**Design and evaluation of an editing and imputation strategy for micro-data from integrated administrative sources: the Italian case of the ARCHIMEDE Project**

Sara Casacci, Italian National Institute of Statistics, casacci@istat.it

Romina Filippini, Italian National Institute of Statistics, filippini@istat.it

**Abstract**

*In order to provide users with detailed statistical information at local level, in recent years the Italian National Institute of Statistics (Istat) has made available micro data collections based on the integration of several administrative sources. In particular, micro data archives have been produced in the context of the ARCHIMEDE project (Integrated Archives of Economic and Demographic Microdata) to enlarge the offer of statistical information on the households socio-economic conditions. Extracting reliable statistical information from multiple data sources is in general a complex task. The present work describes the methodology adopted to ensure more "complete" and "coherent" data dissemination and provides indications regarding the quality of the results produced. Specifically, the different editing and imputation techniques and tools used for the main variables in the database are illustrated. A specific focus concerns the consistency between the two quantitative variables “Earned income” (from fiscal sources) and “Work Intensity” (from social security data). The latter, taking values in [0,1], is defined in terms of amount of work carried out over 1 year. Before applying the error localization and imputation procedures, a detailed analysis based on auxiliary administrative sources has been conducted aimed at properly identifying erroneous cases to be corrected. Finally, impact of imputation at territorial level is evaluated by comparing values of some indicators based on the “raw” data, with the corresponding values based on the final adjusted data.*

**Keywords:** Editing and imputation, administrative data, experimental statistics

**1. Introduction**

Using integrated administrative data gets more and more widespread in Official Statistics because of their richness of information and the reduction of costs. However, the usefulness of administrative data depends on their concepts, definitions and coverage, the quality with which the data are reported and processed and the timeliness of their availability (Statistics Canada, 2010). It is worth to note that most of these elements are not under the control of the National Statistical Institutes.

The integration of data sources is a specific trait of the use of administrative data since, as they are gathered for other purposes, they do not observe all the variables of interest, and they refer to a population covering a part of the target population.

In recent years, the Italian National Institute of Statistics (Istat) has made available to selected users micro data collections based on the integration of several administrative sources. In particular, micro data archives have been produced in the context of the ARCHIMEDE[[1]](#footnote-1) project (Garofalo 2014), to enlarge the offer of statistical information on socio-economic conditions of households.

The main objective of this work is to test an editing and imputation (E&I) strategy for the micro data collection, with the goal to improve the quality of data.

The most relevant type of errors related to the integration process are linkage errors, coverage errors and consistency errors. In this paper we focus in particular on the last two ones. Although the general setting of E&I is still valid for collections of administrative data (Di Zio and Luzi, 2014), the overall process needs to take into account variable definitions and target population of different input sources. The E&I process can help to recognize systematic errors suggesting definitions not completely harmonized or sources under coverage for particular sub-populations.

The paper is organized as follows. Section 2 describes the microdata collection and outlines the proposed strategy and tools for E&I. Section 3 describes the edit rules and the treatment of categorical and numerical variables. A comparison of indicators calculated on “raw” and adjusted data has been carried out in Section 4 in order to assess the impact of imputation. Finally, conclusions are drawn in Section 5.

**2. Data and methods**

*2.1 Data*

The application described in this paper deals with a collection of microdata produced to study the socio-economic situation of individuals and households in Italy at municipal level, based on integrated administrative sources. The microdata collects information on socio-demographic characteristics (age, sex, citizenship, marital status, highest level of education), occupation (type of employment and work intensity[[2]](#footnote-2)) and income (earned income, income from retirement, total income) of individuals. Table 1 lists the administrative sources used as input for constructing microdata collection. The final dataset consists of 59,720,421 individuals.

**Table 1. List of administrative sources**

|  |  |  |
| --- | --- | --- |
| **Source** | **Holding Authority** | **Information** |
| Municipal Population Registers | Italian Municipalities | Demographics |
| Tax Returns Register | Ministry of Economy and Finance | Income |
| Central Register of Pensioners | National Institute for Social Security | Income |
| Social Security Sources (workers) | National Institute for Social Security | Work |
| Social Security Benefits Register | National Institute for Social Security | Income |
| Student Registers | Ministry of Education, Universities and Research | Education |
| Population Census | Italian National Institute of Statistics | Education |

Each administrative source was checked separately[[3]](#footnote-3) before integration (analysis on metadata, checks on coverage, duplicate records, partial non-response, etc.) but no imputation was carried out. Since administrative data do not observe all variables of interest and they mostly refer only to part of the target population, integration[[4]](#footnote-4) allows an informational enrichment through the creation of new variables and a (partial) correction for under/over coverage of some sub-populations (e.g. income earners) (Bakker, 2011). Furthermore, the integration of administrative sources aims to harmonize data under a single common denominator. Nonetheless, some errors affecting the administrative sources are still present. Besides, the integration process can introduce new bias in statistical data, in particular can increase the possible conflicts into the available information (Di Zio and Luzi, 2014). For this reason, an E&I process was developed.

*2.2. Strategy and tools*

To solve the problem of consistency and comparability of information, taking into account the high number of variables and complexity of information in the integrated dataset, we proceeded in subsequent steps, treating separately categorical and numerical variables and identifying subpopulations for which inconsistencies can be recognized as “justifiable”. Responding to specific knowledge needs, some choices in the construction of specific variables led to misleading inconsistencies. For example, the calculation of earned income (added to other income components) was aimed to produce a proxy of household economic vulnerability. To this aim, for self-employed workers negative components (losses) were excluded; this led part of workers (with positive work intensity) being characterized by an earned income equal to 0.

An automatic editing and imputation procedure was implemented to localize and substitute recognized erroneous values: generalized software SCIA[[5]](#footnote-5) and BANFF[[6]](#footnote-6) were used for handling categorical and numeric variables respectively. The nearest neighbor record (according to a specific distance function) to the failed record is chosen among the potential donors, implementing the Fellegi-Holt approach.

**3 E&I at work**

*3.1 Edit rules*

Starting from a complex dataset, including more than 200 variables, some core issues were identified: demographic information, education and occupation. Considering variables related to these issues, a set of 27 edits (between-variables constraints) was defined (Table 2). The number of failed edits involves about 5.1 million records (8.5%) but the main errors regard missing values on ‘Educational level’ (1.9 million) and inconsistency between ‘Earned income’ and ‘Work intensity’ (3.3 million). In both cases, we are dealing with reconstructed variables derived from sources of different nature. We treated separately the two types of error.

**Table 2. Number of failures by groups of edits**

|  |  |
| --- | --- |
|  | **Number of errors** |
| **Inconsistencies:** |   |
| Age class & educational level | 56 |
| School or university enrolment & educational level | 649 |
| Age class & school or university enrolment | 844 |
| Age class & marital status | 1,338 |
| Age class & occupational status | 4,237 |
| Age class & presence of income from work | 14,430 |
| Earned income & work intensity | 3,296,352 |
| **Missing values:** |   |
| Marital status | 25 |
| Educational level | 1,896,494 |

*3.2. Treatment of categorical variables: focus on ‘Educational level’*

The first macro-phase of the E&I strategy concerns the treatment of categorical variables. Although inconsistencies among categorical variables involve a low number of records, several of them (1,896,494) have missing values on the ‘Educational level’. Data are subject to partial non-response as each source covers only some specific part of the overall population. In fact, the highest level of education derives from the integration of student registers - which include the educational qualifications achieved after 2011 - with Population Census, which gathered this information for usually resident population at October 9, 2011. The information on educational level is missing for individuals not resident at October 9, 2011 and who did not receive an educational degree after 2011. Since bias arises when respondents and non-respondents have different characteristics with respect to the investigated phenomena (Little and Rubin, 2002), it is necessary to diagnose the missing data mechanism: for example, most of missing values concerns foreigners (60%). For about 1.9 million records ‘Educational level’ was imputed with SCIA using citizenship and province of residence[[7]](#footnote-7) as stratification variables in donor imputation.

*3.3. Inconsistency between ‘Earned income’ and ‘Work Intensity’*

A specific focus on the inconsistency between the two quantitative variables ‘Earned income’ (from fiscal sources) and ‘Work Intensity’ (from social security archives), was carried out. Before applying the error localization and imputation procedures, a detailed analysis aimed at properly identify erroneous cases to be corrected has been conducted (Figure 1).

At first, an analysis was carried out in order to asses quality of information in the input administrative sources about work intensity and income (e.g. misclassification between employee and retirement income in the fiscal source). Furthermore, admissible cases were identified (e.g. family workers[[8]](#footnote-8) have paid contributions but they often do not perceive an income from work). Finally, supplementary information on wages from Social Security Sources and on companies’ turnover from Istat Business Register was introduced. The presence of these additional variables gives strength to the decision of imputing cases with incomplete information (e.g. under coverage in the administrative sources of some categories of workers with autonomous pension funds[[9]](#footnote-9)).

**Figure 1. Analysis of inconsistencies on numerical variables**



Once the cases to be corrected have been identified, data were processed through the donor imputation with BANFF. Table 3 shows the main features for donor imputation.

**Table 3. Main features for imputation with BANFF**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Errors** | **Auxiliary variables**  | **Matching variables** | **Imputed variable** | **Number of errors** |  |
| Work intensity>0 & Earned income=0 | Wage | Work intensity and wage (layer: province) | Ratio between earned income and wage | 299,020 |  |
| Company’s turnover | Work intensity and company’s turnover (layer: province) | Ratio between earned income and company’s turnover | 109,995 |  |
| Work intensity=0 & Earned income>0 | Wage | Earned income (layer: province) | Work intensity | 316,350 |  |
| Company’s turnover | Earned income (layer: province) | Work intensity | 248,977 |  |

**4. Evaluation of imputation**

Globally, the amount and impact of imputation on the overall distributions is quite small. Some indicators on education and work were calculated on raw data and adjusted data. For benchmark purposes, we also reported Labor Force Survey (LFS) indicators (Table 4). As we can observe, in most cases, the imputation process reduces the differences between indicators from microdata and the benchmark.

**Table 4 – Indicators before and after E&I and Labor Force Survey (LFS) indicators**

|  |  |  |  |
| --- | --- | --- | --- |
| **Indicators** | **Raw data** | **Final data** | **LFS** |
| Percentage of people aged 15-29 years that are not in education or employment | 27.2 | 26.7 | 26.2 |
| Percentage of people aged 30-34 years having completed tertiary education | 27.0 | 26.0 | 23.9 |
| Percentage of people aged 25-64 years having completed at least upper secondary education | 58.3 | 57.8 | 59.3 |
| Percentage of employed people aged 20-64 years  | 56.8 | 56.8 | 59.9 |

Since the micro data collection aims to expand the supply of information of local interest, the impact of error treatment and imputation should be checked at municipal level. The check on ‘Educational level’ reveals an under coverage of information in the Valle d'Aosta region and in Rome (Cartogram 1).

**Cartogram 1. Municipalities by percentage of records with an imputed ‘Educational level’**

****

Cartogram 2 shows that the distribution of imputed ‘Work intensity’ and ‘Earned income’ varies among municipalities. In some cases it appears desirable, e.g. for San Marino Republic's neighboring municipalities[[10]](#footnote-10). We can also observe a higher diffusion of imputation of ‘Work intensity’ in some areas of the South.

**Cartogram 2. Municipalities by percentage of records with an imputed ‘Work intensity’ (a) and with an imputed ‘Earned income’ (b)**

|  |  |
| --- | --- |
| (a) | (b)  |

In order to assess the reliability of imputation at a low territorial level, we calculated four indicators on raw data and on imputed data at municipal level. As we can observe in Table 5, the average absolute differences between them are quite small. An interesting aspect concerns the fact that differences are higher in little municipalities.

**Table 5 -** **Average absolute differences between indicators calculated at municipal level on raw data and indicators calculated on adjusted data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|   | Percentage of people aged 15-29 years that are not in education or employment | Percentage of people aged 30-34 years having completed tertiary education | Percentage of people aged 25-64 years having completed at least upper secondary education | Percentage of people aged 18-59 with very low work intensity |
| Municipalities with less than 2,000 inhabitants | 0.64 | 0.88 | 0.42 | 0.17 |
| Municipalities with 2,000 up to 5,000 inhabitants | 0.56 | 0.58 | 0.30 | 0.11 |
| Municipalities with more than 5,000 inhabitants | 0.55 | 0.70 | 0.42 | 0.08 |
| All municipalities | 0.59 | 0.76 | 0.39 | 0.13 |

**5. Concluding remarks**

This paper deals with the development of a strategy of editing and imputation for microdata collection from administrative sources of the ARCHIMEDE project. The aim is to improve the quality of data, in particular by reducing consistency errors and missing values and to produce an adjusted dataset.

The strategy is based on the use of generalized software (SCIA and BANFF) and relies on the integration of additional data sources aimed at properly identifying erroneous cases to be corrected.

First of all, inconsistencies and missing values on categorical variables involve about 1.9 million records (3.2%). Categorical variables are rarely affected by errors, except for the missing values on ‘Educational level’. In this case a simple donor imputation is implemented. Besides comforting results, a more accurate analysis on missing mechanism and auxiliary sources (e.g. Labor Force Survey) is necessary.

Second, inconsistency between the two quantitative variables (‘Work intensity’ and ‘Earned income’) can be reduced and – partially – resolved introducing information from auxiliary administrative sources.

The impact of the imputation on the overall distributions is quite small at a national level but it considerably varies among municipality.

The future agenda is: (a) to deepen the characteristics of critical sub-populations (e.g. missing data on ‘Educational level’); (b) to test the inclusion of further stratification variables in the imputation phase; (c) to fill still existing inconsistencies in the data (e.g. inconsistencies between income and work with no auxiliary information).

**6. References**

Ambroselli, S. (2015), I codici identificativi univoci all’interno del SIM (Sistema Integrato di Microdati), Istat Working Paper, 5.

Banff Support Team (2003), Functional Description of the Banff System for Edit and Imputation. Statistics Canada, Quality Assurance and Generalized Systems Section technical report.

Bakker, B. (2011), Micro Integration. Statistical Methods Series Document, Statistics Netherlands. Available at: http://www.cbs.nl/NR/rdonlyres/DE0239B4-39C6-4D88-A2BF-21DB3038B97C/0/2011x3708.pdf (Accessed 20 December 2017).

Di Zio M. and Luzi 0. (2014), Editing Administrative Data, in Memobust Handbook on Methodology of Modern Business Statistics. Available at: <https://ec.europa.eu/eurostat/cros/system/files/Statistical%20Data%20Editing-07-T-Administrative%20Data%20v1.0_0.pdf> (Accessed 10 March 2018).

Garofalo, G. (2014), Il Progetto ARCHIMEDE obiettivi e risultati sperimentali. Istat Working Paper, 9.

Little J. and Rubin D. (2002), Statistical Analysis with Missing Data. Wiley, New York.

Reid G., Zabala F. and Holmberg A. (2017), Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ, Journal of Official Statistics, 33 (2), pp. 477–511.

Riccini, M.E. (2004), CONCORD V. 1.0. Controllo e correzione dei dati. Manuale utente e aspetti metodologici, Istat, Tecniche e strumenti.

Statistics Canada (2010), Survey Methods and Practices. Catalogue no. 12-587-X. Available at: <http://www.statcan.gc.ca/pub/12-587-x/12-587-x2003001-eng.pdf>. (Accessed 10 January 2018).

Zhang, L. (2012), Topics of statistical theory for register-based statistics and data integration, Statistica Neerlandica, 66 (1), pp. 41–63.

1. Integrated Archives of Economic and Demographic Microdata. [↑](#footnote-ref-1)
2. ‘Work intensity’ is a measure of the individual participation in the labor market (only regular work derived from administrative sources) during the year. It varies between 0 (no quota of the year worked) and 1 (continuous work in the year). The annual share is calculated as the average of monthly work intensity, derived from the periods for which a social security contribution is observed. [↑](#footnote-ref-2)
3. As suggested by Zhang (2012) and Reid et al. (2017), the first phase of quality assessment includes single sources assessment. [↑](#footnote-ref-3)
4. Record linkage is carried out by a dedicated structure in Istat that identifies a unique code for each record in administrative sources (Ambroselli, 2015). [↑](#footnote-ref-4)
5. A software developed by ISTAT (Riccini, 2004). [↑](#footnote-ref-5)
6. A system of Sas procedures developed by Statistics Canada (Banff Support Team, 2003). [↑](#footnote-ref-6)
7. The distribution of educational level differs on a territorial level. [↑](#footnote-ref-7)
8. People employed in their family business. [↑](#footnote-ref-8)
9. Autonomous pension funds are separate funds (i.e., separate institutional units) established for purposes of providing incomes on retirement for specific groups of employees. [↑](#footnote-ref-9)
10. San Marino Republic is an enclaved microstate surrounded by Italy. Italian resident who work in the San Marino Republic (cross-border workers) are not present in Italian Social Security Sources therefore their work intensity is underestimated. [↑](#footnote-ref-10)