25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11November 2016

Simonetta Cozzi
Marco Di Zio
Danila Filipponi
ISTAT

Session No. 3
Administrative Data/Agencies/Units

An Italian experience to evaluate the quality of integrated administrative data

Abstract

The use of administrative data in the production of official statistics has exceedingly increased in the recent years. If in the past administrative data were mostly used for the purposes of sampling frame construction or as the auxiliary variable in the estimation process, it is now more and more popular to use administrative data also as a direct data source. Effective use of these data has required the development of new methods as well as for the estimation and quality evaluation of statistical products.

This paper reports the Istat experience in the establishing an appropriate methodological framework for assessing the quality of administrative data for statistical purpose. The project, named ARCOLAIO, has developed methodologies for monitoring the quality of administrative data that enter in the different phases of statistical production process. Methodological solutions has been proposed:

- To evaluate the quality of administrative data in the acquisition phase through an approach based on quality indicators
- To analyze the longitudinal stability of administrative data, in terms of metadata and data, and their impact on production process
- To evaluate the quality of statistics based on an integrated use of administrative data

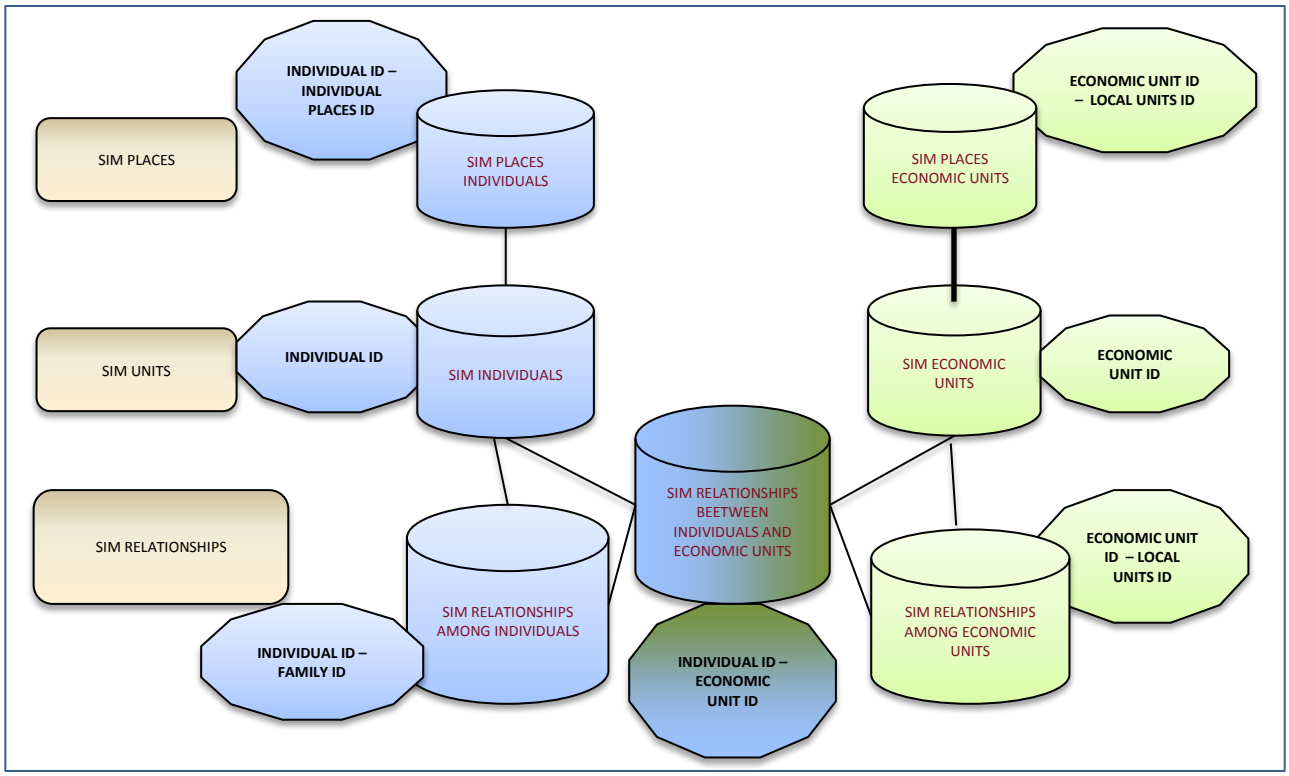
The methodological solutions have been applied to the new Integrated System of Microdata (SIM), recently implemented by ISTAT.

1. The project ARCOLAIO

The ISTAT project, named ARCOLAIO, has studied methodologies for monitoring the quality of administrative data in each phase of a statistical production based on administrative data.
The framework within which the project has been developed is the Integrated System of administrative Microdata (SIM) recently implemented in ISTAT
The Microdata Integrated System (SIM) identifies and integrates the information present in the administrative sources acquired by ISTAT to define an infrastructure usable for the production of social and economic statistics.
The basic objects in SIM are:

- The basic statistical units of interest in official statistics: individuals, economic units and places;
- The variables on these units;
- The relationship between units, between units and variables, and between variables;
- The longitudinal relations

Figure 1 - The Integrated System of Microdata



In this framework, differently from the classical statistical production process, where first the populations and variables are identified and then the data are collected, different populations can be identified by properly combining and treating the basic units, the variables and relationships. Therefore SIM can be thought as a statistical multipurpose population consisting of base elements

(basic units, basic variables, relations) that appropriately combined provide population characterizing the new phenomenon of interest.

The aim of Arcolaio is to define methods for evaluating the quality of SIM, the statistical methodologies to derive statistical populations from an integrated system of administrative microdata and to assess the quality of the obtained outputs. Then, important is the identification of all the potential phases of a statistical production process that uses administrative data according to the SIM approach, the description of data both in terms of units of variables and the identification of the types of errors that can occur in the passage from one state to the next.

The project covers all statistical production phase (from administrative data collection to the statistical outputs) and has been structured in different WP:

- WP1-the Administration Data Documentation System
- WP2- Assessment of the quality of the integrated microdata
- WP3-Assessment of changes in administrative data
- WP4 - Statistical methods in the use of administrative data for statistical purposes and evaluation of the quality.

Knowledge of metadata is an important requirement for a correct interpretation and use of data. A documentation system of structural metadata along with process metadata, is an fundamental tool in a context in which we want to exploit all the available information. The Work Package 1 aims to define an administrative data documentation system, regarding to the structural metadata, process and quality indicators, coherent with the standard documentation systems (GSIM).

The development of methodologies to produce statistical estimate from administrative data is the second important requirement for a "correct" use of secondary data. The use of administrative data for the production of statistics must solve methodological issues related to the secondary nature of the data:

1 the data do not represent a random sample;
2 the phenomenon observed in the administrative source does not always agrees with the statistic concept: (i) because the unit of observation is not the same (ii) the definitions of statistical variables of interest do not coincide with administrative ones (iii) the target population is not the same;
3 the quality control in the production of the data managed by public administration is often missing;
4 Administrative sources are subject to changes in the administrative forms determining longitudinal inconsistency of information.

These issues require the development of tools to evaluate the quality of the input data sources and estimation method for the statistic use of administrative information. The Work Packages 2 and 3 have as objective the development of statistical quality indicators for individual administrative sources and quality indicators of the integration process carried out in SIM. In particular, the Work Package 2 aims to develop quality indicators to evaluate the quality of the available administrative sources exploiting the information gathered by the integration process and the Work Package 3 aims

the development of methods for evaluating the usability of administrative source using the longitudinal information.

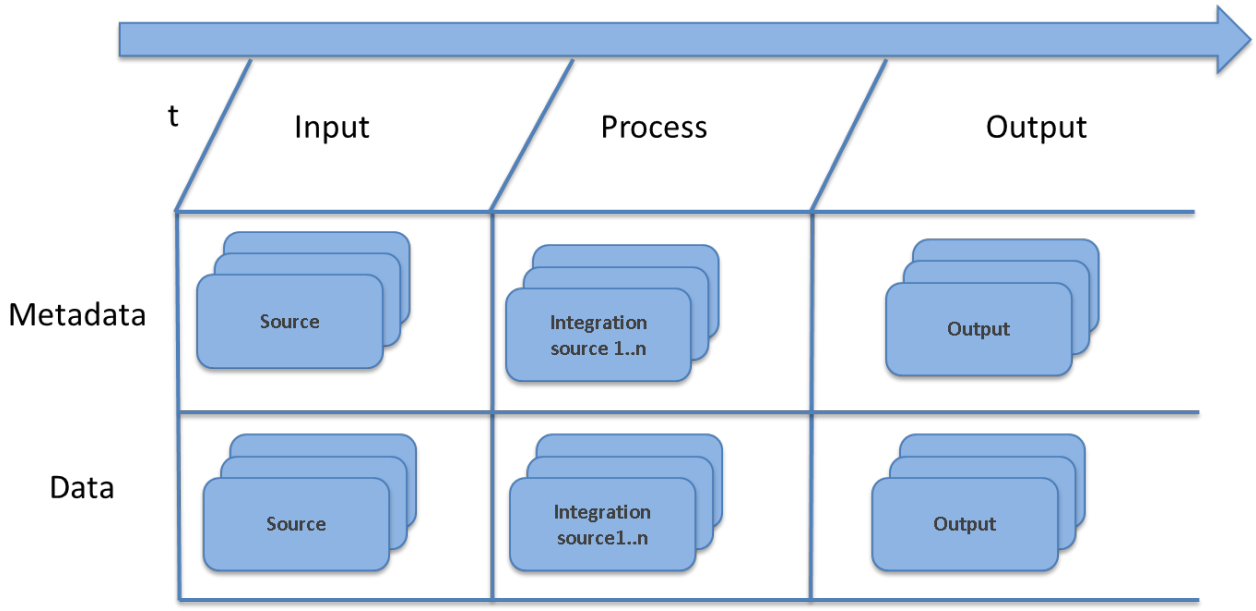
The Work Package 4 aims to (i) specify the inferential context in which estimates based on administrative data are produced; (ii) to define statistical methods for the detection and correction of representation and measurement errors of administrative sources; (iii) to evaluate analyses the quality of the estimates produced by administrative data based on inferential context.

In the paper we will focus on the work packages 3 and 4

2. Focus on WP3 “Assessment of changes in administrative data”

The aim of the WP3 "Assessment of changes in administrative data" it has been to develop methods for evaluating the usability of administrative source using the longitudinal information. Monitoring the changes of AD is important for two main reasons: a) regulatory changes may induce discontinuity producing significant impacts on the statistics production; b) before AD enter into statistical production process an analysis must be carried out to verify the presence of unexpected lack of quality.
Longitudinal aspects involving all stages of the statistical process (input, process, output) that use administrative data at both the metadata and data level (figure 2)

Figure 2 – Temporal dimension



WP3 developed possible methods for validation of an administrative source in the input stage. In particular, outlier analysis has been used for identifying unusual changes over time.
Parametric and non-parametric approach for the identification of outlier has been analyzed and application with simulated and real data has been done.
Specifically, methods have been studied to identify outliers in the contingency tables. In the following, a short theoretical description of the approach used and application of simulated data is described.

2.1 Outliers Identification Procedures for Contingency Tables in Longitudinal Data

Observed cell counts in contingency tables are perceived as outliers if they have low probability under an anticipated loglinear Poisson model[1]. Procedures for the identification of such outliers are derived using the classical maximum likelihood estimator and an non parametric estimator based on median polish.

Outlying observations in a set of data are generally viewed as deviations from a model assumption: the majority of observations -*inliers*- are assumed to come from a selected model (*null model*); few units – *outliers*- are thought of as coming from a different model.

The outliers identification problem is then translated into the problem of identifying those observations that lie in an *outlier region* defined according to the selected null model.

Let consider T categorical variables with possible outcomes

$\{1,.....I_t\},\ t \in \{1,.....T\}$.

Each combination

$i = (i_1,....i_T), \quad with\ \ i \in \mathrm{I} = \otimes_{i=1}^{T} \{1,.....I_t\}$

defines a cell of a contingency table. Under a loglinear Poisson model, the cell counts are considered as a realizations of $(Y_i)\ \ i \in \mathrm{I}$ independent Poisson variables with expected values λ_i

Assuming a Log linear Poisson model, the outlier region for each cell count y_i is defined as

$$out(\alpha_i, \lambda) = \left\{ y_i \in N : \frac{\lambda^{y_i}}{n_{yi}!} e^{-\lambda} < k(\alpha_i) \right\}$$ where N is the set of all non-negative integers and

$$k(\alpha_i) = \sup\left\{\ k > 0 : \sum\nolimits_{n_i} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} 1_{[0,k]}\left(\frac{\lambda^{y_i}}{y_i!} e^{-\lambda} \right) \le \alpha_i \ \right\}$$

The cell count y_i is then an outlier, if it lies in the outlier-region of Poisson's distribution with parameters λ_I

In practice to define the outlier-region and identify the outlying cells, it is necessary to estimate the vector of parameters $\lambda_I = (\lambda_i)\ \ i \in \mathrm{I}$

Loglinear models for contingency table are Generalized Linear Models(GLM), the classical estimator for GLM is the maximum likelihood (ML) estimator. Because of the nature of ML estimator, the regression parameters estimates can be highly influenced by the presence of outlying cells. Some robust alternative have been proposed in literature. A procedure that supplies robust estimates in the analysis of contingency tables is the median polish method(Mosteller& Tukey, 1977; Emerson &Hoaglin, 1983).

A simulation study has been done to illustrate the performances of the developed outlier identification methods (figure 3)

---

[1]Davies L, Gather U (1993) The identification of multiple outliers. Journal American Statistic Association 88:782–792

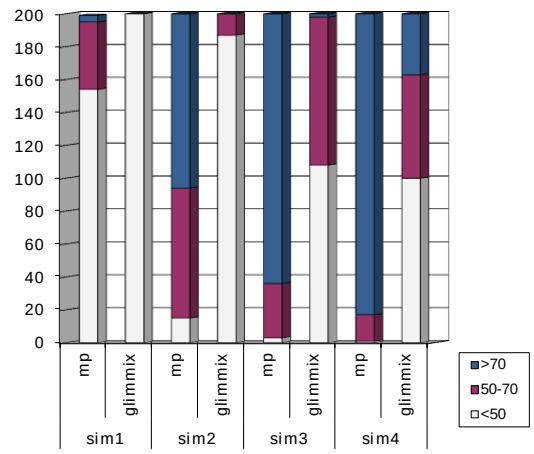
Kuhnt S (2004) Outlier identification procedures for contingency tables using maximum likelihood and L1 estimates. Scand J. Statist. 31:431–442

Kuhnt S (2010) Breakdown concepts for contingency tables. Metrika 71:281–294

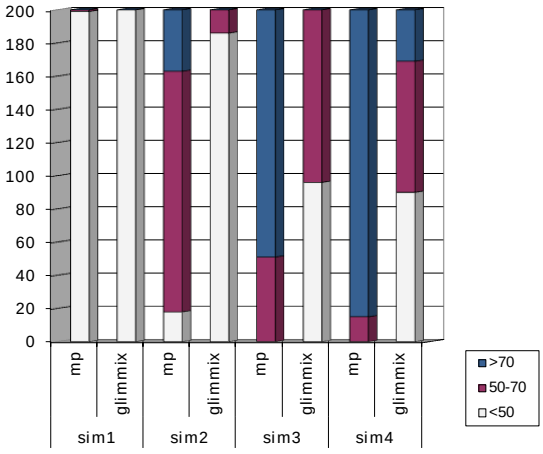
Kuhnt, Rapallo, Rehage(2014) Outlier detection in contingency tables based on minimal patterns, Statistics and Computing May 2014, Volume 24, Issue 3, pp 481-491

Figure 3 - Simulation - Number of tables classified according to the percentage of identified outliers

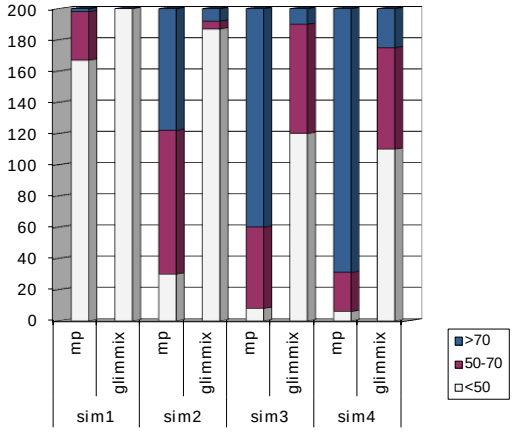
(Table dimension: 20 ×7, simulated outliers 10%)



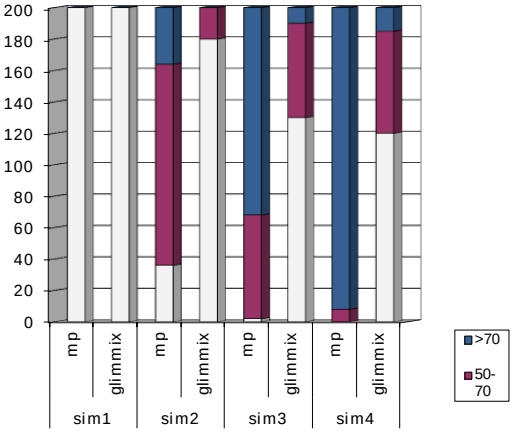
(Table dimension: 50 ×7, simulated outliers 10%)



(Table dimension: 20 ×7, simulated outliers 15%)



(Table dimension: 50 ×7, simulated outliers 15%)



Sim1=distorted data 10%
Sim2=distorted data15%
Sim3=distorted data20%
Sim4=distorted data40%

2.2 Trend analysis

For testing the stability across time of AD, Trend analysis can be used to examine changes in core data elements over time. Trend analysis includes comparisons of counts or proportions over time, as well as more sophisticated time series analysis, smoothing or curve fitting.

The Procedures and a SAS package have been developed by University of Manitoba. In this package, methods for identifying unusual changes over time have been developed using technique that choose the most appropriate model for smoothing the data curve/line. The potential outliers will be flagged on the scatterplot as suspicious points.

Specifically, to perform an outlier analysis, the first step is to choose the best model that fits the data and the second step is to find the outliers based on the selected model. To find the most appropriate model for smoothing data, a set of seven common models which are appropriate for health administrative data were selected: simple linear, quadratic, exponential, logarithmic, SQRT, inverse, negative exponential.

The models were fitted to the aggregated observations over fiscal years or months. For each model, the root mean square error (RMSE) between the model's output and observations was calculated; and the model with minimum RMSE was selected as the optimum model to represent the observations. Secondly, the chosen model was re-fitted to the data to perform an outlier analysis and to calculate studentized residuals without current observation. Significant observations were considered as potential outliers or data quality problems.

An application has been carried out on the administrative data of the Chamber of Commerce used an input file consists the number of enterprise by region and year t (t = 2009, ..., 2013).

Figure 4 - Outlier identification (Italian Provinces)

3. Focus WP4 "Statistical methods in the use of administrative data for statistical purposes and evaluation of the quality".

The main issue of the WP4 is the development of methods for using secondary data into the statistical production. In this context, the main problem is that data are gathered by other organizations for their specific aims, and units and variables refer to specific populations and measurements that are of interest for the body collecting information. Hence, an important task is that of aligning data to fit the NSIs' research interests. The errors made in this transformation step may be broadly refer both to units and measurements . Unit errors and measurement errors as such are particularly important in such a context. A discussion about errors of a statistical production process based on administrative data can be found in Zhang (2012) Different types of errors fit the definition of units and measurement errors , for instance:

- partial information, when some sources refer to a subpopulation of our target population (we will call them "incomplete sources/lists"). This problem is frequently encountered in practice. In fact, the large number of available sources is related to a large number of organizations collecting data, that typically target specific set of units (e.g., specific categories of workers, enterprises having certain legal form,...);
- misclassification, which may be due to differences in the definition of a classification variable or to delays in the registration/cancellation from a list.
- Measurements errors of a variable, which may be due to differences in the definition between the administrative variable and the statistical one.

In this work all these kind of errors are treated into a multi source scenario, that is the estimations are based on multiple data sources referring to different but overlapping populations. This scenario is frequently encountered in practice. In fact, in recent years, the number of available sources for NSIs has been constantly increasing, this fact on one side gives us the opportunity of using statistical methodologies that exploit the information redundancy, on the other side, this abundance is not always associated to a uniform quality of the information. In general, every additional source we are willing to include in the analysis has a lower quality.

Two the possible methodological approach to produce statistical output based on a multi-source information have been exploited within the ARCOLAIO project. In the first we assume that the multiple Administrative and/or Statistical sources provide the value of a same variable of interest for the entire target population or part of it and all the measures are assumed to be imperfect, that is none of them can be considered as a benchmark.

In this case, that could be defined *unsupervised,* a Latent Model approach is particularly suitable. In this framework, it is possible to classify the variables in three groups:

1. variables Y^*representing the "true" target phenomenon. These are the variables that we would observe if data were error free. In general, Y^* are considered latent variables because they are not directly observed.
2. variables Y^g ($g=1,..G$) representing imperfect measures of the target phenomenon. These variables are the ones actually observed from G different data source.

3. covariates X^L and X^M associated respectively to the latent variables Y^* and to the measures Y^g through statistical models.

The statistical model is composed of two components specified via the conditional probability distributions:

$$P(Y^* \mid X^L) \text{ (\textit{latent model})}, \quad (1)$$

$$P(Y^1,..,Y^G \mid Y^*, X^M) \quad \text{(\textit{measurement model})} \quad (2)$$

From the conditional distributions (1) and (2) one can derive the marginal distribution $P(Y^1,..,Y^G \mid X^L, X^M)$ of the imperfect measures.:

$$P(Y^1,..,Y^G \mid X^L, X^M) = \int P(Y^1,..,Y^G \mid Y^*, X^M) P(Y^* \mid X^L) dY^* \quad (3)$$

Then model parameters can be estimated using a likelihood approach, based on the data observed from the G different sources. Once the model parameters have been estimated, we can derive the marginal distributions $P(Y^g \mid Y^*, X^M), \quad g = 1,..,G$ from (2). These distributions can be used to assess the accuracy of each source and the sources can be ranked accordingly. Then the estimates of the true value can be obtained in a hierarchical approach, by selecting the source with the highest level of accuracy.

Using Bayes theorem we can derive the distribution of the latent variables conditional on the available information (*posterior distribution*):

$$P(Y^* \mid Y^1,..,Y^G, X^M, X^L). \quad (4)$$

Then, alternatively to the hierarchical approach, we can use the expectations of the distribution (4) to obtain predictions of the true values for each unit.

In the second approach that could be defined *supervised*, one of the measures (typically a survey measure) is considered as error free and the administrative measures are merely used as auxiliary information. Thus, differently from the unsupervised approach, the error free variable (Y) is modeled as a *response* variable and all the other measures (X) are considered *covariates*. This supervised approach can be adopted in a model based inference as well as in a design based inference. In the latter case, the covariates can be used to specify a *working* model (model assisted approach).The choice of the methodological approach depends both on the informative content and the quality of the available data sources.

Of course, the measure of the output quality depends on the chosen inferential framework. If a latent variable approach is adopted, accuracy measures are naturally provided by the conditional distribution of the latent *true* variable given the available information (e.g., the posterior variance). On the other hand, in a more traditional design based approach, where administrative data play the role of auxiliary variables, evaluation of accuracy is primary based on the estimate of sampling error.