.asp)
- Eurostat (2007). Structure of Earning Survey 2006: implementation arrangements. Eurostat Working Group Labour Market Statistics. Eurostat/F2/LAMAS/41/07.
- ESA 1995. European system of accounts.  
[circa.europa.eu/irc/dsis/nfaccount/info/data/ESA95/ESA95-new.html](http://circa.europa.eu/irc/dsis/nfaccount/info/data/ESA95/ESA95-new.html)
- EU-SILC. European Survey of Income and Living Conditions.  
[www.istat.it/strumenti/rispondenti/indagini/famiglia\\_societa/eusilc](http://www.istat.it/strumenti/rispondenti/indagini/famiglia_societa/eusilc)
- Giaccone, M. (2007). Annual review of working conditions in the EU 2006-2007. *European Foundation for the Improvement of Living and Working Conditions*.
- LFS. Labor Force Survey. [www.istat.it/lavoro/lavret/forzedilavoro](http://www.istat.it/lavoro/lavret/forzedilavoro)
- Palmieri A. “Structure of Earnings Survey for the year 2002 – Quality report”.  
[http://circa.europa.eu/Members/irc/dsis/wages/library?l=/4snationalsmetadatasands/structure\\_earnings/10\\_italy&vm=detailed&sb=Title](http://circa.europa.eu/Members/irc/dsis/wages/library?l=/4snationalsmetadatasands/structure_earnings/10_italy&vm=detailed&sb=Title)
- Rish, I. (2001). An empirical study of the naïve Bayes classifier. *IBM Research Division. RC 22230 (W0111-014)*.
- Rojas, R. (1996). Neural Networks – A Systematic Introduction. *Springer-Verlag, Berlin, New-York, 1996*.
- Rotondi G., Spinelli, V. (2008). The Roadmap for a “Central Hub of Administrative Data”: Design and First Results on the Pilot Case of 770-Form Business Register. Q2008 – *European Conference on Quality in Official Statistics. July 8-11 2008, Rome, Italy*
- Spinelli, V. (2007). Processo di Acquisizione e Trattamento Informatico degli Archivi relative al Modello di Dichiarazione 770. (Data Collection and Automatic Treatment of the 770 Form Tax Register Archives). *Documenti ISTAT n. 4/2007*.  
[www.istat.it/dati/pubbsci/documenti/documenti2007.html](http://www.istat.it/dati/pubbsci/documenti/documenti2007.html).
- Winkler, W.E. (2001). Record Linkage Software and Methods for Merging Administrative Lists. *U.S. Bureau of the Census. Statistical Research Report Series, No. RR2001/03*.

## **ANNEX I - LIST OF VARIABLES**

### **1. Information about the local unit to which the sampled employees are attached**

- 1.1. Geographical location of the local unit (NUTS-1)
- 1.2. Size of the enterprise to which the local unit belongs
- 1.3. Principal economic activity of the local unit (NACE Rev. 1.1.)
- 1.4. Form of economic and financial control
- 1.5. Collective pay agreement
- 1.6. Total number of employees in the local unit in the reference month (*optional*)
- 1.7. Affiliation of the local unit to a group of enterprises (*optional*)

### **2. Information on individual characteristics of each employee in the sample relating to the reference month**

- 2.1. Sex
- 2.2. Age
- 2.3. Occupation (ISCO-88 (COM))
- 2.4. Managerial or supervisory position (*optional*)
- 2.5. Highest successfully completed level of education and training (ISCED 97)
- 2.6. Length of service in the enterprise
- 2.7. Contractual working time (full-time or part-time)
  - 2.7.1. Share of a full-timer's normal hours
- 2.8. Type of employment contract
- 2.9. Citizenship (*optional*)

### **3. Information on working periods for each employee in the sample**

- 3.1. Number of weeks in the reference year to which the gross annual earnings relate
- 3.2. Number of hours paid during the reference month
  - 3.2.1. Number of overtime hours paid in the reference month
- 3.3. Annual days of holiday leave
- 3.4. Other annual days of paid absence (*optional*)

### **4. Information on earnings for each employee in the sample**

- 4.1. Gross annual earnings in the reference year
  - 4.1.1. Annual bonuses and allowances not paid in each pay period
  - 4.1.2. Annual payments in kind (*optional*)
- 4.2. Gross earnings in the reference month
  - 4.2.1. Earnings related to overtime
  - 4.2.2. Special payments for shift work
  - 4.2.3. Compulsory social contributions and taxes paid by the employer on behalf of the employee (*optional*)
    - 4.2.3.1. Compulsory social-security contributions (*optional*)
    - 4.2.3.2. Taxes (*optional*)
- 4.3. Average gross hourly earnings in the reference month

### **5. Grossing-up factors**

- 5.1. Grossing-up factor for the local unit
- 5.2. Grossing-up factor for the employees