Università di Pisa
Dipartimento di Statistica e Matematica
Applicata all'Economia

Report n. 349

Matching Eusilc and administrative data for studying poverty and social exclusion at the local level

Alessandra Coli, Paolo Consolini, Marcello D'Orazio

Pisa, 27 settembre 2011
- Stampato in Proprio –

Via Cosimo Ridolfi, 10 - 56124 PISA - Tel. Segr. Amm. 050 2216231 Segr. Stud. 050 2216317 Fax 050 2216375
Cod. Fisc. 80003670504 - P. IVA 00286820501 - Web http://statmat.ec.unipi.it/

Matching Eusilc and administrative data for studying poverty and social exclusion at the local level

Alessandra Coli[1], Paolo Consolini[2], Marcello D'Orazio[3]

[1]University of Pisa, e-mail: a.coli@ec.unipi.it
[2]Italian Institute of Statistics (Istat), e-mail: consolin@istat.it
[2]Italian Institute of Statistics (Istat), e-mail: madorazi@istat.it

Abstract

Local government needs reliable statistics in order to address policies and actions against poverty and social exclusion. Generally, local government agencies have a huge amount of administrative data which can help calculating such statistics. Moreover local non profit organizations collect useful data to monitor their actions. However such administrative data are seldom integrated in order to convey a synthetic view on poverty, deprivation and social exclusion.
In this paper we try to match survey and administrative data related to the Italian province of Pisa to build a richer database on poverty and social exclusion at local level.

Keywords: data integration, record linkage, exact matching

1. Introduction

This paper describes the data integration procedures set up in order to match sample and administrative data at local level. The application concerns the Pisa province, taking advantage of the significant amount of data made available by the SAMPLE project (European Commission 7th Framework Programme). In particular, on the one hand the data collection benefited from a PI-Silc[1] survey oversampling, on the other it was possible to gain access to administrative databases not frequently used for statistical analysis.
Unfortunately, due to confidentiality issues, we could actually access less than the expected amount of administrative data sets. Moreover some demographic variables were removed form such data sets making our task even more complicated.
The available administrative databases were: i) The Caritas[2] database ii) The Revenue Agency database (RA) ; iii) the provincial Job Centre database (JC).
The Caritas database provides information about the people asking for Caritas social services. Caritas provides support to all asking people independently from the geographical area they live in. Besides, all these people are hardly detectable in the Silc sample being more often homeless people. Due to these reasons the attempt to match the Silc data with the Caritas ones resulted in very few matched units. As a consequence the Caritas data were discarded and only the RA and JC databases were used in the integration procedures.

[1] It-Silc are the Italian Statistics on Income and Living Conditions (Eusilc at European level). In the paper we call PI-Silc the statistics concerning the province of Pisa.
[2] Caritas is a catholic agency which provides assistance to poor, vulnerable and excluded people.

In this paper we describe the complex integration procedures developed to provide the Silc-RA and Silc-JC matched data sets. The final objective is extending the Silc records with variables taken from the RA and JC databases.
Among the main advantages, there is the possibility of grossing up administrative indicators using Silc sample weights. Additionally, the matched data sets allow to give interpretation of administrative indicators taking into account the characteristics of the households whom the individuals belong to. On the contrary, relying on RA and JC indicators only, individuals may be classified as poor or emarginated, in spite of actually living in a wealthy family.

2. The Data

IT-Silc is the best Italian data source for investigating poverty and social exclusion. These statistics cover a large spectrum of aspects (such as housing, health, education, labour, income; see Eurostat, 2009). Data are collected both on individuals and households. As in most of the European countries, the Italian Silc is a system of statistics based both on a sample survey and on administrative information. In particular, in Italy, sample data on income are checked, integrated and sometimes corrected taking into account the evidence stemming from tax registers. Despite its richness, Silc does not allow to estimate accurate indicators at local level. With the SAMPLE project, however, it was possible to enlarge the Pisa province sample from 162 to 818 households for the 2008 wave. The purpose of the oversampling was threefold: getting direct estimates of poverty and social exclusion indicators for the Pisa Province and sub-areas thereof; improving the small area methodology; getting a larger set of units to be linked or matched with local administrative registers.

The Revenue Agency database contains data from taxpayers declarations, concerning the amount and type of earned income. For our purposes we concentrate on individual's declarations only, in particular on tax declarations submitted by employees, pensioners, temporary workers, cooperative workers (the 730 tax returns register) and tax declaration relating mainly self-employed income (Unico p.f. tax returns register). Employees and pensioners can decide not to submit the 730 declaration. In principle it could be possible to trace information on this segment of taxpayers in the declarations submitted by their employers (770 tax return register). Unfortunately we were not able to gain access to this register and therefore we had to restrict our analysis on a subset of taxpayers. The RA archives contain personal data such as the birth date, the residence and the gender of the taxpayer. It is worth stressing that this data source does not cover people who, though earning income, do not present any tax declaration. This may happen for fiscal evasion purposes or because income is lower than the minimum taxable level. Fiscal data were referred to the year 2007. The JC database depicts the local labor market through stock and flows variables.

The job centre updates the database each time a person crosses the labor market: when searching a job, being hired or fired. As a consequence the archive units may be unemployed people but also employed people searching for a new job. When accessing the job centre, people are asked lots of social and demographic information as the date of birth, the education level and the professional status. For employed people moreover it is possible to know the economic sectors they work for, the kind of labor contract etc. The JC data source does not cover people who do not contact the job centre for searching a job. It is worth stressing the different meaning of unemployed people with respect to the official statistics definitions. According to the JC unemployed are people who do not work (or they work earning an income lower than the taxable level) and declare to be disposable to start a job immediately.

3. Integrating Silc data with administrative databases

The RA and JC data sources seem to cover populations similar to the Silc population. As a consequence we expect to find some of the Pi-Silc sampled individuals in the administrative archives. The exact matching (or record linkage) is the technique used to identify and pick up such units.

Record linkage is a technique which compares records contained in two files A and B, in order to determine pairs of records referred to the same population unit. The A and B files are supposed to contain identical units that have to be found according to an identifier (like the social security number) or a set of identifying variables (k variables) present in both files.

Record linkage between two files is very simple provided that each record in both files contains the same identifier and this identifier is recorded without errors. In this case the problem is solved by simply picking out the records (if any) with the same identifier value. This procedure is known as exact matching.

Unfortunately, some complications may occur (Copas and Hilton 1990): (i) Errors may occur because incorrect information is obtained from the individual, or because information is incorrectly recorded. Due to such errors two records for the same person may not agree, and two records which agree may refer to different people. (ii) Some values of the k variables may be missing so that the k-variable may not be known exactly for some of the records in A or B.

Formalizing the linking procedure into a statistical model, it is possible to evaluate the matching by measuring the probability of generating false-matched-pairs and false-unmatched pairs.

Coming to our application, we can state the problem as follows. The Pisa-Silc contains N records, one for each of the N interviewed individuals. On the other hand the RA (or JC) archive contains M records one for each registered subject. Given a set of common variables (k-variables) we have to evaluate the evidence that the i-th record from Silc and the j-th record from RA (or JC) relate to the same person.

Table 1 shows the personal items eligible to be used as k-variables in the three data sources. We dispose of a similar set of variables for PI-Silc and JC data. On the contrary, we cannot access neither the birth month, nor the census enumeration area in the RA data source.

Formalizing the linking procedure into a statistical model, it is possible to evaluate the matching by measuring the probability of generating false-matched-pairs and false-unmatched pairs (Fellegi and Sunter 1969).

Table 1 Personal items to be used as k-variables, in the PI-Silc, RA and JC data sources

Personal item	Data source		
	PI-Silc	RA	JC
Birth day			✓
Birth month	✓		✓
Birth year	✓	✓	✓
Gender	✓	✓	✓
Place of birth (municipality)	✓	✓	✓
Place of residence (municipality)	✓	✓	✓
Place of residence (census enumeration area)	✓		✓
Nationality	✓	✓	✓

3.1 Integrating Silc with the local Revenue agency archive

The PI-Silc and RA integration is particularly difficult because of the lack of identifiers on the RA side (see Tab. 1). In particular the lack of the "birth month" and "Census Enumeration Area" led us to develop a complex integration procedure. Before describing the procedure, it is worth mentioning that revenue database consisted in data from the 730 and the Unico p.f. tax returns registers (see § 2). The first register contains data on employees and pensioners income, whereas the second collects mainly information on self-employed people.

- Step 1: The matching of 730 register and PI-Silc

a. Deterministic matching based on demographic variables
At this stage a pair of records is considered to be a correct link if the two records agree exactly on each element within a collection of identifiers that represent the match key. As already pointed out, available data allow to build a match key made of the following personal items: "gender", "birth year", "birthplace", "place of residence" and "nationality". Due to the weakness of the linking variables we decided to apply the matching on a 730 register subset only, i.e. taxpayers with a joint tax declaration[3]. Thus a 730 unit is linked to a PI-Silc unit if both these conditions are met: i) the compared units share exactly the same match-key value ii) the compared units belong to the same family (i.e, the 730 couple records the same mach-key values as the corresponding PI-Silc couple).
On the one hand the PI-Silc data set accounts for 1246[4] units. On the other hand, the 730 register includes 84.942 taxpayers with single tax return statements and 30.766 with joint tax returns (15.383 couples). On the basis of the integration procedure above described, it was possible to link 136 subjects out of 1246 in the PI-Silc dataset. If we consider the first condition (match key variables) only, we realized that 342 units out of 1246 can not find a link. These units are then definitely dropped down.

b. Deterministic matching based on demographic and income related variables
In order to find a link for the remaining 768 units we run a further deterministic matching taking into account income related information. At first, the mach-key variables are used to classify the PI-Silc and 730 units into homogeneous strata where the new matching procedure is independently run. This means that, for each PI-Silc unit the link is searched among the administrative units sharing the same "gender", "birth year", "birthplace", "place of residence" and "nationality". At this point we take into account two different income related information: the type of income (employed income, pensions) and the level of gross and net income[5]. The matched unit is defined as the administrative unit which earns the same kind of income and whose income level differs from the PI-Silc income for less than 100 euro. When more than one unit was detected, the one with the lowest distance in terms of gross income was chosen.
Applying the two-step procedure above described we matched 369 PI-Silc units.

- Step 2 The matching of Unico p.f. and PI-Silc data sets

We replicated the procedure described at point b) on all the records present in the PI-Silc data set (1476 records) in order to recover employees or pensioners who submitted the Unico p.f. declaration instead of the 730 tax revenue declaration. Through this procedure we identified 34 more links.

---

[3] Italian tax system allows couples to fill tax returns jointly, in this case the tax information of each member is reported on a distinct record.
[4] Actually the PI-Silc data set accounts for 1476 individuals aged 15 or more. Before running matching however, we removed the PI-Silc units that report cask benefits or losses from self-employment, since the Italian tax system obliges the self-employed to fill "Unico p.f. tax returns" instead of 730 tax return.
[5] A "harmonized definition of income" is actually employed. as a basic requirement to link the PI-Silc and 730 units. The harmonization process is quite complex and demanding. A fully detailed illustration goes beyond th scope of this paper.

As a result the final matched file contains 403 matched units out of 1476. The matched data set contains the Silc variables as well as data coming from the RA database (730 and Unico p.f registers). Furthermore a linking probability is provided which helps evaluating the quality of the matching. This probability is an estimate of the probability that the coupled units represent a true link (the two record refer to the same unit) giving the observed values for the matching variables. Note that these probabilities are estimated by using the EM algorithm under the assumption of conditional independence. In practice, it is considered the vector $\gamma_{a,b}$ resulting by the comparison of the observed values for the matching variables on record a and record b; $\gamma_{a,b}^{(j)} = 1$ if a and b show the same value for the variable j and 0 otherwise. The conditional independence assumption means that $\gamma_{a,b}^{(j)}$ is independent from $\gamma_{a,b}^{(k)}$ given that a and b are a true link. The values of the variables are being compared by using the method proposed by Jaro and Winkler (cf. Winkler, 1988). The probabilities are estimated by using the code made available in the package "RecordLinkage" (Borg and Sariyar, 2010) freely available for the R environment (R Development Core Team, 2010).

3.2 Integrating PI-Silc with the local Job Centre archive

The matching procedure is split in two separate procedures depending on the data available in the labour market database. In general, deterministic matching is applied for those units in database which present information concerning the census Enumeration Area (EA) in which they live; when this information is missing a probabilistic record linkage is applied. Note that the information concerning the census EA was not available for all the responding units at the Eusilc survey.
The exact matching procedure is straightforward. For each unit in the Eusilc survey is linked with the corresponding unit in the labour market database sharing the same information as far as the following variables are concerned: municipality, census EA, gender, birth month, birth year and a variable summarizing information about the birth place (country and NUTS3). Due to computational constraints the variable concerning the living municipality is used as a blocking variable (search is restricted to units living in the same Municipality). The units not linked in the exact matching phase are processed and linked in the record linkage phase. In practice units sharing the same values for the subset of the matching variables that not present missing values. In particular, the variables municipality, gender, birth month and year were available for all the units while some units had information missing for the census EA or for the birth place. In this step, again the Municipality is used as a blocking variable.
The first matching step based on exact linkage allowed to identify 529 couples of units corresponding. The second integration step allowed to find some other 404 couples of units. Note that for each linked couple of units it is estimated a probability of the linking quality. This probability is estimated with the same procedure used for the integration of PI-Silc with the Revenue agency archive.

4. Conclusions

The aim of this paper is to integrate data from different databases relating the Italian province of Pisa: sample data from PI-Silc, fiscal data from the tax revenue agency and labour data from the provincial job centre. The ultimate goal is to build a matched database having PI-Silc as the core dataset and the linked administrative data sets as satellites for in depth analysis on specific aspects (labour, income, taxes). The integration benefits from the Silc oversampling for the province of Pisa, made available for the 2008 wave by the SAMPLE research project (European Commission 7th Framework Programme).
The Silc and administrative data sets have been integrated using exact matching and record linkage procedures. Finally, only a limited subset of records is successfully linked: about 27% of the PI-Silc

units find a link in the RA register, about 63% in the JC register. This not entirely satisfying result is due to several reasons which can be summarized as follows:

- The Silc oversampling does not include all municipalities of the Province of Pisa (only 25 municipalities out of 39 have been involved);
- The administrative data sources cover only sub-populations of Silc (under coverage) i.e. part of the taxpayers (RA register) and people asking for the local job center services (JC register)
- There are errors and missing values in the variables used as identifiers; for example the JC register seems to contain outdated information on addresses, which lead to wrong census enumeration area.
- Data sources (RA in particular) contain only "weak" identifiers.

The number of matched records depends significantly on the quality of the administrative data sets and on the possibility of using more detailed identifiers. The work is still in progress, we hope to improve the matching getting more complete data at least for the RA register.
Our future analysis will be devoted to the evaluation the matching quality comparing distributions in the original and matched data sets. Subsequently we will provide indicators based both on Silc and the matched administrative data, with special focus on labour-related indicators.

References

Borg A., and Sariyar M. (2010). RecordLinkage: Record Linkage in R. R package version 0.2-3. http://CRAN.R-project.org/package=RecordLinkage

Copas J. B., F. J. Hilton (1990) Record Linkage: Statistical Models for Matching Computer Records, *Journal of the Royal Statistical Society. Series A* (Statistics in Society), Vol. 153, No. 3 (1990), pp. 287-320.

Eurostat (2009). Cross Sectional UDB (User Data Base). Version 2007-2 from 01-08-09

Fellegi I. P., A. B. Sunter. (1969). A theory for record linkage, *Journal of the American statistical association*. vol. 64, pp. 1183-1210.

Eurostat (2009). Cross Sectional UDB (User Data Base). Version 2007-2 from 01-08-09

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

Sample project (2011) Deliverables 7 "Oversampling Description", Deliverable 9 "Integrating data model – first release", URL http://www.sample-project.eu

Winkler W. E. (1988) "Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage". *Proceedings of the Section on Survey Research Methods, American Statistical Association* 1988, pp. 667–671.

Acknowledgments: this work is funded by the European Commission 7th Framework Programme (www.sample-project.eu) and by the Italian Ministry of Education (PRIN research project "Inferenza con informazione ausiliaria carente: il campionamento da popolazioni elusive e la stima per domini non pianificati, PRIN code 2007RHFBB3_003, 2007).