**Local decisions and new guidelines of the Official Statistics**

**An experiment with *Early School Leavers* to measure the quality of integrated administrative data used to support local policies[[1]](#footnote-1)**

Simona Ballabio, Istat - Italian NSI, ballabio@istat.it

Arianna Carra, Istat – Italian NSI, carra@istat.it

Sara Casacci, Istat – Italian NSI, casacci@istat.it

Daniela Ferrazza, Istat – Italian NSI, ferrazza@istat.it

Flavio Verrecchia, Istat – Italian NSI, verrecchia@istat.it

Alberto Vitalini, Istat – Italian NSI, vitalini@istat.it

Lorena Carmen Viviano, Istat – Italian NSI, lovivian@istat.it

**Abstract**

*The work is based on the experimental use of data from the ArchIMEDe project (ARCHivio Integrato di Microdati Economici e Demografici – i.e. Integrated Archive of Economic and Demographic Microdata) of the Italian Institute of Statistics (ISTAT), created with the aim of expanding the supply of information of regional and local interest through the production of micro-data from administrative sources. In a scenario of possible integration and/or transition from survey sources to administrative sources of Italian official statistics, the first step in the study of phenomena is the problematization of operational measures and existing definitions. Although the data used - coming from secondary sources - are not structurally similar to those of ISTAT sample surveys and are not even directly comparable, the new boundaries of official statistics show an increasing interrelationship which, at local level, requires an integrated reading of phenomena through the different sources available in order to obtain the best information and the best quality. The contribution, conceived as part of the SPoT project (data and methodologies for the development of Statistics for Policies on the Territory), proposes a case study of Early School Leavers, defined as young people between 18 and 24 years of age who do not have a secondary school qualification and who are not enrolled in an education cycle.*

*The paper presents the results of some analyses carried out on residents in Lombardy using spatial association measures and hot spot maps. These results highlight, at the same time, elements which are useful from a substantial point of view for the understanding of the phenomenon at a local level, as well as methodological aspects relating to strengths and weaknesses in the use of secondary sources for experimental statistics to support local policies.*

**Keywords:** Experimental statistics, Administrative data, ArchIMEDe, Early School Leavers, sub-regional and local data.

**1. Introduction**

Official statistics are passing through an important change of perspective, where experimentation on secondary data collection becomes an ordinary activity, strictly connected with the evolution of sample survey data. During last years a lot of human and instrumental resources has been invested at the Italian Institute of Statistics (ISTAT) in order to expand statistics from administrative source. The goal of this effort is to expand ISTAT information offer through the production of register based statistics, which led to implementation of elementary collections of longitudinal and cross section data. When facing an important issue (whether social or economic), public decisionmaker and scientists need to know how the phenomenon is distributed locally; how it is distributed within some sub-populations of interest, in order to shape knowledge and policies as closest as possible to population needs. What is the optimal level of territorial detail to support policies? Sample survey data typically refer to national and regional level, but often user need statistics at district and municipality level. Despite typical deficit in administrative data (Eurostat 2003) they proved to have high potential to spread. In fact, they reach higher levels of detail and they allow to move smoothly within different definitions and different perimeters, unhooked from the administrative structures of the territories. In this work we aim to highlights strengths of administrative data through a case study: ‘ELET people’, Early leavers from education and training (previously known as ESL - Early school leavers). According to Eurostat definition (Eurostat 2003) in this category we include young people between 18 and 24 years of age who do not have a secondary school qualification and who are not enrolled in an education cycle. This topic being so crucial from an European perspective, that the Strategic framework for European cooperation in education and training (ET2020) defines a goal to be achieved by 2020: ELET in the EU-28 should be not more than 10%. This contribution is conceived as part of the SPoT project[[2]](#footnote-2) which refers to ArchIMEDe data[[3]](#footnote-3) in order to obtain both substantial information about ELET and peculiar features regarding quality of statistical information. We used 2012 data, also because of its proximity to the census year that allows us to have robust reference data.

**2. Data and Methods**

*2.1 SIM Highlights*

Centralized management of administrative data is an important activity towards modernizing production processes with the aim of maximizing production efficiency and informational content of statistics.

In this context, the Integrated Micro-data System (SIM – Sistema Integrato di Microdati) is the main repository of administrative data in ISTAT and it is conceived to support transversely the Institute's production processes. SIM hosts *micro data*, since they refer to the basic units of official statistics: Individuals, Economic Units, Places. Relations among units of the same type lead to 'Complex units' (e.g. households, groups of enterprises); relations between units of different types lead to ‘Micro-data systems’, e.g. Leed systems - Linked employer employee data (Individuals-Economic Units), Students-School/University systems (Individuals-Economic Units), Mobility systems (Individuals-Places).SIM is an *integrated system* since (a) it identifies each object in the administrative sources with a unique and stable (over time) ID number[[4]](#footnote-4); (b) it defines, for each object, the logical and physical relationships among different administrative data sources (Ambroselli and Garofalo 2015).

*2.2 ARCHiMeDe*

Given this amount of information, Istat developed ArchIMEDe project with the purpose of exploring the informative contents of the SIM, in order to produce new statistical microdata collections - made available to selected users - for territorial planning and evaluation of public policies at national, regional and local levels (Garofalo 2014).

The application illustrated in this paper is based on a collection of microdata produced from integrated administrative sources to study the socio-economic conditions of individuals and households. The dataset compiles information on socio-demographic characteristics (e.g. age, sex, citizenship, highest level of education), occupation (e.g. individual participation in the labor market during the year and type of employment), education (school and university enrollment) and income of individuals. The integration of administrative sources (Municipal Population Registers, Tax Returns Register, Central Register of Pensioners, Social Security Archives, Social Security Benefits Register, Student Registers) permits an informational enhancement and a partial correction for under/over coverage of some sub-population. Furthermore, it aims to harmonize data under a single common denominator. However, some 'errors' (under-over coverage, missing values, inconsistencies) and informational gaps are still present and comparability with surveys is not guaranteed. Despite these limitations, the information produced within the project allows to expand significantly the territorial detail to which data are disseminated.

*2.3 Methods*

As mentioned above, we use ETEL population incidence on the total population (18-24). To visualize the distribution of the phenomenon on a territory, Hot spot maps were drawn up using the value of the indicator in the various areas, with the aim of making immediately visible the contrast between areas where the phenomenon is more intense (hot, red and orange areas) and areas where it is less intense (cold, blue areas).Furthermore, a specific spatial correlation indicator of LISA family -Anselin Local Moran's Indicator- has been used in order to perform a cluster and outlier analysis of municipalities. which allows to identify concentrations of high values, concentrations of low values, and spatial outliers. It can help answer different kinds of questions:Are there municipalities in a study area with anomalous patterns? Where are the unexpectedly high value of ETEL across the study area? Are there clusters of municipalities with similar characteristics with respect the ETEL indicator?

Are there municipalities in a study area with anomalous patterns? Where are the unexpectedly high value of ETEL across the study area? Are there clusters of municipalities with similar characteristics with respect the ETEL indicator?

**3. Results**

As seen below, archives store enormous quantity of data at high level of detail. Researchers can surf this ocean, picking up variables they need and combining them in a flexible way. This is what we did to build ETEL indicator: we built a derived variable by mixing single variables referred to conceptual dimensions (age, instruction, education).In cartogram 1 incidence of ETEL on 18-24 population is shown as a Hot spot map, which enables us to go deep to a municipal analysis level.

There are some clearly distinguishable ‘hot’ areas in the north-west of Pavia area, in the south of Brescia area and on the border with the Bergamo area. In addition, ’hot’ areas are emerging in the area of Insubria (from now on ‘Insubria’), at the swiss border. information at a subregional level offer an important tool for policymaker in order to scale efforts where more needed.

**Cartogram 1. Incidence of ETEL population, by Municipalities, Lombardia, 2012**



**Source**: Experimental ArchIMEDe data.

**Note**: Orange and red indicate "hot" areas, which may be critical. Red indicates municipalities whose index values are greater than or equal the upper quartile plus 1.5 the length of the interquartile range. Orange indicates the remaining municipalities belonging to the last quartile of the index distribution considered.

Evidence is confirmed by LISA indicator which identifies areas with high spatial autocorrelation. Red areas highlight clusters of municipalities that have a high indicator value and are close to municipalities with a high indicator value (high-high); while blue areas highlight clusters of municipalities with low indicator values that are close to municipalities with a low indicator value (low-low).

’Hot’ areas in Cartogram 1 are further specified: in addition to the Insubria, the area of Pavia, Lodigiano, and above all, Brescia, Bergamo and Mantova.

Could this information be useful for local policy making? Is the observed result due to a real phenomenon? In order to answer these question, it is useful to reflect about quality in *these* administrative data. ArchIMEDe does not yet contain data on regional training. A future integration of this specific repository is a must for further project developments. In addition, there is another emerging consideration about the quality of archive data which is, however, linked to a 'latent' competence and could unveil a systematic error source There is an high risk of misleading reading without a direct knowledge of the area, which make some evidences sound like counter-intuitive. Insubria has historically an high level of social and human capital area, in contradiction with the supposed higher intensity of ETEL. Researchers hypothesize that an overestimation could be caused by the phenomenon of “study and work across borders”. In particular, the occurrence of “frontier workers” is difficult to capture, both from the official sample survey data (ISTAT – Labour force survey), due to problems of precision of estimates, and from ArchIMEDe, due to the current absence of adequate secondary sources.

However, reliable statistics are available for further integration: ISTAT (Population Census); FSO (Swiss Federal Statistical Office) and FSOAT (Statistics Office of the Canton of Ticino). Considering the ten-yearly frequency of the population census, probably ArchIMEDe should be better implemented with data from foreign sources (e.g. SIMIC – Central Migration Information System of Switzerland – FSO) in order to improve the quality of information on an annual basis. By the way, transition to a permanent census [[5]](#footnote-5)is likely to open new reflections on how to integrate administrative and sample survey data. Administrative data allow researchers to go beyond standard definition of concepts and to discover different perspective on topics. In this case, further than citizenship and gender we considered Income, – which is not usually grasped in ETEL official statistics– as a useful variable trough which look at the target population. It can be considered as a proxy of the engagement level in labour market, or/and as an indicator of self –support capacity.

***Cartogram* *2*. *LISA – Local Indicator of Spatial Association, ETEL Incidence, Lombardia, 2012***

|  |  |
| --- | --- |
|  |  |

**Source**: Experimental ArchIMEDe data.

Rankings of ETEL Incidence have been constructed considering 12 *sub-regional district*[[6]](#footnote-6) of Lombardia Region (Table1). Different rankings emerge according to different population profiles: Total population, Only Italian population, non-income population, Female population, both italian end foreing. These rankings, beyond confirming critical areas highlighted above, provides some keys to differentiate the territory and its policies:

* In Mantova e Brescia the criticality seems linked to an important presence of foreign citizens. Infact, if we look at the column of Italian citizens only, the Mantova e Brescia falls significantly in the ranking;
* Varese won the podium considering only non-preceptors of income: this evidence should be linked ‘work and employment across borders’ phenomenon of Insubria;
* Bergamo is the first district considering only Italians, and becomes sixth when focusing only Italian females.This could be linked to gender differences in low-skilled job opportunities. This is seems confirmed also by the the ranking structure of non-income earners.

A mix of local data and investigation of sub-populations allows us to hypothesize "tailored" policies according to specific socio-demographic characteristics. For instance, this could bring to a collaboration with local Entrepreneurs’ association where ETEL are connected to low-qualification works actractiveness (see also ‘Tees Valley Routeways’ program in UK), or forthermore with non-profit associations in areas where the problem is due to the integration of foreigners (eastern Lombardy).

**Table 1. Ranking of ETEL incidence in different sub-population, by District, Lombardia, 2012**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Provincia** | **Total** | **Italian** | **Non income population** | **Italian Women** | **Foreign Women** |
| **Mantova** | **1** | 6 | 2 | 7 | **2** |
| **Brescia** | **2** | 5 | 1 | 4 | 3 |
| Bergamo | 3 | **1** | 5 | 6 | 4 |
| **Pavia** | 4 | **2** | 6 | **1** | 9 |
| Lodi | 5 | 7 | 4 | 5 | 12 |
| Como | 6 | 3 | 7 | **2** | 7 |
| **Varese** | 7 | 4 | **3** | 3 | 8 |
| Milano | 8 | 9 | 9 | 8 | 11 |
| Cremona | 9 | 12 | 8 | 12 | 5 |
| Lecco | 10 | 11 | 11 | 11 | 6 |
| Sondrio | 11 | 8 | 12 | 9 | **1** |
| Monza e della Brianza | 12 | 10 | 10 | 10 | 10 |

**Source**: Experimental ArchIMEDe data.

**5. Conclusions**

Generally speaking, quality is defined as “the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs”(Eurostat, 2003). Referring to official statistics, Eurostat defines 6 criteria to ‘measure’ quality: relevance; accuracy; timeliness and punctuality; accessibility and clarity; comparability and coherence. “Data quality is the perception of data's fitness to serve its purpose in a given context.” (Rothbard A., 2015)

Knowledge of the territory and, above all, clarity (Eurostat, 2003) - that refers to the data’s information environment whether data are accompanied with appropriate metadata - can be useful for the improvement of processes and methods of construction of secondary databases of administrative data aimed at the statistical exploitation. For example, for the information gap in Insubria - due to legislation -the solutions of perspective can be both methodological and institutional. (e.g. Italy-Switzerland projects. See the INTERREG Project "The labour market of the cross-border insubrious area. Integrated Statistical Yearbook", Year 2006).

So, while waiting for specific information gaps to be filled regarding Accuracy of the registries (Eurostat, 2003), it is necessary to exclude from the analyses data sets not denoting the closeness of computations to the exact or true values

Moving from the conceptual framework of the survey to that of register-based statistics, in addition to the well known weakness in integration of secondary data, systematic errors,should also be considered issues related to operational definitions. For istance, being not possible to build register about presence of population, should become necessary to prepare new definitions (i.e. resident and actively population present for work or study reasons).

These weaknesses, which can be overcome as experimentation grows, are counterbalanced by relevance (Eurostat, 2003). Register based statistics can reach an high quality level in meeting current and potential users’ needs. Namely, the chance of carrying out analyses at a very ‘small’ territorial level and diversifying by sub-population are precondition for defining targeted local policies, tailored to users' needs so more efficient and more effective. and are efficient and more effective.

**5. References**

Ambroselli S. and Garofalo G. (2015), Reversing the flow: from an integrated system of administrative microdata to an infrastructure for the users, NTTS 2015. Available at:https://ec.europa.eu/eurostat/cros/system/files/Ambroselli\_etal\_NTTS2015abstract\_SIM\_ARCHIMEDE\_0.pdf (Accessed 18 March 2018).

Anselin, L., Syabri, I. and Kho, Y. (2006), GeoDa: An Introduction to Spatial Data Analysis. Geographical Analysis, 38, pp. 5–22.

Anselin, L. (1995), Local Indicators of Spatial Association—LISA. Geographical Analysis, 27, pp. 93–115.

Eurostat, Working Group ‘Assessment of quality in statistics’ (2003), Definition of quality in statistics, Methodological documents, item 4:2, Sixth meeting 2-3 October 2003,Luxembourg, Available at: <http://ec.europa.eu/eurostat/documents/64157/4373735/02-ESS-quality-definition.pdf> (Accessed 19 may 2018).

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

Lavigne M. and Nadeau C. (2014), A Framework for the Evaluation of Administrative Data, Proceedings of Statistics Canada Symposium 2014. Available at: <https://www.statcan.gc.ca/eng/conferences/.../14284-eng.pdf> (accessed 19 may 2018).

Rothbard A., (2015), Quality issues in the use of administrative data records, University of Pennsylvania. Available at: <https://www.aisp.upenn.edu/wp-content/uploads/2015/06/Data-Quality-Paper_Final.pdf> (Accessed 19 may 2018).

Runci M.C., Di Bella G., Galiè L. (2016), Il sistema di integrazione dei dati amministrativi in Istat, Istat Working Papers n.18. Available at: <https://www.istat.it/it/archivio/193056> ( Accessed 19 may 2018).

1. The paper is the joint responsibility of the authors. Paragraphs 2.1 and 2.2 are attributed to Sara Casacci, paragraph 3 to Daniela Ferrazza. [↑](#footnote-ref-1)
2. SPoT - Data and methodologies for the development of Statistics for Policies on the Territory, Interregional project. [↑](#footnote-ref-2)
3. ArchIMEDe - ARCHivio Integrato di Microdati Economici e Demografici – i.e. Integrated Archive of Economic and Demographic Microdata. [↑](#footnote-ref-3)
4. The process of data integration is carried out by applying specific record linkage strategies depending on the quality of the identifying variables of the basic units (Runci, Di Bella, Galiè, 2016). [↑](#footnote-ref-4)
5. Istat, Piano Generale di Censimento <https://www.istat.it/it/files//2016/03/PGC-CENSIMENTO-POPOLAZIONE-E-ABITAZIONI.pdf> [↑](#footnote-ref-5)
6. In Italian ”Provincia”, local administrative Institutions including different Municipalities. [↑](#footnote-ref-6)