rivista di statistica ufficiale

n. 1-2
2014

Temi trattati

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 1-2
2014

Temi trattati

Istat

Per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione della Rivista di Statistica Ufficiale
All'attenzione di Gilda Sonetti
Istat – Via Cesare Balbo, 16 – 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 1-2/2014

Periodico quadrimestrale
ISSN 1828-1982

Registrato presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Via Cesare Balbo, 16 – Roma

Preface

This special issue of the *Rivista di Statistica Ufficiale* gathers a selection of papers that were originally presented at the "Giornate della Ricerca Metodologica", a research workshop which was held in March, 20-21, 2013 at Istat, the Italian NSI. The workshop gave to experts of statistical methodology and survey statisticians from the NSI as well as from universities an opportunity to discuss recent advances in methodologies for producing official statistics. About 250 people attended the "Giornate della Ricerca Metodologica" throughout the different sessions of the workshop.

The workshop featured 14 papers of Istat experts in different reference areas of methodologies for statistical surveys, from sampling and small area estimation to unit and item non-responses and non-sampling errors treatment, from statistical uses of administrative data to estimation with longitudinal data and surveys, to time series analysis and seasonal adjustment. This issue brings together six articles which passed the reviewing process. The contributions vary in scope and subject.

The paper by *De Vitiis et al*. addresses the issue of estimating demographic and social characteristics of a hard-to-reach population, using non-standard sampling techniques (indirect sampling) and parameter estimation (weight-sharing method). This topic has received increasing attention at NSIs because of emerging needs for surveying subpopulations (homeless, immigrants) which are not easy to detect (because, e.g., of their high mobility) or are only rarely recorded and for which therefore an adequate sampling frame is not available. The authors show how the above sampling strategy has been adopted for the survey of homeless people, which was carried in 2011–2012 for the first time in Italy. Following the indirect sampling approach, in which target population units are reached through random selection of services or facilities that they contact, the estimation is performed through the weight sharing method, based on the links connecting the frame of services with the population of homeless.

Toti et al. present an application of a fairly recent method for outliers identification in the case of multivariate data (Forward Search, FS). It is well known that the presence of missing and/or outlying values in sample surveys (i.e. observations lying "far away" from the main part of a dataset and, possibly, not following the assumed model) may unduly affect inferences from sampled data to the parameters of interest in the population. Outliers can be of fundamental interest in many applications, and thus their identification should also be considered as a goal in itself. Indeed, automatic rules for outlier identification may be troubled by the presence of well-known problems like, e.g., the masking effect (an outlier is undetected because of the presence of another adjacent outlying observation) and the situation may be even worse for multivariate outliers. FS is an iterative procedure which allows to identify groups of outlying observations, even in the case of multivariate data, and to search for structures of heterogeneity in the data, in order to yield robust estimates of the parameters of interest to a predetermined level of confidence. The authors illustrate the results of an application of FS, using a properly developed statistical software solution, for outlier identification in Italian marriage surveys in 2011, comparing micro and macro data.

Righi et al. look into the problem of obtaining valid inference for variance estimation in large scale surveys in presence of imputed data. This is an important issue for survey statisticians because, e.g., of item nonresponses or other errors which typically affect surveys during data collection, introducing an additional component of variability, due to

imputation, during editing phase, of these unobserved or incorrect values. In order to assess some of the variance estimators that explicitly take into account the process of imputation, the authors performed a Monte Carlo experiment on real (business) survey data comparing, under random hot-deck imputation, two well-known methods for variance estimation (bootstrap and multiple imputation, MI) with a relatively new one, which is based on an extension of the standard jackknife technique. The results of the simulation experiment show that the modified jackknife estimator has good performance with respect to the standard methods, yielding nearly unbiased variance estimates while resulting easier to implement and less computer intensive than bootstrap and MI.

The paper by *Rocci and Serbassi* focuses on the important issue, especially for short-term business surveys, of how demographic and other changes to which the target population units are subject may affect estimation of variation of short-term indicators in panel surveys. Indeed, the observed variation may be due both to the intrinsic characteristics of the target parameter that changes over time (and which is of interest for the survey) and the different structure of the population, in terms of number of units or of its composition observed in the two moments. This issue is of particular interest when dealing with business longitudinal data, because of events (like births, deaths, mergers, …) that frequently affect the profile of population units over time. A special case is represented by short-term surveys which produce infra-annual (monthly, quarterly, …) estimates based on a panel sample which is renewed over a much longer time interval (e.g. when changing the base year of the estimated indices). The authors present the results of a simulation study with real data from the Italian monthly survey on employment, working hours, wages and labor costs in large enterprises, in order to assess the representativeness of the panel and to measure the effects of the treatment of legal changes concerning panel units on the longitudinal dynamics of the indices.

The paper by *Comune* illustrates the results of a longitudinal study on income distribution and poverty at local (municipality) level, for the years 2005-2008, using annual tax returns and other administrative data on individuals and households. Estimating poverty at municipal level is a challenging topic in that official poverty estimates are usually available at a much wider territorial detail. A sample of households has been selected from local population register and then, for all individuals in the sampled households, the matching records from tax returns database have been added – after checking for duplicates and ruling out multiple records for the same individual in tax returns database – in order to yield estimation of personal and family income. The author reviews how typical problems with panel surveys, like attrition, were dealt with using following-up rules, aimed at adjusting the sample to make it representative with respect to the major individual and household characteristics in the target population. Other key quality issues, like coverage of the target population and underreporting of income in the administrative data sources have been coped with in the study. Using EU-SILC "at risk of poverty rate" definition and a *local* relative poverty line based on (disposable) income estimated from tax returns records observed in the municipality, the paper shows the resulting estimates of poverty rates from a cross-sectional as well as a longitudinal perspective, analyzing changing status over time through poverty transition matrices for individuals and households. The results show coherence with similar surveys aimed at estimating poverty in small areas using local poverty lines based on income.

The final paper in this issue, by *Cerroni et al.*, deals with the critical problem of defining a conceptual framework for measuring quality of administrative data (AD) and registers, in view of their use for statistical purposes. One of the fundamental problem with using AD is that, normally, statistical concepts do not fully match the administrative ones, which have been defined for other purposes, whereas NSIs and other users of AD for statistical purposes have little or no influence on the definitions and the production processes of the administrative data holders. As a result, little is known *a priori* on the quality of AD. The authors illustrate the results obtained within the European research project BLUE-ETS in developing a conceptual framework of the administrative data quality when AD are considered for possible use as the input source of the statistical process. The framework developed consists of three high level views (referred to as hyperdimension) on the quality of administrative sources, identified, respectively, as Source, Metadata and Data. Each hyperdimension is composed of several dimensions of quality and each dimension contains a number of quality indicators. The first view mainly focuses on the exchange of the data source with the data source holder, while the second view focuses on the metadata of the data in the source. The third view, which is the main subject of the paper, focuses on the quality of the data used as input in the statistical process, yielding a corresponding set of indicators, which are grouped according to five quality dimensions (Technical checks, Accuracy, Completeness, Time-related and Integrability). The authors present the results of a case-study relative to calculating the input quality indicators for the data from the Italian Social Security Database, one of the most relevant AD source used in Istat in current statistical processes, introducing the Quality Report Card, which is a useful tool to display the outcomes of the indicators in a standardized and easy readable format, thus providing potential users with a quick overall evaluation of the data sources quality.

From the above outline it appears that the papers gathered in this issue offer a reasonable balance of contributions of a predominantly methodological character with papers of intrinsically applied type. The statistical processes considered in case studies and applications in the various papers are equally split between business and households/population surveys. Most of the articles cope with different aspects of measuring quality of (survey) data, from estimating variance when imputing for item nonresponses (the paper by *Righi et al.*), to getting rid of outlying observations in sampled data (*Toti et al.*), from measuring the impact of demographic and other events concerning population units on parameter estimates in panel surveys (in *Rocci and Serbassi*) to overall evaluation of the quality of AD sources (*Cerroni et al.*) when these are considered for possible use in a statistical process. Some papers deal with important emerging information needs (surveying homeless, in *De Vitiis et al.*, estimating poverty at local areas, the paper by *Comune*), while two papers explicitly deals with using AD as an auxiliary or alternative source to survey data.

In closing we would like to express our appreciation to all the authors and reviewers for their contribution to this special issue. We also want to thank Stefania Rossetti, who acted as Editor of the *Rivista di Statistica Ufficiale*, for her role in initiating this special issue and facilitating its progress at every step of the way.

Tommaso Di Fonzo and Alessandro Pallara
Associate Editors of the special issue

A methodological approach based on indirect sampling to survey the homeless population

Claudia De Vitiis[1], Stefano Falorsi[2], Francesca Inglese[3], Alessandra Masi[4], Nicoletta Pannuzi[5], Monica Russo[6]

Abstract

The Italian National Institute of Statistics carried out the first survey on homeless population. The survey aims at estimating the unknown size and some demographic and social characteristics of this population. The methodological strategy used to investigate homeless population could not follow the standard approaches of official statistics usually based on the use of population lists. The sample strategy for the homeless survey refers to the theory of indirect sampling, based on the use of a sampling frame indirectly related to the target population. Following the indirect sampling approach, the estimation is performed through the "weight sharing method", based on links connecting the frame of services with homeless population.

Keywords: indirect sampling, weight sharing method, link.

1. Introduction

The survey on homeless people, conducted for the first time in Italy in 2011-2012, is an important component of a research project on the condition of people living in extreme poverty. The research project[7], launched in 2009, was aimed at the dual purpose of building an archive of the system of formal and informal services, public and private, existing in the country aimed to meet the needs of people living in extreme poverty and studying the phenomenon of homeless people spread over the Italian territory.

To achieve these objectives three separate and successive surveys were carried out with the following purposes: (i) the construction of an archive relative to the population of centers (organizations/entities) providing services to people living in extreme poverty (2009-2010), (ii) the acquisition of detailed information on the services provided by each one of the listed centers (2010-2011), (iii) the realization of the sample survey on homeless people (2011-2012).

[1] Istat, e-mail: devitiis@istat.it.
[2] Istat, e-mail: stfalors@istat.it.
[3] Istat, e-mail: fringles@istat.it.
[4] Istat, e-mail: masi@istat.it.
[5] Istat, e-mail: morusso@istat.it.
[6] Istat, e-mail: pannuzi@istat.it.
[7] The Ministries of Health and Labour and Social Policies, Istat, the Italian Federation of associations for the Homeless (fio.PSD) and Caritas took part in the project.
The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

The homeless population is composed of all individuals who are experiencing housing problems due to the impossibility and/or inability to obtain independently and maintain a home in the true sense (*ETHOS typology*[8]). In this population people living in public spaces (streets, shacks, abandoned cars, caravans, sheds), in night dormitories, in hostels for homeless people or temporary housing, in housing for specific social support interventions (for homeless individuals, couples and groups) are included.

The availability of information about centers providing services to people living in extreme poverty was the fundamental information for the design of the survey on homeless people. In fact, because a direct list of individuals belonging to the target population was not available, the sampling strategy could not be based on a classic direct sampling approach.

In fact, homeless people cannot be captured by means of standard household surveys, as they are not listed in population registers. Investigating this particular population requires, therefore, reaching statistical units in a different way, for example by individuating centers where they regularly go to receive the services they need. This situation can be dealt with in the general context of *hard-to-reach* population methods (Marpsat and Razafindratsima, 2010), in particular by means of sample designs classified *time-location sampling*, in which target population units are reached through the random selection of places (the centers) and instants of time.

The centers chosen for the survey are those providing night shelter and canteen services: the former are by definition services frequented exclusively by homeless people, while the latter are services frequented not only by homeless people but also by people who are living in a state of extreme poverty. However, even for the considered centers the list of users was generally not available (for some of them, especially night shelter, a nominal list existed, but these list are not stable enough), not allowing the direct selection of individuals.

For all the above reasons, the methodology adopted for the survey finds its theoretical basis in the *indirect sampling* which is founded on the idea of using a sampling frame referred to a population that is linked indirectly to the target population. In this approach, therefore, the sampling design is defined on the basis of information contained in a list that refers to a different population from that of interest, assuming an exact correspondence between the units contained in the two populations (provisions and persons) with reference to a specific time interval. An alternative approach proposed for a similar context is the Centre Sampling (Baio *et al.* 2011), which allows to select a sample of individuals from the users of centers and estimate the inclusion probabilities of sample units on the basis of center attendance profiles of interviewed units. Anyway, for the purpose of a nationwide survey in the context of official statistics, we chose to base the sampling strategy on the availability of the list of centers, constructed to this aim in the first two phases of the project, containing crucial information about provisions provided to homeless people, to be used as *reference universe* for the survey.

Thus, the sampling list is represented by provisions, meals and beds, provided by the listed services. In this perspective, the population of interest is restricted to homeless people who frequent services of night shelters and canteen services. This is a similar approach to the one followed for the first French survey on homeless people (Ardilly, Le Blanc – 2001a), which obtained satisfactory results. The method used to estimate the parameters of the population of

[8] Fonte: http://www.feantsa.org/files/indicators_wg/ETHOS2006/EN_EthosLeaflet.pdf

interest is the *weight sharing method* which is based on the knowledge of the links between units belonging to the two populations. The links between the sample units, i.e. provisions, and survey units, homeless people, are one-to-one, in the sense that each service provided can be associated, at a given instant of time, only to a homeless person.

To ensure a correct sharing of the weights, the map of the links between persons and centers has to be known: for each interviewed person the list of all visited centers had to be collected for a fixed period of time. The length of the observation period for each person has been set in a week, so that the estimate of the number of links refers to an average week. The survey instrument to collect appropriately all the information to map the links was a daily diary, in which the places where people ate and slept in the seven days preceding the interview were collected.

The methodology used for this study presents some limitations: the first is related to the coverage of the phenomenon, because part of the population of interest, consisting in homeless people living in public spaces and which do not use services or to night shelters nor canteen services, is not surveyed; the second is due to the fact that the services, particularly the canteen, are also frequented by people who do not belong to the population of interest. Both problems can introduce bias and/or increased variability (in case of over-coverage) in the estimate of the population of interest, the first implying a problem of under-coverage of the phenomenon and the second a problem of over-coverage due to the fact that the population caught at the centers does not reflect the true target population.

The paper is organized as follows. In section 2 the theoretical context of the survey is outlined. In section 3 the preliminary survey steps are presented to describe the construction of the archive of centers. Section 4 is devoted to the description of the sampling design, while section 5 illustrates the estimation procedure. Finally section 6 reports the main results of the survey about homeless people.

2. The methodological approach

2.1 Outline

As is known, a sample survey can be carried out through the selection of a sample from a list of units belonging to the target population of interest. In the classical theory of sampling, since the selection of the sample is random, it is possible to derive the probability of selection of the sample units. This ensures that the accuracy of the results produced by the survey is assessable.

In many situations, unfortunately, a list pertaining to the target population is not available but a list that is closely linked to it exists. The lack of a sampling frame of the target population is a somewhat widespread in the national statistical institutes (such as when a list of clusters of units is available) and this is the larger context within which lies the theory of indirect sampling, in particular applied for the *hard-to-reach populations*.

In indirect sampling the list from which the sample is selected consists, in general, of a different population from the target population (Lavallée, 2007), but linked to the population of interest. Following this approach, the inclusion probabilities of units surveyed cannot be determined directly and, therefore, an indirect and peculiar way is utilized. The estimation method adopted in indirect sampling is, in fact, the weight sharing method that

uses both the inclusion probabilities of units belonging to the list and the links between the units of the sampling frame and the units of the target population. For a correct application of the method it is necessary to know the links that are established between the units of the two populations.

The estimation method is formalized in the following on the basis of the text by Lavallée (2007), but with reference to a target population not composed of clusters, or better, in which the cluster size is equal to one because homeless population consists of individuals.

2.2 Indirect sampling

In a methodological context based on "indirect sampling" the sampling population, U^A, is considered, different from the target population, U^B, for which the estimates of the parameters of interest must be produced.

From the population U^A containing N^A units, a sample s^A of size n^A is selected in order to produce estimates referred to population U^B containing N^B units.

The link between the two populations is identified through an indicator variable which can take two values:

- $l_{ik}^{AB} = 1$ if unit *i* belonging to U^A is linked to unit *k* belonging to U^B;
- $l_{ik}^{AB} = 0$ if unit *i* belonging to U^A is not linked to unit *k* belonging to U^B.

Let π_i^A be the inclusion probability for unit *i* selected in the sample s^A, $\pi_i^A > 0$ for each $i \in U^A$.

For every unit *i* selected in the sample s^A, the unit *k* of U^B which has a correspondence with the unit *i*, $l_{ik}^{AB} = 1$, is identified. In this way the sample s^B, composed of units belonging to U^B, is obtained as it follows

$$s^B = \left\{ k \in U^B \middle| \exists\, i \in s^A,\ l_{ik}^{AB} = 1 \right\}.$$

2.3 Estimation based on weight sharing method

In indirect sampling, the method used for calculation of weights of units is the *weight sharing method* or *generalized weight sharing method* (GWSM, Deville and Lavallée 2006). This method allows the reconstruction of the weight of the units of the sample s^B starting from the sampling weight associated to the units selected in the sample s^A. It takes into account the inclusion probability, π_i^A, of the selected sample units through a calculation based on the relationship between the units listed in the sampling frame and the units belonging to the population of interest.

If the links between the units of the two populations are not one-to-one, the procedure for estimating the parameters of interest is very complex. In this case it is necessary to take into account some circumstances: each unit $k \in U^B$ can be associated with multiple units belonging to U^A; the population U^A can be not perfectly overlapped on U^B, in the sense that some units of U^A can be connected with not eligible units for U^B.

An incorrect reconstruction of the links between units belonging to the population U^B and the units belonging to the population U^A involves the risk of introducing biases in the estimation of the parameters of interest, caused by incorrect counts or by an alteration of the probability selection system. The existence of at least one link between all units of the target population U^B and the units of the population U^A is a necessary condition to ensure the unbiasedness of the estimator based on the weight sharing method.

The weight sharing method provides for each unit *k* of a final weight that is calculated in several steps:

1. the initial weight is calculated as the sum of the direct weights of the units of U^A selected in the sample s^A having non-zero link with the unit *k*, $l_{ik}^{AB} = 1$,

$$ {}_1w_k^B = \sum_{i=1}^{N^A} w_i^A t_i^A l_{ik}^{AB} , $$

being

$w_i^A = 1/\pi_i^A$ and π_i^A the inclusion probability of generic unit *i* selected in s^A defined on the basis of the sample design;

$t_i^A = 1$ if unit $i \in s^A$ and $t_i^A = 0$ if unit $i \notin s^A$;

2. for each unit *k* of U^B the total number of link with units of population U^A is obtained

$$ L_k^B = \sum_{i=1}^{N^A} l_{ik}^{AB} ; \tag{2.1} $$

3. the final weight w_k^B is calculated as

$$ w_k^B = \frac{{}_1w_k^B}{L_k^B} = \frac{1}{L_k^B} \sum_{i=1}^{N^A} w_i^A t_i^A l_{ik}^{AB} . \tag{2.2} $$

From the expression of the weight w_k^B, it is evident that the estimation of the parameters of interest of the population U^B is based on the sample s^A and on the existence of links between populations U^A and U^B. In order to estimate the total of a generic variable of interest y^B on the target population U^B,

$$ Y^B = \sum_{k \in U^B} y_k^B , $$

the following estimator can be used

$$ \hat{Y}^B = \sum_{k \in U^B} w_k^B y_k^B . \tag{2.3} $$

The duality of the estimator $\hat{Y}^B$ with respect to U^A and U^B (demonstrated in Lavallée, 2007) allows to express the estimator of the parameter Y^B as a function of the units belonging to U^A and, at the same time, as a function of the units belonging to U^B. In fact the total Y^B can be written also as

$$Y^B = \sum_{i \in U^A} \sum_{k \in U^B} \frac{l_{ik}^{AB}}{L_k^B} y_k^B . \tag{2.4}$$

This equation allows to define for each unit $i \in U^A$ the variable

$$z_i^A = \sum_{k \in U^B} \frac{l_{ik}^{AB}}{L_k^B} y_k^B , \tag{2.5}$$

from which the equality between the totals of the two variables z_i^A e y_k^B originates

$$Z^A = \sum_{i \in U^A} z_i^A = \sum_{k \in U^B} y_k^B = Y^B .$$

The total $Y^B = Z^A$ can be estimated on variable z_i^A as

$$\hat{Z}^A = \sum_{i \in U^A} w_i^A \, t_i^A \, z_i^A = \hat{Y}^B . \tag{2.6}$$

By substituting (2.5) in (2.6) it is possible to obtain

$$\hat{Z}^A = \sum_{i \in U^A} w_i^A \, t_i^A \sum_{k \in U^B} \frac{l_{ik}^{AB}}{L_k^B} y_k^B = \hat{Y}^B . \tag{2.7}$$

It is clear that the second summation of the formula (2.7) refers to sample s^B and therefore the estimator is

$$\hat{Y}^B = \sum_{s^B} y_k^B \left(\sum_{s^A} w_i^A \, t_i^A \, \frac{l_{ik}^{AB}}{L_k^B} \right) = \sum_{s^B} w_k^B \, y_k^B . \tag{2.8}$$

Properties of GWSM about unbiasedness and variance evaluation using usual Horvitz-Thompson estimators are described and proved by Deville and Lavallée (2006).

3. The first operational phases

The map of - formal or informal, public or private - services to homeless people represented the sample base for the homeless people survey. The list of services represented the sampling frame and also the places where the interviews were conducted (Pannuzi, 2009).

The map has been obtained through two phases: a) a census of the organization and services (centers) offering services to homeless people; b) an in-depth survey on the services providers in order to collect information, both quantitative and qualitative, about their users.

Because the homelessness phenomenon is mainly spread in the wider cities, the census was conducted on a sub-set of Italian municipalities, selected on the base of their demographic size. They are 158 municipalities, including all the municipalities with over 70,000 inhabitants, the provincial capitals with more than 30,000 inhabitants, and the municipalities bordering on the municipality with more than 250,000 inhabitants.

3.1 Census of organizations and services for homeless people

In the 158 municipalities, the census collected information on all the services which provide homeless people supports: for primary needs (one-off financial contributions, drugs distribution, clothing distribution, food distribution, shower facilities and personal hygiene, soup kitchen, street unit), night shelter (self-managed housing, shelter, residential communities, residential communities for night shelter, dormitories, emergency dormitories), day shelter, social secretariat, social support measures.

The purpose was to build a database which, for each detected service, contains all the necessary information: service typology, service details (number of bed spaces, average number of meals provided per day, average number of clients per day), supplier organization denomination, address, phone, possible organizations on behalf of the service is provided, organization representative, organization type. That information is being obtained by a CATI survey, through interviewers selected by fio.PSD and trained by Istat.

Starting from the information contained in the pre-existing Istat, Caritas and fio.PSD databases, the survey updated and completed the picture by adding new organizations, reported by the already interviewed organizations. The added organizations are being interviewed in the same way, with a snowball technique in order to catch the maximum number of centers, even informal, supplying services to the homeless.

3.2 Survey on service providers

Once the database of the centers was complete, the services and, hence, the organizations were surveyed by a CAPI interview. A deep reference frame on the situation of the active services and organizations on the territory for the homeless has been drawn.

The information mainly regarded: the basic organizational and service details (contact details and location); the type of organization (whether municipal or other public bodies, private, NGO, etc.); the geographical area served; the users main characteristics (age, gender, citizenship, household type, presence of any physical or mental restrictions); the service access criteria; the provision of any support to exit from the homelessness; the collection of data by the organization or the service; the funding sources and the share of resources for the homeless; the staff information; the cooperation among the services and the interactions with other organizations, especially with social and health units; the participation of the organization in workshops, seminars about the homeless problems; the client participation in the organizational activities.

The information collected at the services about daily services, number of users and homeless users, opening days and time, represented the key information for the definition of the sampling frame for the survey on homeless people.

4. The sampling design for the survey on homeless people

4.1 Objectives of the survey and reference population

The main objective of the third phase of the project, the survey on homeless people, was to estimate - at national level - the unknown size of the target population, along with some demographic and social characteristics of the homeless people. The survey was planned with the aim of measuring the extent of a phenomenon never observed in our country, if not in local contexts, and to obtain information on the process underlying the causes of social exclusion of homeless people.

The observation field of the survey was delimited within the geographical scope of the 158 municipalities, identified on the basis of population size, and all the centers (more than 800) in which night shelter and canteen services are provided, listed in the archive of centers. The services of night shelters are self-managed housing, shelters, dormitories, emergency dormitories, residential communities and semi-residential community. The canteen services include both those providing a meal for lunch and those providing a meal for dinner.

From the peculiarities of the phenomenon under observation a need arose to use a definition of the population connected to the moment in which the condition of homelessness has occurred and to define an observational period delimited in a precise interval of time. The homeless are, in fact, a population subject to constant changes that may be caused by different phenomena. It is a population that is renewed over time and from one period to another the size and composition of the population may change due to demographic shifts or as an effect of the evolution of society in which, for example, the working condition is strongly precarious. In order to identify correctly the homeless people at the survey stage, it was adopted as reference period of the occurrence of the event that led an individual to sleep in centers that provide night shelters or on the street, the month before with respect to the day of the interview. As the size of the homeless population strongly depends on the period in which the phenomenon is observed, the period of observation has been established in 30 days, a period considered long enough to ensure that the majority of homeless people used at least once the services involved in the investigation. The choice of the reference period of the survey was determined by considerations related to the organization of the network for the data collection and evaluation of experts who assumed that the part of population missed in a winter month was of limited scale. For these reasons, the period of time fixed for the survey of the phenomenon was fixed as the 30 days between November 21st and December 20th 2011.

The delimitation of observation field and reference period for the survey led to a precise definition of the target population. In fact, it consists of all the homeless people who receive at least one provision provided at one of all centers C that provide night shelters or canteen services during the fixed period of time J in the 158 municipalities constituting the territorial scope of reference.

4.2 Sampling design

The sampling list is composed of all the centers - operating in the territorial scope of reference - that provide night shelter and canteen services. The centers are the places where services are provided. The services involved in the survey were classified into three types: nightly accommodation, canteens providing meals for lunch and canteens providing meals

for dinner. In a center there may be multiple services, but in this case each of them were treated as a separate center.

To each center *c* (*c = 1, ..., C*) the opening days in a month were associated, in order to build the center-day couples constituting the statistical units of the list. The services thus defined are associated to each center so as to be identified by a unique center code, ensuring the univocal correspondence between provisions and individual, as each individual can receive only one provision in each center during a single time interval (called "day").

The sampling design adopted for the survey is a two-stage stratified random sampling: the strata are the centers, or, in other words, all centers are included and each of them represents a separate stratum of services. This choice was determined by the limited number of centers, which allowed to visit them all, and from efficiency considerations of the design, since in this way it was avoided a selection stage. The first stage units are the opening days of the centers *C* in the reference period of the survey *J*, each corresponding to a cluster of provisions (meals or beds). The final statistical units (or secondary statistical units) are represented by the provisions (each of them connected through a one-to one relationship to one individual) delivered in the opening-days defined in the list.

In order to ensure the coverage of all the days of the reference month, the allocation of the sample days to centers was made randomly, from the list of opening days of each service. The aim of the sample scheme was to achieve a sample of provisions selected with equal probabilities, since the goal of the survey was to produce estimates at national level. Moreover, information on the variables of interest that could lead the definition of a sample design of a different type (for example with non proportional allocation or selection with unequal probabilities) was not available.

4.3 Sample size and allocation

The sample size was determined in relation to the amount of monthly provisions provided to homeless people in the considered centers. This set of provisions constitute the reference universe for the survey, i.e. the population U^A used for reaching the target population U^B. The number of provisions was estimated using information on the services contained in the archive of the centers. In particular, this information concerned the amount of average daily provisions provided by the services, the opening days of the services in the month, the number of total users of the services in a year, the number of homeless people who were users in one year.

The amount of total monthly provisions was estimated through the information deriving from the CAPI second phase survey, considering the average number of daily provisions for each center and the number of opening days of services in a month, while the estimate of monthly provisions provided to homeless has been obtained by applying to the estimate of monthly provisions, the percentage of homeless users.

The sample size was fixed at about 5,400 units after an assessment based on the calculation of the expected sampling error of the estimate of the population size at national level, which was the planned territorial domain. This evaluation was performed, due to the unavailability of prior information about the variability of the individual attendance of services, on the basis of a likely and simplified conjecture on the variability of the individual number of links, assuming that the variability of the number of links is one of the main sources of variability for the estimates, together with the time variability. The

expected sampling cv for the estimate of the population size was around 3.5%.

In order to achieve a self-weighting sample, the sample size to be assigned to each center has been calculated on the basis of a fixed sample fraction defined as

$$f^{A}(J) = \frac{n^{A}(J)}{\widetilde{N}^{A}(J)},$$

where $n^{A}(J)$ indicates the overall sample size and $\widetilde{N}^{A}(J)$ the total estimated monthly number of provisions provided to the population of interest in the period *J*.

Having thus defined the proportional allocation of the sample between the centers, the total number of provisions attributed to the each service is closely related to the size of the center. The sample size in each center *c* is thus determined as

$$n_c^{A}(J) = \widetilde{N}_c^{A}(J) f^{A}(J).$$

4.4 Assigning the number of sample days to centers

After assigning to each center the total number of provisions to be selected, $n_c^{A}(J)$, to be sampled in the reference month of the survey *J*, the number of provisions for every occasion of investigation *g* and the number of days the sample to be attributed to each center were defined on the basis of fixed rules.

The number of provisions to be selected for every survey occasion has been calculated taking into account the maximum number of interviews that each interviewer would be able to conduct in the selected center-day during opening hours. This number has been fixed at 4 interviews per day per interviewer. For small centers a number of interviews ranging from a minimum of one to a maximum of 4 was attributed; centers of medium size were assigned a number of interviews equal to 4; to larger centers, the number of interviews attributed goes from a minimum of 4 to a maximum of 12 (carried out by 1 to three interviewers).

The number of sample days to centers was assigned by fixing that the maximum number of occasions not exceeded 15 days. In this way two different needs were met: on the one hand to seize variability of the phenomenon through the flow of people in services during the reference period of the survey, and on the other hand to limit the presence of interviewers in services throughout the reference period of the survey.

4.5 Selection of provisions

For each selected primary unit, defined by the couple center-day, users of service were selected at random by the interviewers each time on the spot. The selection was carried out, following a systematic procedure, from the list of the users or from the waiting incoming line, allowing the replacement of the persons selected that do not belong to homeless population.

The interviewer selected at random individuals to be interviewed using a sampling interval defined on the basis of two quantities: the number of people to be interviewed and the number of daily predicted provisions for the specific center, obtained from the second phase survey. Strictly speaking, the sampling interval should be calculated on the actual number of users present on the day of the interview, but the interviewer, especially in the

canteens where the flow is not predictable, did not know this number in advance and had therefore to use "expected" information. However, in order to calculate a posteriori the exact probability of inclusion of selected provisions, at the end of each survey day the interviewer took note of the effective number of users of the service on that occasion.

4.6 Inclusion probabilities

The first stage inclusion probability of the primary units, identified by the couple consisting of the generic *g* sample day in the center *c* (during the period *J*), is defined as

$$\pi_{cg}^{A} = \frac{d_c(J)}{D_c(J)}, \tag{4.1}$$

where $d_c(J)$ is the number of sample days of the center *c* and $D_c(J)$ is the number of opening days of the center *c* in the reference period *J*. The probability of inclusion of the selected provisions in the center *c* conditional on the selection of the day *g* is given by

$$\pi_{i|cg}^{A} = \frac{n_{cg}^{A}}{\tilde{N}_{cg}^{A}}, \tag{4.2}$$

where n_{cg}^{A} is the number of provisions on the day selected in the center *cg* and $\tilde{N}_{cg}^{A}$ is the total number of effective provisions provided to homeless people in the center-day *cg*.

Finally, the probability of inclusion of the provision *i* in the center-day *cg* is defined as

$$\pi_{cgi}^{A} = \pi_{cg}^{A}\,\pi_{i|cg}^{A}. \tag{4.3}$$

5. Estimation procedure

5.1 Outline

In the estimation phase, for the application of the weight sharing method, some preliminary operations were required to calculate the weights of the provisions in the sample s^A and for an appropriate mapping of links between the two populations.

To compute properly the weights of provisions, the inclusion probability of each provision for each sample center-day (*cgi*) were calculated and correction factors have been introduced to take account of total non-response, determined by the non-response of centers and survey days. Some centers, in fact, refused to collaborate to the survey (center non-response) and for some others interviewing homeless people in the selected days was not possible (day non-response). Moreover, as during the field work individuals who refused to cooperate were replaced, this last form of non-response was not treated in the estimation phase.

Besides the classic form of non-recording of information on sampling units, in the homeless survey non-response was determined also by failure to collect essential information for the reconstruction of the links between people and provisions, contained in the diary section of the questionnaire. This type of partial non-response constitutes, in fact, a problem for the identification of the link (link non-response).

The two linked populations consist the first of all provisions provided in the set of sample center-days in the period *J*, $U^A(J)=\bigcup_{cg} U^A_{cg}(J)$, and the second of all the homeless people who visit the set of center-days in the period *J*, $U^B(J)=\bigcup_{cg} U^B_{cg}(J)$.

In practice, for each sampled person the links with the centers belonging to the reference list were collected over a period of time limited to a week; this was considered, in fact, an observation period in which the behavior of a person about the attendance of the centers can be assumed regular compared for example to that observed in a single day. The collection of information over a period of a week allows to limit the link non-response, more than the observation of a longer period of time.

The links were reconstructed on the basis of information collected in the questionnaire used for the interviews, that contained a diary section in which questions were asked about the centers frequented by the individual in the week prior to the day of the interview: for each day of the week, the interviewed person indicated the center, among those considered as the reference universe, in which he/she had lunch, dinner and/or slept. The reference week is, therefore, not a fixed week but for each individual, is a sliding group of seven days. The availability of such information has made it possible to establish links between the two populations, to associate to each individual *k* the weights of the provisions received in the sample center-day and determine the total number of links between the individual and population of provisions.

5.2 The sampling weights of selected provisions

The first operation carried out for the construction of the weights of provisions involved the calculus of the inclusion probability of provisions provided to homeless people selected in the sample s^A. To this end some pieces of information have been used, both obtained during survey, regarding the number of provisions provided at each center-day sample *cg*, and collected in the second phase, regarding the information on the homeless user in centers. The importance of combining these two pieces of information is determined by the fact that the canteen services are also frequented by individuals who are not part of the population of interest. Therefore, the number $\widetilde{N}^A_{cg}$ of actual provisions delivered to homeless people in sample center-days was calculated by multiplying the number of total provisions of the specific day by the proportion of homeless users of the center.

It is worth underlining that the high variability of the influx to the centers does not guarantee that the probability of inclusion of the statistical units in the sample is constant, since the number of interviews per day is fixed a priori. This probability, therefore, may actually vary among selected days both within a center and among centers.

5.3 Total non-response treatment

As regards the lack of participation of centers, which in fact constitutes a non-response of strata, a post-stratification of direct weights has been carried out, in order to ensure that the final weights sum to the known total number of provisions provided to homeless population in a generic month. In fact, in the context of indirect sampling, when no known totals on the target population are available, the size of the reference population consists of

the total number of monthly provisions provided in the centers considered, with reference to the period of observation.

The post-strata have been defined through the analysis of data, separately for the two types of centers. For centers that provide night shelter was considered the geographical distribution of the centers and the size of the centers in terms of provisions per month, for canteen services only the size of the centers were taken into account, thus defining the classes by size in terms of provisions.

For each class the post-correction factor was calculated as the ratio between the total archive number of monthly homeless provisions (the known total) ${}_{ps}\widetilde{N}^{A}(J)$ and the number of monthly homeless provisions of respondent centers ${}_{ps}N_{R}^{A}(J)$,

$$ {}_{ps}f = \frac{{}_{ps}\widetilde{N}^{A}(J)}{{}_{ps}N_{R}^{A}(J)}. $$

The correction for first stage unit non-response of the sampled days was carried out by the calculation of a correction factor for each center, γ_{cg}, considering the effective number of days of interview $d_c^*(J)$,

$$ \gamma_{cg} = \frac{d_c(J)}{d_c^*(J)}. $$

The final weight associated to the provision can be expressed in the form

$$ w_{cgi}^{A} = {}_{ps}f \frac{1}{\pi_{cg}^{A}} \gamma_{cg} \frac{1}{\pi_{i|cg}^{A}} . \qquad (5.1) $$

5.4 Link reconstruction

Proper mapping of the individual links between interviewed sampling units belonging to U^B and the provisions belonging to U^A is an important and delicate phase of the estimation procedure, because through this operation it is possible associate to each unit *k* selected in the sample s^B the weights of the provisions linked to it.

In the survey, 20 percent of the cases has presented a problem of identification of the link (link non-response). Not all individuals respondents were able to provide the retrospective information requested about their attendance of the centers in the week preceding the interview date. The total number of required links may vary from a minimum of 1 (if the unit frequented only the center where the survey found it) to a maximum of 21 link (3 per day for 7 days). In these cases it was necessary to treat the problem to avoid the risk of multiple counting of the same people.

The imputation of missing data in the diary was obtained through a probabilistic intra-record procedure, based on the fact that: i) the behavior of homeless people in the use of services is regular; ii) the geographical characteristics and socio-demographic characteristics of homeless people with partially completed diary, mostly related to the use of services, are not significantly different than the rest of the population.

This allowed to assign to each individual the weights of the provisions associated with the sample center-day in which he has been caught and to estimate the total number of links connecting the individual to the population of provisions.

5.5 Parameters and estimation

The main parameters of interest in the target population - defined in the survey reference period *J* - are total of "fixed" variables, i.e. variables that do not vary over time *J* (age, sex, nationality, etc.).

An important parameter of interest is represented by the unknown size of the population $U^B(J)$.

The total of a generic variable of interest y^B, depending on the period *J*, is

$$Y^B(J) = \sum_{k \in U^B(J)} y_k^B ,$$

while the unknown size $N^B(J)$ of the target population $U^B(J)$ is obtained by setting

$$y_k^B = 1, \ \forall k \in U^B(J).$$

Defining the application *K* that relates all the provisions served in the period *J* in the set of centers *C* to each individual who has received it

- *K*: (provisions) $\rightarrow$ (individual)
- $i = K(i)$

the parameter of interest $Y^B(J)$ can be rewritten in the form

$$Y^B(J) = \sum_{k \in U^B(J)} y_k^B = \sum_{i \in U^A(J)} \frac{y_{K(i)}^B}{L_{K(i)}^B(J)}, \tag{5.2}$$

where $L_{K(i)}^B(J) = \sum_{i \in U^A(J)} l_{ik}^{AB}$ is the number of links of the individual *k* with the provisions received during the period *J* in centers *C* (Ardilly, Le Blanc – 2001a).

From (5.2) it is possible to derive that the generic variable of interest y^B takes the same value for all the provision related to the individual *k*, that is $K(i) = k$. The total $Y^B(J)$, rewritten in the form

$$Y^B(J) = \sum_{k \in U^B(J)} \frac{y_k^B}{L_k^B(J)} \left[\sum_{i \in U^A(J); K(i) = k} 1 \right], \tag{5.3}$$

points clearly out that the quantity in square brackets is the total number of provisions received by unit *k* in the period *J* in the set of centers *C*.

Writing the variable $y^B_{K(i)}$ as y^A_i, the total of variable y^B can be estimated on the population of provisions $U^A(J)$ by the transformed variable z^A_i, defined on the provisions of the population $U^A(J)$, as described in section 2,

$$z^A_i = \sum_{k \in U^B(J)} \frac{l^{AB}_{ik}}{L^B_k(J)} y^B_k \quad . \tag{5.4}$$

Rewriting (2.7) in the specific context here considered, we obtain the expression of the estimator of $Y^B(J)$ in the form

$$\hat{Y}^B(J) = \sum_{s^B} y^B_k \left(\sum_{s^A} w^A_i \, t^A_i \frac{l^{AB}_{ik}}{L^B_k(J)} \right) = \sum_{s^B} w^B_k \, y^B_k \ , \tag{5.5}$$

in which, for simplicity, it is omitted to express the provision index in the full form (*cgi*) relative to the two stages of sampling. For each $k \in s^B$

$$w^B_k = \frac{1}{L^B_k(J)} \sum_{s^A} w^A_i \, t^A_i \, l^{AB}_{ik} \ ,$$

where w^A_i is the weight associated with the unit center-day-provision (*cgi*) selected in the sample s^A with inclusion probability π^A_{cgi}.

Since in practice to each person is given a weekly weight, as it is based on the links referred to a week, the estimation of interest parameters related to the whole reference period, *J*, was only possible through a strong hypothesis about stability in the behavior of individuals among weeks, each of which can, therefore, be considered representative of the whole period. Under this hypothesis it is possible to expand the links observed for a week to the total period *J* by simply multiplying the number of weekly links by the number of survey weeks (Ardilly, Le Blanc – 2001b).

In conclusion, the final weight for individual *k* is expressed as

$$w^*_k = \frac{1}{\mathrm{H} \, L^B_k(\eta)} \sum_{s^A} w^A_i \, t^A_i \, l^{AB}_{ik} \ ,$$

where

- $L^B_{K(i)}(\eta) = \sum_{i \in U^A(\eta)} l^{AB}_{ik}$
- η is the generic week
- H is the number of weeks in month *J*
- $U^A(\eta)$ is the part of population $U^A(J)$ restricted to week η.

5.6 Variance estimation

The variance of estimator $\hat{Y}^B(J) = \hat{Z}^A(J)$ can be calculated on sample s^A using the equivalence between y^B and transformed variable z^A defined in (5.4).

Considering the sampling design adopted for homeless people survey, variance estimation for estimator $\hat{Y}^B(J)$ can be obtained through the expression derived for a two stage sampling scheme with equal probabilities, both at first and second stage. This last condition is valid only approximately, as inclusion probabilities are nearly constant.

In order to describe the context in which this expression has been applied it is necessary to introduce the following symbolic notation (for sake of brevity the reference period of the survey *J* is not indicated)

N_{cg}^A Number of provisions (secondary sample units) belonging to day *g* (primary sample units - opening days) of center *c* (stratum);

n_{cg}^A Number of provisions selected belonging to day *g* of center *c*;

z_{cgi}^A Value of the variable z^A on provision *i* belonging to day *g* of center *c*;

$Z_{cg}^A = \sum_{i=1}^{N_{cg}^A} z_{cgi}^A$ Total of the variable z^A calculated in day *g* belonging to center *c*.

The estimator $\hat{Z}^A$ defined in formula (2.6) expressed considering the two sampling stages as

$$\hat{Z}^A = \sum_{c=1}^{C} \sum_{g=1}^{d_c} \sum_{i=1}^{n_{cg}} \frac{z_{cgi}^A}{\pi_{cgi}^A}, \tag{5.6}$$

can be rewritten - substituting in (5.6) the inclusion probabilities defined in (4.3) - in the form

$$\hat{Z}^A = \sum_{c=1}^{C} \frac{D_c}{d_c} \sum_{g=1}^{d_c} \frac{N_{cg}^A}{n_{cg}^A} \sum_{i=1}^{n_{cg}^A} z_{cgi}^A = \sum_{c=1}^{C} \frac{D_c}{d_c} \sum_{g=1}^{d_c} \hat{Z}_{cg}^A = \sum_{c=1}^{C} \hat{Z}_c^A ,$$

in which:

$$\hat{Z}_{cg}^A = \frac{N_{cg}^A}{n_{cg}^A} \sum_{i=1}^{n_{cg}^A} z_{cgi}^A \quad \text{and} \quad \hat{Z}_c^A = \frac{D_c}{d_c} \sum_{g=1}^{d_c} \hat{Z}_{cg}^A .$$

An estimator of the variance of $\hat{Z}^A$ (Cicchitelli *et al*, 1999) is given by

$$\hat{V}\left(\hat{Z}^A\right) = \sum_{c=1}^{C} \hat{V}_c\left(\hat{Z}_c^A\right) = \sum_{c=1}^{C} D_c^2 \frac{D_c - d_c}{D_c} \frac{1}{d_c} \frac{\sum_{g=1}^{d_c} \left(\hat{Z}_{cg}^A - \sum_{g=1}^{d_c} \hat{Z}_{cg}^A / d_c \right)}{d_c - 1} + \sum_{c=1}^{C} \frac{D_c}{d_c} \sum_{g=1}^{d_c} \left(N_{cg}^A\right)^2 \frac{N_{cg}^A - n_{cg}^A}{N_{cg}^A} \frac{1}{n_{cg}^A} \hat{S}_{cg}^2 , \tag{5.7}$$

where $\hat{S}^2_{cg}$ is the estimate of variance of the second stage of z^A in sample center-day cg

$$\hat{S}^2_{cg} = \frac{1}{n^A_{cg} - 1} \sum_{i=1}^{n^A_{cg}} \left(z^A_{cgi} - \sum_{i=1}^{n^A_{cg}} z^A_{cgi} / n^A_{cg} \right)^2 .$$

From (5.7) it can be observed that the variance of estimator $\hat{Z}^A$ (equivalent to $\hat{Y}^B$) depends on first stage variability (deriving from selection of sample days) and on variability of links collected on individuals associated to selected provisions. Actually, the second source of sampling variability is connected to the attendance of centers of homeless people.

6. Results

6.1 Organizations and services

The results of the data collection carried out during the preliminary phases of the project (the census of organizations and the subsequent total survey on service providers, cfr section 3) show that in 158 Italian municipalities, where the survey was carried out, there were 727 organizations providing, in 2010, services to homeless people (Istat, 2011), as summarized in Table 1 where the total number of services and the total number of users are reported for different typologies of services.

These organizations acted in 1,187 locations for a total of 3,125 services, given that each location provided, on average, 2.6 services. A third of the services gives an answer to primary needs (food, clothes or personal hygiene), the 17% provides a night shelter, whereas the 4% offers day shelter. Social secretariat and social support services are very widespread on the territory (24% and 21%, respectively). The users of support services for primary needs are twenty times higher than those using night shelters and they are twice the users of social secretariat and social support services. Public organizations directly provide 14% of the total services and they reach 18% of the total users. The services provided by private organizations with public financing together with the public organizations, represent two thirds of the total services and they reach two thirds of the total users. Among the service of social secretariat and social support , public services reach around one third of the users; the figures rise to 75% and 90%, respectively, if the private organizations with public financing are added. A maximum 10% of the users are reached by public services for primary needs and for night shelter, the services provided by private sector with public financing represent a further 48% among service for primary needs and 58% among night shelters. Services located in Lombardia and Lazio reach, together, almost 40% of the national users (20% and 17% respectively); the users of services in Milano represent 63% of the users in Lombardia, whereas the municipality in Roma serves 91% of the Lazio users. Both Sicilia and Campania reach around 10% of the total users.

Table 1 - Services and users by service macro-typology – Survey on services providers - Year 2010 (percentage composition and absolute values)

	Absolute values		Percentage compositions	
	Services	Users	Services	Users
Support services for primary needs	1,061	1,305,236	34.0	49.9
Night shelter	520	76,657	16.6	2.9
Day shelter	128	47,202	4.1	1.8
Social secretariat	754	568,161	24.1	21.7
Social support services	662	618,734	21.2	23.7
Total	3,125	2,615,990	100.0	100.0

Source: ISTAT

6.2 Homeless people

The third data collection phase, consisting in interviewing the homeless people, estimated that in November and December 2011 47,648 homeless people used a canteen or night-time accommodation service at least once in the main 158 Italian municipalities (Istat, 2013). The confidence interval within which the number of homeless people may vary, with a probability of 95%, is of between 43,425 and 51,872 people, corresponding to an estimated coefficient of variation around 4.5%, evaluated through the variance estimation procedure described in section 5.6 (Istat 2014).

The estimated number of homeless people corresponds to approximately 0.2% of the population regularly registered in the municipalities covered by the survey. However, it should be noted that this group includes individuals that are not registered by town halls or who are officially resident in municipalities other than those where they actually live. The proportion of homeless people out of the total number of residents was highest in the North-west[9], where homeless people corresponded to approximately 0.35% of the resident population, followed by the North-east with 0.27%, the Central Italy with 0.20%, the Italian islands (0.21%) and the South and the Islands (0.10%).

Most homeless people were male (86.9%) and under the age of 45 (57.9%), two thirds had a maximum level of lower secondary school education and 72.9% stated that they lived alone. The majority were foreign citizens (59.4%) and the most common countries of origin were Romania (11.5% of the total), Morocco (9.1%) and Tunisia (5.7%).

More than half the homeless people who access services (58.5%) live in the North (38.8% in the North-west and 19.7% in the North-east), just over a fifth (22.8%) in the Central Italy and only 18.8% in the South and Islands area (8.7% in the South and 10.1% in the Islands). The result in terms of geographical breakdown, however, stems from the considerably higher concentration of the population in big cities. The higher percentages observed in the North-west and the Central Italy essentially depend on the fact that Milano and Roma account for as much as 71% of the samples surveyed. As many as 44% of homeless people use services based in Roma or Milano: 27.5% in Milano and 16.4% in Roma. Less than 10% of homeless people had problems interacting with the interviewers, and were not able to answer the interview for problems associated with physical or clear disabilities (incapacity, disease or mental disability) and/or dependency issues (7.1%) and for difficulty interacting due to their limited knowledge of the Italian language (2.2%).

[9] These geographical areas are used here and elsewhere in this paper for the purpose of brevity, and would be more fully described as follows: “in the municipalities in the North-west where the survey was conducted”.

Detailed information was collected on those capable of answering the interview, regarding socio-demographic aspects, relations with family, relatives and friends, type of employment, health conditions, use of services and main source of subsistence[10].

On average, homeless people were aged 42.2; only 5.3% were over 64. Foreign citizens are younger than Italians (36.9 against 49.9 years). The age difference also meant that the duration of homelessness tended to be higher among Italians: around half the foreign citizens (49.7%) had been homeless for less than six months, against a third (32%) of Italians; while "only" 9.3% had been homeless for at least four years, against a quarter (24%) of Italians. Overall, the result was an average duration of homelessness of 2.5 years, lower for foreign citizens (1.6 years) and higher for Italians (3.9 years).

The fact that homeless foreign citizens tended to be younger was also linked to higher average levels of education: as many as 43.1% had at least a secondary school diploma (9.3% had a university degree) against 23.1% of Italians; nonetheless, 6.1% of foreign citizens stated that they were illiterate. More than half the Italian citizens (51.5%) had no more than compulsory schooling (lower secondary school diploma).

Among the homeless, 7.5% stated that they had never had a home; of these, similar numbers stated that before becoming homeless, they had lived with friends and/or relatives, in a travelers' camp or similar or who lived in shared lodgings, institutions for minors, disabled or other. These were mainly foreign citizens (72.3%) or younger people (the average age was of 37.4); 28.8% had been homeless for at least two years, 58.5% lived alone and 30.7% with friends or relatives.

Before becoming homeless, 63.9% lived in their own home, a percentage which rose to 73.2% among Italians. Out of the foreign citizens, 20% were already homeless before arriving in Italy. More than a quarter (28.3%) are in employment particularly occasional or temporary work (24.5%) and in low-qualified jobs. 71.7% of homeless people did not work at all, more than half of homeless people (51.5%) stated that they did not work because they couldn't find a job. Only 6.7% had never had a job (a quarter of these were female, two thirds were foreign citizens and under the age of 35). The majority of homeless people (53.4%) receive financial aid from the support network of family, friends or volunteer associations, which in many cases represent their only source of subsistence; 17.9% of homeless people did not have any source of income. The loss of employment was one of the most important factors in the gradual process of social exclusion that leads to "homelessness", along with separation from spouses and/or children and, to a more limited extent, health issues. As many as 61.9% of homeless people had lost a stable employment position, 59.5% had separated from their spouse and/or children and 16.2% stated they were in bad or very bad health. Moreover, very few had not experienced any or only one of these events, confirming the fact that homelessness is caused by a combination of factors.

The lifestyle of homeless people is reflected in the fact that three quarters lived alone (78.3% of Italians and 71.9% of foreign citizens); the foreign citizens were more likely to live with family other than spouses or children, or with friends (20.5% against 12.1%); while only a very small number lived with their partner, spouse and/or children.

[10] The analyses presented below refer only to homeless people capable of answering the interview.

As many as 78.3% of foreign citizens stated that they were in contact with a family member: however, 35.5% had contact only through the internet or by telephone or letter (essentially with parents, spouses and/or children), only 42.8% stated that they managed to see their family and 21.5% did so less than once a year. The number of Italians who had family contact fell to 58.6%; nonetheless, 50.8% stated that they did see them (only 7.8% did so through the internet, by telephone or letter), 8.8% saw them less than once a year and 14% saw them at least once a week and the same number one or more times each month.

In terms of maintaining relations with people who have their own home, "living" in their country of origin only appeared to have limited benefits for Italians: 76.2% stated they had friends and 66% had non-homeless friends; for foreign citizens, the percentages were 71% and 57% respectively.

In the 12 months prior to the interview, in addition to the services provided where the interview took place, 89.4% of homeless people had used at least one canteen service, 71.2% had used a night-time shelter and 63.1% a shower and personal hygiene service (with lower percentages for the use of medical services, day-time shelters and street units). The foreign citizens made more use of canteen (91.3% against 86.5%) and personal hygiene services (67.5% against 56.7%), also due to the greater frequency with which they were forced to sleep on the street, in other public spaces or in make-shift lodgings.

Nearly half (45%) of the homeless people had used employment services (without any substantial differences between Italians and foreign citizens), while Italians tended to make greater use of social services (53.7% against 30.3% of foreign citizens) and health services (64.1% against 48.2%). In the month before the interview, 61.3% of homeless people had used a night-time shelter and 24.4% had also used a day-time shelter: 41% were forced to sleep in an outdoor public space at least once and 26.7% in an indoor public space; around a quarter had slept in a vehicle, shack or abandoned building. Foreign citizens, more than Italians, were more likely to have been forced to sleep in public spaces (73.5% against 59.1%) or make-shift lodgings (48.7% against 39.0%).

Table 2 - Homeless people[11] by citizenship and other characteristics - Year 2011 (percentage composition and absolute values)

	Foreign citizen	Italian	Total
Gender			
Male	87.6	86.2	87.0
Female	12.4	13.9	13.0
Age group			
18-34	46.5	10.4	31.8
35-44	27.7	22.0	25.3
45-54	17.4	30.3	22.7
55-64	7.0	26.5	14.9
65 and over	-	10.9	5.3
Who do you live with?			
Alone	71.9	78.3	74.5
With children and/or spouse/partner	7.6	-	8.4
With other family and/or friends	20.5	12.1	17.1
Where they lived before becoming homeless			
At home	57.5	73.2	63.9
With relatives or friends	18.7	11.5	15.8
Other	23.7	15.3	20.3
Employment			
Employed	27.8	29.2	28.3
Unemployed	72.2	70.8	71.7
- Never been employed	7.7	5.4	6.7
Type of event			
Disease	13.7	19.8	16.2
Separation from spouse and/or children	54.4	67.0	59.5
Loss of stable employment	55.9	70.6	61.9
Total (=100%)	25,658	17,561	43,219

Source: ISTAT

7. Concluding remarks

The adopted approach represents an important innovation for the Italian official statistics because of two main reasons: for the first time the homeless population is surveyed at national level on the whole Italian territory and a new methodological instrument, such as the indirect sampling, is experimented for a large scale survey, obtaining very encouraging results. The analysis of the outcome of the methodological approach has highlighted some features of the implementation which can be taken into account for the improving of future realization of the survey on homeless, which will be carried out in 2014. In particular, one relevant aspect is that for future applications it will be possible to exploit the information about variability of links, associated to interviewed individual, and of the flow of users, in order to plan a more efficient sampling design. Another important aspect to improve is the user selection procedure, especially in the canteens, as the analysis of contact reports showed some difficulties in the selection of homeless people among all users, realized through the replacement of non-homeless persons. These critical aspects are examples of what can be improved on the basis of past experience. Other limitations of the methodological approach, such as coverage issues, can be overcome only by conducting control investigations using different survey techniques.

[11] Net of people with interaction difficulties during the interview.

References

Ardilly, P. and D. Le Blanc. 2001. Sampling and weighting a survey of homeless persons: a French example. *Survey Methodology*. Vol. 27, n.1, 109-118 (2001a).

Ardilly, P. and D. Le Blanc. 2001. "Échantillonnage et pondération d'une enquête auprès de personnes sans domicile: un exemple français" *Techniques d'enquête*. Vol. 27, n. 1, pp. 117-127 (2001b).

Baio, G., Blangiardo, G.C. and M. Blangiardo. 2011. Centre Sampling Technique in Foreign Migration Surveys: a Methodological Note. *Journal of Official Statiscs*. Vol. 27, n. 3, pp. 451-465.

Cicchitelli G., A. Herzel and G. E. Montanari. 1992. *Il campionamento statistico*. Bologna: il Mulino.

De Vitiis C., S. Falorsi, F. Inglese and M. Russo. 2011. "Indirect sampling in the First Italian Survey on Homeless Population". Paper presented at the ITACOSM Conference 2011, Pisa, 27-29 June.

De Vitiis C., S. Falorsi, F. Inglese, and M. Russo. 2012. "Estimating the Homeless Population through Indirect Sampling and Weight Sharing Method". Paper presented at the SIS Conference 2012, Roma, 19-21 June.

Deville J.C. and P. Lavallée. 2006. Indirect Sampling: The Foundation of the Generalized Weight Share method. *Survey Methodology*. Vol. 32, n. 2, 165-176.

Istat. 2011. *Services to homeless people*. 21 november 2011, Available at http://www.istat.it/en/archive/45837.

Istat. 2011. *I servizi alle persone senza dimora*. 3 novembre 2011, Available at http://www.istat.it/it/archivio/44096.

Istat. 2013. *The homeless*. 10 June 2013, Available at http://www.istat.it/en/archive/92503.

Istat. 2014. *The national research on homeless people conditions*. 2014. http://www.istat.it/it/archivio/127256.

Lavallée, P. (1995) 'Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method', Survey Methodology, Vol. 21, No. 1, pp. 25-32.

Lavallée P. 2007. *Indirect Sampling*, Springer, New York.

Marpsat M, Razafindratsima N. 2010 'Survey methods for hard-to-reach populations: introduction to the special issue' Methodological Innovation Online.; 5(2):3–16

Pannuzi N., A. Masi and I. Siciliani. 2009. *Surveys of homeless persons - the case of Italy, Counting the homeless*. Peer Review, 12-13 November.

Xiaojian X. and P. Lavallée. 2009. 'Treatments for link nonresponse in indirect sampling', *Survey Methodology*. Vol. 35, n.2, 153-164.

Metodi di Forward Search per la ricerca di outlier: un'applicazione ai dati Istat sui matrimoni nel 2011[1]

Simona Toti,[2] Romina Filippini,[3] Francesco Amato,[4] Claudia Iaccarino [5]

Sommario

La Forward Search (FS) è un metodo iterativo capace di individuare gruppi di dati anomali, nel caso di dati di regressione o multivariati. Il presente lavoro applica l'approccio FS ai dati Istat relativi ai matrimoni celebrati nel 2011 e alle caratteristiche degli sposi. Lo scopo è l'identificazione di eventuali Province o Comuni con caratteristiche anomale rispetto al resto dell'Italia. I modelli rispetto ai quali si è condotta l'analisi FS sono: il modello multivariato, il modello composizionale e il modello di regressione. Dal confronto dei risultati emerge la forte dipendenza della definizione di dato anomalo dal contesto in studio. L'analisi è stata effettuata con l'ausilio del software SiRiO (Sistema Ricerca Outlier), sviluppato in ISTAT per rendere fruibile il sistema d'analisi anche all'utilizzatore meno esperto.

Parole chiave: outlier, Forward Search, indagine Istat sui matrimoni, SiRiO.

Abstract

Forward Search (FS) is an iterative method to detect the presence of groups of outliers in the case of regression or multivariate data. In the present paper, FS is applied to Istat data on marriages celebrated in 2011 and to the characteristics of partners. The aim is to identify Provinces and Municipalities with significantly different characteristics from the other Italian Provinces and Municipalities. The analysis was conducted referring to three FS models: the multivariate, the compositional and the regression model. The comparison of the results obtained highlights the importance of the reference framework to define an observation as outlier. The analysis was performed using SiRiO (Outlier Research System), a software developed in Istat to make the FS analysis available for non- expert users.

Keywords: outlier, Forward Search, Istat marriage data, SiRiO.

[1] Gli autori ringraziano Alessandra Reale e Sabrina Prati per l'incoraggiamento e il supporto costante. Quanto pubblicato impegna esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

[2] Istat, e-mail: toti@istat.it.

[3] Istat, e-mail: filippini@istat.it.

[4] Istat, e-mail: framato@istat.it.

[5] Istat, e-mail: iaccarin@istat.it.

1. Introduzione

L'identificazione dei dati anomali è un problema importante nell'analisi statistica dei dati: la presenza di osservazioni atipiche può infatti dare origine a risultati distorti (Becker e Gather, 1999). Con dati anomali o outlier si intendono le osservazioni che si trovano "lontane" dalla maggior parte delle altre osservazioni e che seguono un modello diverso da quello assunto per il fenomeno analizzato.

La Forward Search (FS) è un metodo statistico capace di individuare gruppi di dati anomali, anche nel caso di dati multivariati, e di ricercare strutture di eterogeneità nei dati (Atkinson *et al.*, 2004).

Nell'ambito dell'attività istituzionale dell'unità Istat di Controllo e Correzione dei dati censuari (MTC/D) è stata sviluppata una competenza metodologica e il software SiRiO per affrontare il problema dell'identificazione dei dati anomali tramite FS, in tre casi differenti: un modello regressivo (confronto fra archivi); un modello gaussiano multivariato; un modello log-normale composizionale. Mentre le prime due specifiche nascono dalla collaborazione con il gruppo dell'Università di Parma, l'unità MTC/D ha esteso il metodo al caso di dati composizionali ed ha sviluppato l'applicazione web per rendere fruibile il sistema d'analisi in modo semplice ed intuitivo.

Il presente lavoro nasce dalla collaborazione dell'unità MTC/D con l'unità Strutture familiari e ciclo di vita (DEM/C) e riporta i risultati dell'analisi dei dati relativi alle indagini Istat sui matrimoni celebrati nel 2011.

2. L'approccio Forward Search per la ricerca dei dati anomali

La FS è un metodo iterativo il cui principale risultato è quello di individuare, all'interno di un gruppo di osservazioni, il sottoinsieme privo di dati anomali a partire dal quale ottenere stime robuste dei parametri di interesse, ad un prefissato livello di confidenza.

Il primo passo della procedura consiste nell'individuazione di un sottoinsieme di unità con valori vicini al centro della distribuzione dei dati, su cui verranno calcolate opportune statistiche. Quindi, nei passi successivi, il sottoinsieme viene incrementato di una unità alla volta, in modo tale che il sottoinsieme sia sempre composto dalle osservazioni più vicine al centro. Su ogni nuovo sottoinsieme di osservazioni vengono ripetute le stesse analisi statistiche. La procedura garantisce che i dati anomali entreranno a far parte del sottoinsieme soltanto agli ultimi passi.

Ad ogni passo, una buona sintesi dell'andamento della procedura è dato dalla più piccola distanza dal centro della distribuzione delle unità non appartenenti al sottoinsieme di volta in volta analizzato: tale misura è detta "segnale" della procedura stessa. Riportando sull'asse delle ascisse i passi dell'algoritmo (ossia la dimensione del sottoinsieme) e sull'asse delle ordinate il segnale, il grafico della spezzata che unisce i punti mostrerà un picco in corrispondenza del passo immediatamente precedente all'inclusione del primo dato anomalo, se presente.

La conoscenza, anche approssimata, della distribuzione di probabilità della minima distanza utilizzata, sotto l'ipotesi nulla di normalità dei dati, permette la costruzione di bande ad un prefissato livello di confidenza per il segnale.

2.1 Il modello multivariato

Il modello gaussiano multivariato assume che ogni osservazione campionaria sia la realizzazione di una stessa distribuzione normale multivariata con vettore delle medie μ e matrice di varianza e covarianza Σ, ed è rispetto a tale modello che si definiscono i dati eventualmente anomali.

Per valutare la distanza delle osservazioni dal vettore delle medie, centro della distribuzione, ad ogni passo, si ricorre alla distanza di Mahalanobis:

$$d_i = \sqrt{(x_i - \bar{x}_m) S_m^{-1} (x_i - \bar{x}_m)} \tag{2.1}$$

dove $\bar{x}_m$ è il vettore delle medie campionarie e S_m è la matrice di varianza e covarianza campionaria. Tali parametri vengono stimati con le sole osservazioni presenti nel sottoinsieme, a quella iterazione.

È da notare che, mentre la stima della media non è influenzata dal troncamento del campione, la matrice di varianza e covarianza campionaria sottostima la matrice di varianza e covarianza della popolazione. Per correggere tale distorsione la FS prevede l'uso del fattore di espansione di Tallis (1963).

Sotto l'ipotesi nulla di normalità delle variabili, è possibile costruire delle bande di confidenza per la minima distanza di Mahalanobis: fin quando la distanza delle osservazioni è stimata sulla base di un campione coerente con l'ipotesi nulla, il segnale resta entro la banda; se, negli ultimi passi della procedura, il segnale esce dalle bande, questo indica l'entrata nel sottoinsieme di stima di un'unità non omogenea, dunque di un'unità anomala.

L'utilizzo nella FS degli stimatori di massima verosimiglianza (media e matrice di varianza e covarianza campionarie), sensibili alla presenza di osservazioni anomale, rende subito evidente l'introduzione del primo outlier segnalato dalla presenza di picchi nella traiettoria del segnale.

La distribuzione della minima distanza di Mahalanobis non è nota in letteratura. Tuttavia, è possibile ottenere bande di confidenza molto accurate per tale statistica facendo riferimento alla teoria delle statistiche d'ordine sotto l'ipotesi di normalità delle osservazioni. Le bande di confidenza così ottenute sono basate sulla distribuzione F di Fisher (Riani *et al.*, 2009).

Poiché non è sempre possibile fare l'assunzione di normalità dei dati, la procedura della FS introduce un passaggio propedeutico all'analisi vera e propria, nel quale si utilizza la trasformata di Box-Cox in associazione alla FS, per riportare i dati sotto l'ipotesi di normalità.

2.2 Il modello composizionale

L'analisi di dati composizionali trova applicazione ogni volta che l'oggetto di interesse è un vettore v-dimensionale le cui componenti siano variabili reali positive, che possono essere viste come porzioni di un totale fissato. Se dunque l'interesse è nell'ampiezza di una componente relativamente alle altre, allora confrontare differenti composizioni si traduce nel confrontare ogni valore con tutti gli altri tramite rapporti (Aitchison, 1986). Esistono diverse funzioni che applicate ai dati composizionali v-dimensionali vincolati a una somma fissata, restituiscono vettori indipendenti in R^{v-1}. Tra queste la trasformata ILR (Isometric Log Ratio) garantisce una perfetta isometria tra lo spazio composizionale e quello reale

(Egozcue *et al.*, 2003). Inoltre se si assume per i vettori di dati osservati una distribuzione log-normale multivariata, la trasformata ILR restituisce vettori a v-1 componenti indipendenti nello spazio reale con distribuzione multinormale (Aitchison e Shen, 1980).

Per i dati trasformati, il modello di riferimento sarà dunque una distribuzione normale multivariata. In questo modo il problema composizionale diventa un problema multinormale ed è quindi possibile adottare la stessa procedura d'analisi descritta nel paragrafo precedente. La procedura FS per dati composizionali prevede dunque:

1. un passo iniziale di applicazione della trasformata ILR ai dati;
2. l'applicazione della procedura FS multivariata gaussiana.

2.3 Il modello regressivo

Un problema classico dell'analisi di dati amministrativi è quello del confronto tra fonti: avendo a disposizione per una data variabile continua, rilevata su una certa popolazione, una fonte attendibile, cioè con valori non affetti da errore, la si vuole utilizzare per valutare l'attendibilità di un'altra fonte, potenzialmente affetta da errori. Una possibile rappresentazione della relazione tra le due fonti, è data dal modello regressivo a intercetta nulla e coefficiente di regressione unitario. I punti "distanti" da tale retta esprimono una discrepanza tra le due fonti, indicativi di una potenziale situazione di errore nella fonte scelta come variabile dipendente.

Per valutare la distanza delle osservazioni dal modello di regressione lineare semplice si utilizzano i residui studentizzati[6]:

$$r_i = \frac{e_i}{\sqrt{S_m{}^2(1-h_i)}} \tag{2.2}$$

dove e_i è il residuo del modello di regressione, S_m la matrice di varianza e covarianza campionaria, h_i i valori di leva ossia gli elementi della diagonale principale della matrice cappello:

$$H = X(X_m{}^T X_m)^{-1} X^T \ . \tag{2.3}$$

L'utilizzo nella FS della tecnica, poco robusta, dei minimi quadrati (stimatori OLS, Ordinary Least Squares) per la stima dei parametri del modello, diventa un punto di forza nella metodologia. Ad ogni passo, il più piccolo residuo delle unità non appartenenti al sottoinsieme di stima del modello è il segnale della procedura. L'introduzione di dati anomali determina forti picchi nella spezzata che monitora il segnale.

È possibile costruire delle bande di confidenza del minimo residuo per identificare le osservazioni omogenee rispetto al modello e quelle che, avendo associata una distanza talmente elevata da risultare esterna alla banda, risultano significativamente lontane dal modello, dunque relative a osservazioni anomale. La distribuzione del minimo residuo non è nota in letteratura ma una sua approssimazione è stata derivata sulla base della distribuzione delle statistiche d'ordine sotto l'ipotesi di normalità dei dati (Atkinson e Riani, 2006).

[6] Si parla in effetti di residui studentizzati di cancellazione per evidenziare il fatto che le quantità a numeratore ed a denominatore sono calcolate su sottoinsiemi diversi del campione (Atkinson *et al.*, 2004).

3. SiRiO: Sistema Ricerca Outlier

Il pacchetto Matlab FSDA, composto da funzioni per l'analisi FS di regressione e multinormale, è disponibile all'indirizzo http://www.riani.it/MATLAB.htm. Nell'ambito della collaborazione tra Istat (Unità di Controllo e Correzione dei Censimenti) e Università di Parma sono state sviluppate procedure *ad hoc* per il caso del confronto fra fonti e per il modello multinormale. Per il modello composizionale è stata sviluppata all'interno dell'Istat una procedura Python che applica la trasformata ILR ai dati (Palombi *el al.*, 2011) il cui output viene analizzato dalla procedura multinormale.

L'applicazione web SiRiO, sviluppata in Java, si interfaccia con le tre procedure in modo da renderle fruibili senza necessità di specifiche competenze Matlab e Python. Ogni utente può creare diversi progetti; ogni progetto è caratterizzato da uno o più insiemi di dati di partenza e da uno specifico tipo di analisi. Per ogni insieme di dati analizzati sono resi disponibili grafici e tabelle, scaricabili come file in formato zip. Attualmente l'applicazione è disponibile sulla rete interna Istat[7].

4. L'applicazione della FS ai dati Istat sui Matrimoni 2011

La presente analisi utilizza l'applicativo SiRiO per la ricerca degli outlier tramite il metodo FS, con l'obiettivo di individuare la presenza di eventuali comportamenti anomali, a livello provinciale e comunale, relativi all'evento matrimonio e alle caratteristiche degli sposi.

In particolare, è stata analizzata a livello provinciale la condizione professionale degli sposi e delle spose separatamente (fonte: Rilevazione dei Matrimoni). Lo scopo di questa prima analisi, è quello di mettere a confronto la distribuzione del numero di matrimoni per condizione professionale (Occupato, In cerca di occupazione, Inattivo), separatamente per sposo e sposa, individuando le Province che hanno un profilo che si discosta da quello dalle altre. Si è dunque utilizzato il modello multivariato classico e quello composizionale della FS per la ricerca dei dati anomali.

La seconda analisi ha riguardato il confronto tra i dati ricavati dai modelli individuali (fonte: Rilevazione dei Matrimoni) e quelli ricavati dai modelli riepilogativi (fonte: Rilevazione mensile degli eventi demografici di Stato Civile). Lo scopo è di individuare i Comuni che hanno scostamenti tra i valori riportati nelle due fonti, significativamente diversi da quelli degli altri Comuni della Provincia. In questo caso, si è utilizzato il modello di regressione della FS, assumendo come variabile dipendente il numero di modelli individuali e privi di errori i valori riportati nei modelli riepilogativi.

L'analisi dei risultati, oltre a individuare le Province con informazione statistica e distribuzione della nuzialità, anomale, permette di apprezzare le differenze dei vari modelli FS.

[7] Gli utenti Istat interessati all'utilizzo di SiRiO possono richiedere le credenziali di accesso inviando una mail a framato@istat.it.

5. Le fonti

Le informazioni sui matrimoni celebrati in Italia vengono rilevate da due indagini Istat: la Rilevazione dei Matrimoni (Modello Istat D.3) e la Rilevazione mensile degli eventi demografici di Stato Civile (Modello Istat D7.a).

La *Rilevazione dei Matrimoni* (D.3) è di fonte Stato Civile e fa quindi riferimento alla popolazione presente. È un'indagine individuale ed esaustiva che ha per oggetto tutti i matrimoni celebrati in Italia e che consente di analizzare il fenomeno della nuzialità con riguardo alle principali caratteristiche del matrimonio e degli sposi.

Le informazioni raccolte riguardano sia le notizie sul matrimonio, quali la data, il rito di celebrazione (religioso o civile), il comune di celebrazione e il regime patrimoniale scelto dagli sposi (comunione o separazione dei beni), sia le informazioni relative a ciascuno degli sposi, quali la data e il comune di nascita, il comune di residenza al momento del matrimonio, il luogo di residenza futura degli sposi, lo stato civile precedente, il livello di istruzione, la condizione professionale, la posizione nella professione, il ramo di attività economica e la cittadinanza.

La *Rilevazione mensile degli eventi demografici di Stato Civile* (D7.a) raccoglie, a livello comunale, le informazioni relative alle nascite, ai matrimoni e ai decessi dichiarati presso gli uffici di Stato Civile. In questo caso, quindi, i dati raccolti non sono individuali ma riepilogativi. Nello specifico, nel caso dell'evento "matrimonio", la rilevazione fornisce per ciascun mese il numero di matrimoni celebrati in ogni singolo Comune, secondo il rito religioso o civile.

6. Un'analisi multivariata: la condizione professionale degli sposi 2011

Dalla *Rilevazione dei Matrimoni* (D.3) è possibile ricavare alcune informazioni demo-sociali per ciascuno degli sposi. I dati rilevati a livello comunale sono stati analizzati per Provincia. In particolare, è stata considerata la variabile Condizione professionale, separatamente per sposo e sposa, allo scopo di individuare le Province con caratteristiche anomale.

6.1 Descrizione dei dati

La variabile Condizione professionale, inizialmente suddivisa in nove categorie, è stata ricodificata in tre modalità: Occupato, In cerca di occupazione (che aggrega: disoccupato; in cerca di prima occupazione) e Inattivo (che aggrega: ritirato dal lavoro; casalinga, solo per le spose; studente; inabile al lavoro; in servizio di leva o servizio civile, solo per gli sposi; altro). Il numero totale di matrimoni analizzati è pari a 204.830 nelle 110 Province italiane.

I dati relativi agli sposi e alle spose sono stati analizzati separatamente, sia con modello multivariato gaussiano che composizionale. Nel seguito sono illustrati solo i risultati dell'analisi sulle spose. Per quanto riguarda gli sposi, l'analisi non ha evidenziato nessuna Provincia anomala.

6.2 Analisi dei risultati: modello gaussiano

Le tre variabili considerate a livello provinciale sono: il numero di spose occupate, il numero di spose in cerca di occupazione, il numero di spose inattive. La matrice di dati, analizzata dalla procedura, contiene dunque in riga le Province e in colonna i valori di ciascuna delle tre variabili.

Nella procedura implementata in SiRiO il primo passo prevede l'applicazione della procedura Box-Cox per la normalizzazione dei dati, nel caso che i dati non rispettino l'ipotesi nulla di multinormalità. Poiché la trasformata di Box-Cox risente dell'eventuale presenza di dati anomali, anche per la ricerca della migliore stima del parametro di tale funzione, si procede utilizzando l'approccio FS.

In figura 1 sono riportati i grafici forniti da SiRiO, che rappresentano i valori relativi alle 110 Province prima e dopo la trasformazione. In particolare: sulla diagonale principale sono riportati gli istogrammi delle tre componenti della variabile Condizione professionale; fuori dalla diagonale i grafici di dispersione a coppie di componenti. I valori relativi alle osservazioni anomale sono indicati sul grafico da un cerchio.

Mentre i dati non trasformati presentano una distribuzione asimmetrica a valori positivi (Figura 1.a) quelli trasformati mostrano una distribuzione approssimativamente normale, per le singole componenti (Figura 1.b). L'analisi ha individuato una sola Provincia (Ogliastra, individuata dal cerchio in figura 1.b) con frequenze relative alle tre componenti della variabile Condizione professionale, che si discostano significativamente dai valori delle altre 109 Province.

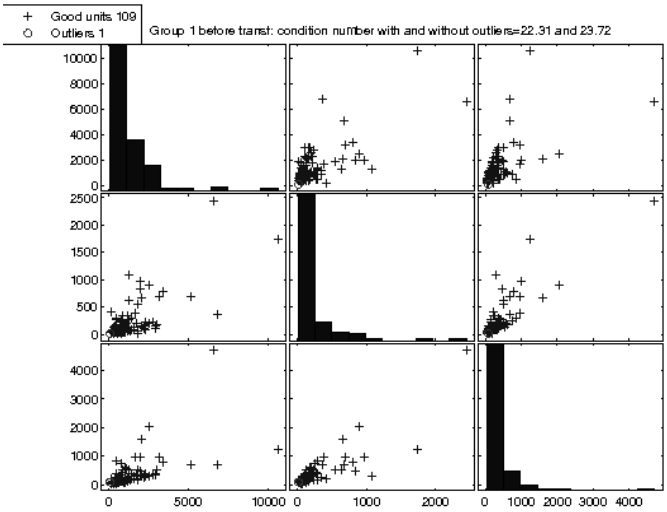
La lontananza dell'osservazione anomala dal centro della distribuzione può essere più o meno estrema a seconda della componente. Infatti, nell'analisi multivariata, ad ogni unità viene associato un valore, quello della distanza di Mahalanobis, come sintesi delle distanze delle singole componenti. Dunque può accadere che un'osservazione anomala abbia valori estremi solo su alcune componenti, mentre risulti vicina al centro della distribuzione, sulle altre.

Oltre all'output grafico, SiRiO fornisce per ogni analisi effettuata:

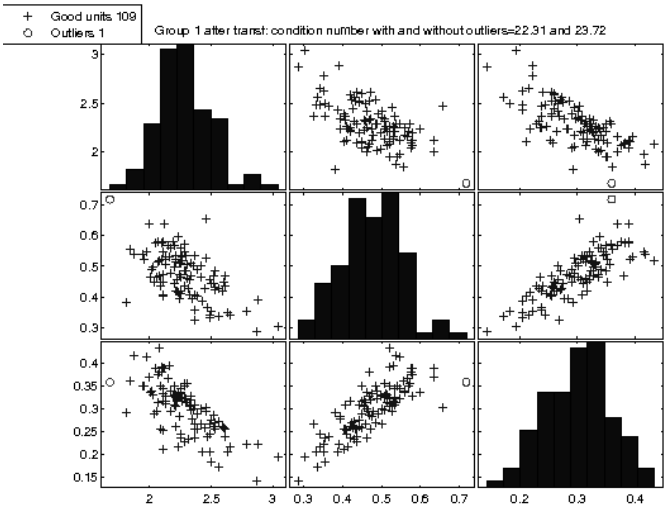
1. risultati di sintesi sul campione analizzato: numero di osservazioni del campione, numero di osservazioni anomale e valori dei parametri della trasformata Box-Cox;
2. risultati puntuali per ogni osservazione: valore iniziale, valore trasformato con Box-Cox, statistica test e relativo p-*value*.

Figura 1 – Grafico fornito da SiRiO per l'analisi FS multivariata: distribuzione univariata e dispersione bivariata delle tre modalità della condizione professionale (Occupato, In cerca di occupazione, Inattivo) relativa alle spose. a) Dati grezzi b) Dati trasformati secondo la procedura FS Box-Cox

a)



b)



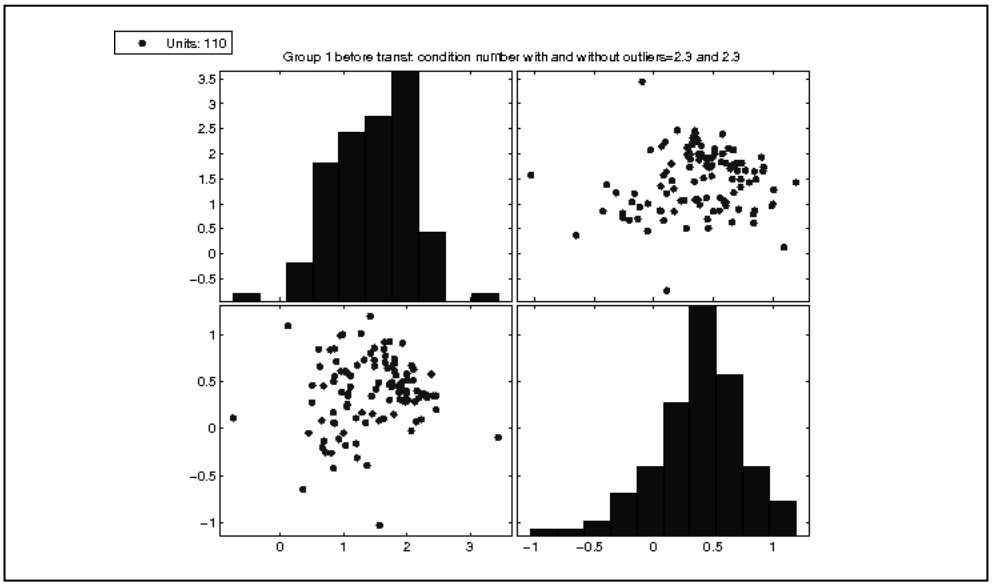
Fonte: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

6.3 Analisi dei risultati: modello composizionale

Per il modello composizionale, SiRiO richiama il codice Python che opera la trasformata ILR sui dati. Le tre modalità della variabile Condizione professionale, dopo l'operazione di chiusura sulle frequenze, ossia la trasformazione delle frequenze assolute in frequenze relative, rispetto al totale dei matrimoni per Comune, vengono ridotte a due componenti ILR. A questo punto, i dati trasformati vengono analizzati dalla procedura Matlab per dati multinormali. La matrice di dati, analizzata dalla procedura, in questo caso contiene in riga le Province e in colonna i valori delle due componenti.

La figura seguente, fornita da SiRiO, riporta i valori relativi alle 110 Province dopo la trasformazione ILR.

Figura 2 – Grafico fornito da SiRiO per l'analisi FS composizionale: distribuzione univariata e dispersione bivariata delle due componenti della variabile condizione professionale relativa alle spose dopo la trasformata ILR



Fonte: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

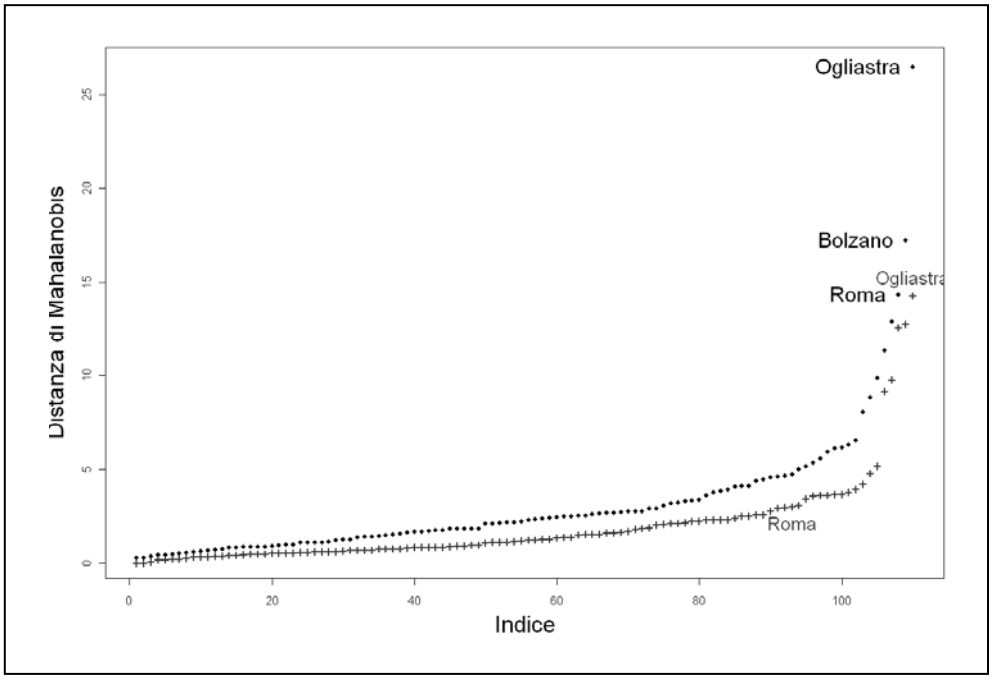
In questo caso la procedura non rileva Province anomale, in particolare la Provincia dell'Ogliastra non risulta significativamente distante dal centro della distribuzione.

E' possibile confrontare i 2 ranghi ottenuti dall'ordinamento delle unità in base alla distanza di Mahalanobis, fornita in output da SiRiO, dall'analisi multinormale e dalla composizionale. In particolare, in figura 3 sono indicate: con il punto le distanze di Mahalanobis ordinate, ottenute dall'analisi multinormale; con la croce le distanze di Mahalanobis ordinate, ottenute dall'analisi composizionale.

Si può notare come il valore relativo all'Ogliastra si allontani sensibilmente dal resto dei valori solo nell'analisi multivariata, basata sulle frequenze assolute delle tre modalità. I valori relativi all'Ogliastra sono infatti sensibilmente inferiori in valore assoluto dalla media delle restanti 109 province (Tavola 1). Se si considerano le frequenze relative o

percentuali, rispetto al totale dei matrimoni per Comune, la differenza fra il profilo dell'Ogliastra e il profilo medio è decisamente meno evidente. La distanza di Mahalanobis relativa all'Ogliastra, nell'analisi composizionale, è comunque l'ultima nell'ordine crescente ma non risulta significativamente lontana dal gruppo principale delle altre osservazioni.

Figura 3 – Distanze ordinate di Mahalanobis della variabile Condizione professionale delle spose relativa alle 110 Province italiane, 2011, ottenute da SiRiO per l'analisi multivariata gaussiana (punto) e per quella composizionale (croce).



Fonte: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

Tavola 1 – Valori assoluti e percentuali della variabile Condizione professionale per le Province di Ogliastra e Roma e media delle osservazioni

	Occupati	In cerca di occupazione	Inattivi	Totale
		Valori assoluti		
Ogliastra (outlier)	74	8	86	168
Roma	10.602	1.732	1.244	13.578
Media (su 109 Province)	**1.326**	**218**	**333**	**1.877**
		Valori percentuali		
Ogliastra	44,0%	4,8%	51,2%	100,0%
Roma	78,1%	12,8%	9,2%	100,0%
Media (su 109 Province)	**70,6%**	**11,6%**	**17,8%**	**100,0%**

Fonte: Rilevazione dei Matrimoni (Modello Istat D.3). Anno 2011.

Nella tavola 1 sono riportati anche i valori relativi alla Provincia di Roma, che permettono di evidenziare le differenze nel significato dell'analisi composizionale rispetto a quella multinormale. Infatti, mentre i valori assoluti relativi a tale Provincia, si discostano sensibilmente (ma non significativamente) da quelli medi, i valori percentuali risultano in linea con il profilo percentuale medio italiano. Ciò è confermato dalla diversa posizione occupata dalla Provincia di Roma, nei due ordinamenti delle distanze di Mahalanobis ottenute dai due modelli. Nel caso multivariato, Roma è tra le Province con distanza di Mahalanobis più elevata (terzultimo valore, in ordine crescente) tanto da risultare lontana dalla maggioranza delle osservazioni (figura 3). Nel caso composizionale invece, il valore della distanza per questa Provincia, risulta vicino alla maggioranza delle distanze relative alle altre unità.

Mentre l'analisi composizionale permette di fare confronti fra le unità indipendentemente dall'ordine di grandezza dei valori assunti delle stesse, nell'analisi multivariata l'unità di misura gioca un ruolo fondamentale nel confronto fra gli elementi dell'insieme. È dunque importante avere chiaro l'obiettivo dell'analisi prima di procedere alla ricerca dei dati anomali.

7. Un confronto tra fonti: il numero di matrimoni per Comune dai modelli Istat D.3 e D7.a

I dati individuali raccolti mediante la Rilevazione dei Matrimoni (D.3) vengono sottoposti a una serie di controlli di tipo quantitativo (copertura della rilevazione) e qualitativo. È proprio nell'ambito dei controlli quantitativi che si utilizzano le informazioni ricavate dalla Rilevazione mensile degli eventi demografici di Stato Civile (D7.a): su base mensile e annuale il numero di matrimoni per Comune risultante dalla raccolta dei modelli individuali deve corrispondere a quello ricavato mediante i modelli riepilogativi.

Nella presente analisi vengono messe a confronto le due indagini precedentemente descritte per individuare i Comuni che abbiano differenze significative nel numero di matrimoni ottenuto mediante le due diverse fonti, nel 2011.

Una prima analisi condotta a livello nazionale, non ha identificato alcun Comune anomalo. Di seguito sono riportati i risultati ottenuti stratificando i Comuni per Provincia.

7.1 Descrizione dei dati

I dati a disposizione riguardano 7.282 Comuni divisi in 110 Province. Questa dimensione iniziale è stata ridotta in base a due criteri:

1. sono stati eliminati i Comuni con un numero di matrimoni inferiore a 10 in entrambi le fonti, poiché nelle operazioni di controllo di copertura della rilevazione, non è ritenuta rilevante una differenza tra le due fonti minore di 10;
2. si sono poi eliminate le Province con meno di 10 Comuni, per rendere la dimensione dello strato d'analisi ragionevole per una ricerca di dati anomali.

La dimensione finale dei dati analizzati è quindi di 3.457 Comuni, per 103 Province, su cui sono state rilevate le due variabili: numero di matrimoni di fonte *Rilevazione mensile degli eventi demografici di Stato Civile* (variabile indipendente) e numero di matrimoni di fonte *Rilevazione dei Matrimoni* (variabile dipendente).

7.2 Analisi dei risultati

L'esplorazione grafica evidenzia un andamento lognormale delle distribuzioni dei dati per Provincia, caratterizzata da una forte asimmetria e dalla positività dei valori assunti. Poiché l'analisi presuppone una distribuzione normale per la variabile dipendente (D.3), è stata applicata la trasformata logaritmica.

L'analisi ha individuato 44 Province con almeno un Comune anomalo. Nella tavola seguente è riportato l'elenco delle 44 Province, il numero dei comuni anomali ed il numero totale dei Comuni nello strato. Per questi Comuni si ipotizza una potenziale situazione di errore, che richiede opportuni approfondimenti.

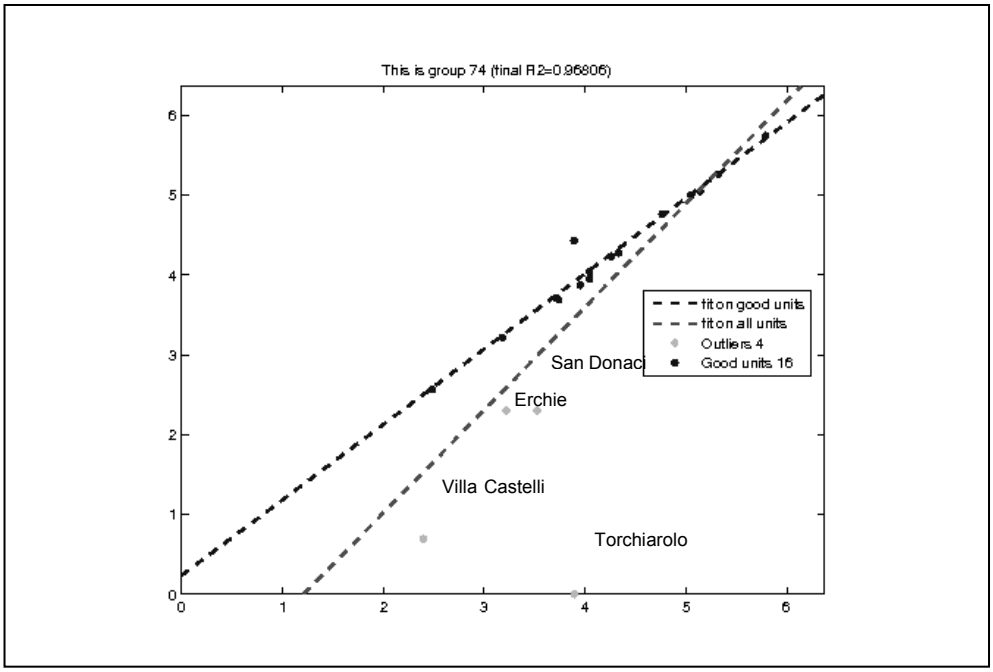
Tavola 2 – Province con almeno un Comune anomalo

Provincia	n. anomali/ tot.strato	*(%)*	Provincia	n. anomali/ tot.strato	*(%)*	Provincia	n. anomali/ tot.strato	*(%)*
Brindisi	4/20	*20.0%*	Salerno	6/141	*4.3%*	Foggia	1/58	*1.7%*
Bari	6/39	*15.4%*	Gorizia	1/24	*4.2%*	Lodi	1/58	*1.7%*
Medio Campidano	3/24	*12.5%*	Monza e Brianza	2/54	*3.7%*	Savona	1/59	*1.7%*
Barletta-Andria-Trani	1/10	*10.0%*	Catania	2/56	*3.6%*	Campobasso	1/62	*1.6%*
Caltanissetta	2/21	*9.5%*	Taranto	1/28	*3.6%*	Sondrio	1/68	*1.5%*
Cagliari	6/66	*9.1%*	Potenza	3/89	*3.4%*	Mantova	1/69	*1.4%*
Rimini	2/26	*7.7%*	Lecce	3/95	*3.2%*	Pavia	2/151	*1.3%*
Teramo	3/41	*7.3%*	Milano	4/129	*3.1%*	Brescia	2/186	*1.1%*
Nuoro	3/45	*6.7%*	Frosinone	2/71	*2.8%*	Roma	1/98	*1.0%*
Latina	2/32	*6.3%*	Siena	1/36	*2.8%*	Bergamo	2/222	*0.9%*
L'Aquila	4/68	*5.9%*	Chieti	2/76	*2.6%*	Avellino	1/112	*0.9%*
Rieti	2/41	*4.9%*	Palermo	2/77	*2.6%*	Udine	1/121	*0.8%*
Trapani	1/22	*4.5%*	Pisa	1/39	*2.6%*	Varese	1/131	*0.8%*
Parma	2/46	*4.3%*	Viterbo	1/51	*2.0%*	Como	1/143	*0.7%*
Piacenza	2/46	*4.3%*	Pesaro e Urbino	1/56	*1.8%*			

Fonte: Rilevazione dei Matrimoni (Modello Istat D.3) e Rilevazione mensile degli eventi demografici di Stato Civile (Modello Istat D7.a). Anno 2011.

Come esempio dell'output grafico fornito da SiRiO, consideriamo il caso della Provincia di Brindisi. In figura 4 è riportato il numero di matrimoni per Comune relativamente alle due fonti: in ascissa è riportata l'informazione aggregata (fonte D7.a) e in ordinata la somma ricavata dalle informazioni individuali (fonte D.3). Dei 20 Comuni analizzati, l'analisi FS ha individuato 4 Comuni anomali, identificati sul grafico dai punti chiari. Le linee tratteggiate rappresentano le rette di regressione stimate con e senza i valori relativi ai Comuni anomali. Lo spostamento dall'una all'altra retta, evidenzia l'effetto delle osservazioni anomale sulle stime dei parametri di regressione.

Figura 4 – Grafico fornito da SiRiO sull'analisi FS di regressione: esempio della Provincia di Brinidsi



Fonte: Rilevazione dei Matrimoni (Modello Istat D.3) e la Rilevazione mensile degli eventi demografici di Stato Civile (Modello Istat D7.a). Anno 2011.

8. Conclusione

Attraverso l'analisi robusta dei dati anomali è possibile individuare le osservazioni che hanno un comportamento diverso dal gruppo cui appartengono. Non esiste l'unità anomala in senso assoluto, ma un'unità è anomala rispetto ad un gruppo di riferimento e alle caratteristiche in studio.

Nel presente lavoro, si è utilizzata la tecnica FS per la ricerca di dati anomali univiariati e multivariati, applicata ai dati Istat sui matrimoni 2011.

L'analisi multivariata, condotta considerando i valori assoluti (modello gaussiano) e i valori relativi (modello composizionale) ha restituito risultati diversi, a sottolineare il ruolo centrale delle caratteristiche dei dati, oggetto di studio.

L'analisi di regressione, condotta per strato, restituisce 92 Comuni anomali. La stessa analisi, condotta senza stratificazione, non restituisce alcun Comune anomalo, a sottolineare l'importanza del gruppo di riferimento.

Riferimenti bibliografici

Amato F., Filippini R., Francescangeli P., Scalfati F. e Toti S. 2013. "SiRiO: una web application per la ricerca di dati anomali multivariati". Poster presentato all'Undicesima Conferenze Nazionale di Statistica, Roma 20-21 febbraio.

Aitchison J. 1986. *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. London (UK): Chapman & Hall Ltd.

Aitchison J. e Shen S.M. 1980. "Logistic-Normal Distributions: Some Properties and Uses". *Biometrika*, 67, 2: 261–272

Atkinson A.C., Riani M. e Cerioli A. 2004. *Exploring Multivariate Data With the Forward Search*. New York: Springer Verlag.

Atkinson A.C, Riani M. 2006. "Distribution Theory and Simulations for Tests of Outliers in Regression". *Journal of Computational and Graphical Statistics*, 15: 460-476.

Becker C. e Gather U. 1999. "The Masking Breakdown Point of Multivariate Outlier Identification Rules". *Journal of the American Statistical Association*, 94: 947-955.

Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueras G. and Barcelò-Vidal C. 2003. "Isometric Logratio Transformations for Compositional Data Analysis". *Mathematical Geology*, 35, 3: 279–300.

Palombi F., Toti S., Filippini R. e Tomeo V. 2011. "A Forward Search Algorithm For Compositional Data". Key Invited Working Paper: Conference of European Statisticians (UNECE), Ljubljana 9-11 maggio.

Reale A., Torti F. e Riani M. 2012. "Robust methods for correction and control of Italian Agriculture Census data". Relazione presentata alla XLVI Riunione Scientifica della SIS, Roma, 20-22 giugno.

Riani M., Atkinson A.C. e Cerioli A. 2009. "Finding an unknown number of multivariate outliers". *Journal of the Royal Statistical Society, Series B Statistical Methodology*, 71: 447–466.

Tallis G.M. 1963. "Elliptical and radial truncation in normal samples". *Annals of Mathematical Statistics*, 34: 940-944.

Toti S., Palombi F. e Filippini R. 2011. "Outlier detection via Compositional Forward Search: application to the preliminary data of the 2011 Italian Agricultural Census". Relazione presentata al convegno: Convegno intermedio SIS, Bologna 8-10 giugno.

Methods for variance estimation under random hot deck imputation in business surveys

Paolo Righi *Stefano Falorsi* *Andrea Fasulo* * †

Abstract

When imputed values are treated as if they were observed, the precision of the estimates is generally overstated. In the paper three variance methods under imputation are taken into account. Two of them are the wellknown bootstrap and Multiple Imputation. The third is a new method based on grouped jackknife easy to implement, not computer intensive and suitable when random hot deck imputation is performed. A simulative comparison on real business data has been carried out. The findings show that the proposed method has good performances with respect to the other two.

Keywords: Bootstrap, Multiple Imputation, Jacknife, Extended DAGJK, Replicate weights, Monte Carlo simulation

1. Introduction

Variance estimation has to take into account an additional complexity element: the unit and item nonresponse that commonly trouble the large scale surveys. Unit nonresponse is customarily handled by forming weighting classes using auxiliary variables observed on all the sampled elements. Then adjusting the survey weights of all respondents within a weighting class by a common nonresponse adjustment factor, with different adjustment factors in different classes (Kalton and Kasprzyk 1986).

Imputation is the commonly used approach to compensate for missing (item nonresponse) or invalid values in sample surveys (Kalton and Kasprzyk 1986). In the paper the random hot deck imputation is considered.

When unit and item nonresponse correction is performed extra variability is introduced in the sampling errors. Modifications of Taylor and resampling methods for contemplating unit nonresponse are quite straightforward while item nonresponse is a ticklish issue. Analyses performed on imputed values treated as if they were observed, can be misleading when estimates of the variance do not include the variability component due to imputation. As a result, the precision of estimates is overstated, and subsequent statistical analyses can be misleading (e.g., confidence intervals have lower than nominal levels).

The approaches proposed in literature to obtain valid variance estimators in presence of imputed data are divided according to several classifications. A first common classification distinguishes among linearization (or Model Assisted techniques see Särndal 1992), resampling (Shao and Tu 1995; Wolter 2007) and the Multiple Imputation (Rubin 1987) methods being the first two categories used for the complete data variance estimates as well.

The resampling techniques in presence of item nonresponse can be divided according to the standard classification used for the complete data (Wolter 2007). Then, bootstrap (Efron 1994; Shao and Sitter 1996; Saigo et al. 2001; Shao 2003), balanced repeated replication

* Italian National Statistical Insitute (Istat), e-mail: parighi@istat.it; stfalors@istat.it; fasulo@istat.it.

† The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

(Rao and Shao 1999), random group (Shao and Tang 2001) and jackknife (Rao and Shao 1992; Rao 1996; Yung and Rao 2000; Chen and Shao 2001; Skinner and Rao 2002; Saigo 2005) methods, can be distinguished.

In the literature there is not a common judgement on which is the best approach or method. This is the main reason why all these methods are investigated in a lot of different contexts involving the sampling design, the estimator, the domains of interest and the imputation process. Furthermore the dimension of the survey and the type and the number of parameters to be estimated have to be taken into account. For large scale surveys, the function to be estimated, the complexity of sampling design, the imputation procedure and the cost-effectiveness issues drive the choice of a specific variance estimator.

The paper focuses on the operational conditions for implementing the methods in the data production in Official Statistics. In particular two issue are raised: the setting up of theoretical framework for applying the method; the computational aspect, that is always troublesome for large scale surveys.

Three methods are deeply investigated: the bootstrap under imputation, the Multiple Imputation and finally a new method based on a jakknife technique. The linearization is kept out because can be problematic to consider in a standardized data production process in which the timeliness is a pressing data quality dimension.

Section 2. introduces the three methods. In particular, section 2.1 gives a literature review on the jackknife techniques with the random hot deck imputation. Section 2.2 proposes a new variance estimator taking into account the item non responses based on a grouped jackknife technique. This is the innovative output of the paper. Section 2.3 and 2.4 are respectively devoted to the bootstrap estimator and Multiple Imputation procedure. In section 3. we compare the three methods by means of a Monte Carlo simulation based on real business survey data. Some concluding remarks are given section 4.

2. Bootstrap, Multiple Imputation and jacknife techniques with random hot deck imputation

The variance estimation process under imputation depends on the kind of imputation procedure has been used. For instance, if a jackknife type estimator has been chosen, the form of the final estimator can change according to the imputation.

In the paper the hot deck procedure is taken into account. This class of methods is one of the most popular in the survey sampling. Various specifications of the method are proposed in the literature. In the simplest form of hot deck imputation, a random sample of size $\bar{r}$ (the number of nonrespondents) is selected from the sample of respondents to an item y, and the associated item y values are used as donors. The accuracy of imputation depends on the nonresponse model (the imputation classes) and on the simple or weighted random selection of the donors.

Since the hot deck imputation is a form of regression imputation (Kalton and Kasprzyk 1986) the analysis of the variance estimators with this imputation technique is not so restrictive.

Eventually, we consider the standard imputation procedures based on real observed values. Hence, we do not explore variance estimators for the imputation procedures where imputed values depend on variables already imputed in previous steps (Raghunathan et al. 2001). That means the hot deck imputation classes are defined on the observed values.

2.1 Jackknife variance methods under imputation

Jackknife variance estimation in presence of item non response has been extensively studied in literature.

One of the first papers was made by Burns (1990). He proposed to perform a new imputation for each jackknife sample according to the same procedure applied on the overall sample. Then, being n the sample size, the procedure needs $n + 1$ imputation steps.

Rao and Shao (1992) showed that the procedure can lead to serious overestimation for large sample size. They propose a consistent jackknife variance estimator in presence of imputed data (for the Horvitz-Thompson estimator) by means of hot deck methods assuming equal response probabilities within imputation classes. The method is suitable for stratified random sampling and stratified multistage sampling design even in the general case in which the imputation classes cut across the sampled clusters. The approach, named *adjusted jackknife*, performs only one single imputation on the full sample and it adjusts the imputed values for each pseudo-replicate before applying the standard jackknife variance formula for stratified design. Then, the technique is much more efficient in terms of computation with respect to the Burns approach.

The consistency of the adjusted jackknife (for smooth functions such as totals and means) is shown assuming equal response probabilities within imputation class and performing independently within each class the hot deck imputation.

In presence of variable inclusion probabilities in the stratum (such as in the multistage sampling designs) the properties holds when the weighted hot deck is implemented. Weighted hot deck select a donor from the imputation cell with probability proportional to the sampling weight. Conversely if a simple random sampling is used to select a donor then the estimator, under imputation, of the target parameter will be biased and it's variance estimator as well.

The adjusted jackknife method needs for each unit an imputation flag on the data set.

Further enhancements of the adjusted jackknife are given by Rao and Sitter (1995) that examine the jackknife with ratio imputation in the model based framework. Rao (1996) gives new results about the adjusted jackknife variance estimator with imputed survey data. As far simple random sampling is concerned, Rao shows some properties when ratio and regression imputation are used for estimating totals or means.

In particular, the adjusted jackknife variance estimator is design consistent and it is also design and model unbiased under the imputation model.

Regarding stratified multistage sampling Rao shows the properties of the adjusted jackknife variance estimator when the mean imputation in the imputation classes is used. Within each imputation class, the weighted mean of the interest variable computed on the respondents is assigned to all missing responses. The technique assumes the best predictor of the missing values is obtained by a homoscedastic mean superpopulation model. If the model holds for the respondents and under uniform response within each class the adjusted jackknife is design consistent.

Yung and Rao (2000) extend the analysis of the adjusted jackknife variance estimator under imputation when poststratified or generalized regression estimator are used. The weighted mean imputation and weighted hot deck stochastic imputation within imputation classes have been studied.

When the weighting classes are the poststrata, the estimator and corresponding jackknife variance estimator are simply computed on the respondents. The authors show also the jackknife estimator when weighting classes cut across the poststrata and they give the proof of the asymptotic consistency. The property holds also when generalized regression estimator and weighted hot deck imputation are used.

Furthermore, the authors investigate the properties of the jackknife variance estimators under weighting adjustment for unit non response. The properties are essentially studied in

the stratified multistage sampling design. In case of unit non response, the authors assume a set of weighting classes. Within each class a uniform response mechanism is supposed. Moreover, non response adjustment is performed before the poststratification adjustment so that the known totals are benchmarked.

This comprehensive theoretical framework encompassing general point estimators and unit and item non response makes jackknife techniques appealing .

2.2 The modified Extended DAGJK under hot deck imputation

The adjusted jackknife method has a remarkable reduction of computational effort with respect to the imputation procedure but it is still computer intensive, because of the number of replications in the estimation procedure. For reducing the computations a commun strategy is to combine units into variance strata and perform a grouped jackknife instead of a standard delete one-unit jackknife (Rust 1985; 1986; Rust and Rao 1996). The family of techniques delete groups of units rather than one unit at time for reducing calculation effort. The creation of a replicate group can be done within design stratum or, combining design strata into superstratum, taking groups of units within superstratum that cut across design strata.

There is limited theoretical guidance on how the grouping should be done and much is based on heuristic knowledge. A common assumption in the literature is to form equal sized groups. Moreover, to take into account a nonnegligible sampling fraction, it can be useful to form superstrata with design strata having similar sampling fractions (Valliant et al. 2008). Finally, grouped jackknife methods distinguishes themself by the computation of the replicate weights as well, augmenting the possible variance estimators.

Currently there is no empirical evidence showed in literature, suggesting the best grouped jackknife. Then we should underline the importance of having evidence of the empirical properties of these methods in practical applications in the Official Statistics context.

As concern grouped jackknife methods taking into account item nonresponse few have been written in the literature. Brick et al. (2005) show a grouped adjusted jackknife according to Rao and Shao approach in case of Horvitz-Thompson estimator. Di Zio et al. (2008) propose the definition of the the Rao and Shao adjustment and the Delete A Group Jackknife (DAGJK). Miller and Kott (2011) investigate a DAGJK with imputed data with a different approach.

In the following we introduce a new method combining the DAGJK technique with the Rao and Shao adjustment.

Delete A Group Jackknife (Kott 1998; 2001) is a variance estimation technique computationally less intensive than classical jackknife and it can be applied also in case of large scale surveys. DAGJK is within the strategies aiming at reducing the number of jackknife replications, while maintaining adequate precision of variance estimates. It assumes an unique superstratum formed by all the design strata and the replicate groups have units belonging to different design strata. Then, the method does not present implications on the definition of the groups and does not require analysis to form superstrata. This analysis can become cumbersome for large scale and complex business surveys and may affect the timeliness of data production. For this reason variance estimation techniques implemented by a sort of automated process and leading to *good* statistical results may be preferred to better techniques but more complex to be implemented.

Consider the stratified simple random sampling commonly used in the business surveys, where in each stratum h are included N_h units. A sample of $n_h \geq 2$ units is drawn from each stratum independently across strata. Let $d_{hk}(> 0)$ be the basic weight of unit k in stratum h, denoted as hk, the estimator of the parameter of the total θ is $\hat{\theta} = \sum_{hk \in s} d_{hk} y_{hk}$, being y_{hk} the value of the variable of interest. The DAGJK technique divides the overall or

parent sample, s into Q mutually exclusive *replicate* or *random* groups, hereinafter denoted by $s^1, \ldots, s^q, \ldots, s^Q$. Given the subsample s^q, the sample sizes in the strata are indicated as $n_1^q, \ldots, n_h^q, \ldots, n_L^q$. The complement of each s^q is called the *jackknife replicate group* $s^{(q)} = s - s^q$, being $n_1^{(q)}, \ldots, n_h^{(q)}, \ldots, n_L^{(q)}$ the strata sample sizes of $s^{(q)}$. The variance estimator is based on the following jackknife procedure:

1. units are randomly ordered in each stratum;
2. from this ordering the units are systematically allocated into Q groups;
3. for each unit hk, Q different sampling weights (*replicate sampling weights*) are computed;
4. given the qth set of the replicate weights the qth replicate estimate is $\hat{\theta}^{(q)} = \sum_{hk \in s} d_{hk}^{(q)} y_{hk}$.
   where $d_{hk}^{(q)}$ denotes the qth replicate weight of unit hkth ;
5. the DAGJK variance estimation is given by

$$v(\hat{\theta}) = \frac{Q-1}{Q} \sum_{q=1}^{Q} (\hat{\theta}^{(q)} - \hat{\theta})^2. \tag{1}$$

The standard DAGJK replicate weights are given by

$$d_{hk}^{(q)} = \begin{cases} d_{hk}, & \text{when } k \in h \text{ and no unit of } h \text{ belongs to group } q \\ 0, & \text{when } k \in q \\ [n_h/(n_h - n_h^q)]\, d_{hk} & \text{otherwise.} \end{cases} \tag{2}$$

There is not an optimal value for Q. When the number of random groups has to be chosen it needs to consider that increasing the number of random groups the variability of the variance estimation is restricted but the computational effort is augmented. In general, it is common practice a choice between 15 and 80 (Kott 1998; Rust 1985), considering that when Q is greater than 15 the Student's t distribution is approximated quite good by the normal distribution.

The statistical properties in terms of bias and variability of the variance estimates depends on the values of Q, n_h and in the case of WOR designs on the sampling fraction in each stratum. In the latter case, with large sampling fractions the (1) produces conservative variance estimates. Neverthless, Kott (2001) shows that even if the finite population correction factor is negligible but $n_h < Q$ for some strata the (1) is still an upward biased variance estimator. For instance, if all $n_h > 5$ but $n_h < Q$ the relative bias of (1) with weights (2) for Horvitz-Thompson estimator is at most 20%. The upperbound of bias is given by $[Q/(Q-1)] max_h \{n_h/[n_h - 1]\}$ which is itself bounded by $max_h \{n_h/[n_h - 1]\}$. The relative upward bias is equal to $max_h \{n_h/[n_h - 1]\} - 1 = max_h \{1/[n_h - 1]\}$.

Kott developed a different expression of the replicate weights defining the Extented DAGJK (EDAGJK). For the Horvitz-Thompson estimator the replicate weights of EDAGJK assume the following expression,

$$d_{hk}^{(q)} = \begin{cases} d_{hk}, & \text{when } k \in h \text{ and no units of } h \text{ belongs to group } q \\ d_{hk}[1 - (n_h - 1)Z] & \text{when } k \in q; \\ d_{hk}(1 + Z), & \text{otherwise.} \end{cases} \tag{3}$$

where $Z^2 = Q/[(Q-1)n_h(n_{h-1})]$.

With the Greg estimator the replicate weights are given by $w_{hk}^{(q)} = d_{hk}^{(q)}\gamma_{hk}^{(q)}$ and the replicate qth GREG estimate is $\hat{\theta}_{greg}^{(q)} = \sum_{hk\in s} y_{hk} w_{hk}^{(q)}$. The correction factor $\gamma_{hk}^{(q)}$ may be calculated according different ways. Let consider the following expressions:

$$\gamma_{hk}^{(q)} = 1 + \left(\mathbf{X} - \sum_{hk\in s} \mathbf{x}_{hk} d_{hk}^{(q)}\right)\left(\sum_{hk\in s} \frac{\mathbf{x}_{hk}\mathbf{x}'_{hk} d_{hk}^{(q)}}{c_{hk}}\right)^{-1} \frac{\mathbf{x}_{hk}}{c_{hk}}. \tag{4}$$

and

$$\gamma_{hk}^{(q)} = \gamma_{hk} + \left(\mathbf{X} - \sum_{hk\in s} \mathbf{x}_{hk} d_{hk}^{(q)}\gamma_{hk}\right)\left(\sum_{hk\in s} \frac{\mathbf{x}_{hk}\mathbf{x}'_{hk} d_{hk}^{(q)}\gamma_{hk}}{c_{hk}}\right)^{-1} \frac{\mathbf{x}_{hk}\gamma_{hk}}{c_{hk}}. \tag{5}$$

Kott (2006) offers some suggestions on the factor has to be used.

In order to take into account item nonresponse in variance estimation, we propose a modified version of EDAGJK based on the Rao and Shao adjustment for hot deck imputation.

The modified variance estimator is :

$$v(\hat{\theta}_I) = \frac{Q-1}{Q}\sum_{q=1}^{Q}(\hat{\theta}_I^{(q)} - \hat{\theta}_I)^2 \tag{6}$$

where,

$$\hat{\theta}_I = \sum_{hk\in s_R} w_{hk} y_{hk} + \sum_{hk\in s_{\bar{R}}} w_{hk} y_{hk}^*, \tag{7}$$

is the estimator with imputed hot deck values y_{hk}^*, being s_R and $s_{\bar{R}}$ the sample of respondents and non respondents.

$\hat{\theta}_I^{(q)}$ is defined as

$$\hat{\theta}_I^{(q)} = \sum_{g=1}^{G}\left\{\sum_{hk\in s_{R_g}} w_{hk}^{(q)} y_{hk} + \sum_{hj\in s_{\bar{R}_g}} w_{hj}^{(q)}\left(y_{hj}^* + \hat{y}_{Rg}^{(q)} - \bar{y}_{Rg}\right)\right\} \tag{8}$$

in which: g ($g = 1, \ldots, G$) indicates the gth imputation cell; s_{R_g} and $s_{\bar{R}_g}$ are respectively the respondents and non repondents in the cell g; $w_{hk}^{(q)}$ are the replicate base or Greg weights. Finally, $\hat{y}_{Rg}^{(q)} = \sum_{hj\in s_{R_g}} w_{hj}^{(q)} y_{hj} / \sum_{hj\in s_{R_g}} w_{hj}^{(q)}$ and $\bar{y}_{Rg}^{(q)} = \sum_{hj\in s_{R_g}} w_{hj} y_{hj} / \sum_{hj\in s_{R_g}} w_{hj}$.

Note that the imputation procedure is performed only on the parent sample.

2.3 Bootstrap variance methods under imputation

Bootstrap method in presence of imputed data has been deeply studied in the relevant papers by Efron (1994) and Shao and Sitter (1996).

Starting from the evidence that the naive approach (treating the imputed values as observed values and using the standard bootstrap) does not capture the inflation in variance due to imputation and serious variance underestimation is possible they showed some procedures

together with reimputing bootstrap datasets. In particular circumstances such approaches define a valid approximation to the distribution of $\hat{\theta}_I$. Let Y be the observed data set, being the estimator $\hat{\theta} = f(Y)$, and let $\hat{\theta}_I$ the estimator with imputed values given by the (7), the bootstrap variance estimator replaces B bootstrap estimates, $\hat{\theta}^{(b)}$ $(b = 1, \ldots, B)$, by $\hat{\theta}_I^{(b)}$, where the index (b) denote the estimate based on the bth resampling. Each estimate $\hat{\theta}_I^{(b)}$ is computed according to the same procedure implemented for the overall sample.

For the bth bootstrap sample the procedure by Shao and Sitter can be described as follows:

1. draw a simple random sample $\left\{ y_{hk}^{(b)} : k = 1, \ldots, n_h - 1 \right\}$ (where y_{hk} denotes the value of y for the unit k beloging to stratum h) with replacement from the sample $\{\tilde{y}_{hk} : k = 1, ..., n_h\}$, independently across the strata, where $\tilde{y}_{hk} = \{y_{hk} : (hk)\} \in s_R \cup \{y_{hk}^* : (hk) \in s_{\bar{R}}\}$
2. apply the same imputation procedure used in costructing the imputed survey data. Denote the bootstrap analogue of $\hat{\theta}_I$ by $\hat{\theta}_I^{(b)}$

$$\hat{\theta}_I^{(b)} = \sum_{s_R^{(b)}} w_k^{(b)} y_k + \sum_{s_{\bar{R}}^{(b)}} w_k^{(b)} y_k^{*(b)}. \tag{9}$$

   where $y_k^{*(b)}$ is the imputed value using the bth bootstrap data and $w_k^{(b)}$ is $n_h/(n_h - 1)$ times the survey weight of unit k.

The bootstrap variance estimator $v(\hat{\theta}_I)$ when has no explicit form may be approximated by

$$v(\hat{\theta}_I) \approx \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\theta}_I^{(b)} - \bar{\hat{\theta}}_I^{(b)} \right)^2 \tag{10}$$

in which $\bar{\hat{\theta}}_I^{(b)} = (1/B) \sum_{b=1}^{B} \hat{\theta}_I^{(b)}$ with $b = 1, \ldots, B$.

Efron (1994) shows that the process generates asymptotically valid variance and distribution estimator for complex sampling designs. The same result was established in Shao and Sitter (1996) for stratified sampling with large n_h. The assumption of negligible sampling fraction in each stratum means that the procedure give consistent variance estimator when with replacement sampling design is implemented. Nevertheless, in business surveys is not uncommon to define very detailed strata with small sample size. In such cases some precautions must be taken. The problems is known also when complete data set is used and a stratified random sampling without replacement is the utilized design. A simple and heuristic approach is to collapse the original strata forming variance strata. Shao and Sitter investigate some bootstrap methods that deal with the small n_h's. In particular they analyze the rescaling bootstrap proposed by Rao and Wu (1988) showing that the method does not work with imputed values.

The simulation study carried out by the authors produces valid approximation with deterministic imputation. In case of random imputation, such as hot deck imputation, upward biased estimates are obtained when some n_h are very small.

To overcome the problem Saigo et al. (2001) have developed a third type of modified bootstrap, the repeated half-sample bootstrap, which together with reimputing bootstrap data sets produces a valid approximation of the distribution of $\hat{\theta}_I$, regardless of whether the imputation is random or not and whether n_h is small or not. In the paper by Shao (2003) are well illustrated the different bootstrap methods and their associated problems.

Eventually, in Shao and Sitter (1996) for avoiding the complete reimputation process for each replication has been proposed a slightly different bootstrap method. Such technique preserves the asymptotic properties except for variance estimate of quantiles.

2.4 Multiple Imputation

Multiple Imputation (MI) was first proposed and thoroughly described in Rubin (1978). More recently the book by Little and Rubin (2002) offers a concise and complete description of the method.

MI is a procedure replacing each missing value by an ordered vector composed of $M \geq 2$ possible values. The ordering assumes that the first components of the vectors for the missing values are used to create one completed data set, the second components of the vectors are used to create the second completed data set and so on. Each completed data set is investigated using standard complete-data methods. To analyze the repetitions within one imputation model to yield a valid inference under the posited reasons for missing data, the M complete-data based on the M repeated imputations are then combined to create one repeated-imputation inference.

Let $\hat{\theta}_m$ and W_m $(m = 1, \ldots, M)$ be the mth complete-data estimate and its variance of the parameter θ obtained by imputation under one model for nonresponse. The MI estimate is given by

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m. \tag{11}$$

The variability associated with this estimate has two components: the average within imputation variance

$$\bar{W}_M = \frac{1}{M} \sum_{m=1}^{M} W_m \tag{12}$$

and the between-imputation component,

$$B_M = \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \bar{\theta}_M\right)^2 \tag{13}$$

where with vector θ the $(\cdot)^2$ is replaced by $(\cdot)^T(\cdot)$.

The total variability associated to $\bar{\theta}_M$ is given by

$$T_M = \bar{W}_M + \frac{M+1}{M} B_M. \tag{14}$$

With scalar parameter the approximate reference distribution for interval estimates and significance tests is a t distribution

$$\left(\theta - \bar{\theta}_M\right) T_M^{-1/2} \sim t_d, \tag{15}$$

where the degrees of freedom,

$$d = (M-1)\left(1 + \frac{1}{M+1} \frac{\bar{W}_M}{B_M}\right)^2. \tag{16}$$

are based on the Satterthwaite approximation (Rubin and Schenker 1986) . An improved expression of the degree of freedom for small data sets is given, for example, in Little and Rubin (2002).

An estimate of the fraction of missing information γ_M about θ due to nonresponse is given by

$$\hat{\gamma}_M = (1 + 1/M)\frac{B_M}{T_M}. \tag{17}$$

Rubin (1987), shows that the relative efficiency of an estimate based on M imputations to one based on $M = \infty$ number of imputations is approximately $1 + \gamma/M$ to 1, where γ is the rate of missing information. Assuming a fraction of 50% missing information an estimate based on $M = 3$ imputations has a standard error that is about 8% higher than one based on $M = \infty$, because $\sqrt{1 + 0.5/3} = 1.0801$. Schafer (1998) states that unless the fraction of missing information is higher than 50% there is little benefit in using more than 5 to 10 imputations.

See Kim et al. (2006) for more details on bias of the MI.

2.4.1 Multiple Imputation and the single imputation procedure: the ABB method

A basic issue of the MI is the single imputation method repeated M times. Theoretically, the method assumes the Y_{mis}' are M repetitions from the posterior predictive distribution of Y', each repetition being an independent drawing of the parameters and missing values under appropriate Bayesian models for the data and the posited response mechanism. In practice, three aspects of the imputation method have to be considered:

- if the underlying imputation model is explicit or implicit;
- if the underlying imputation model is ignorable or nonignorable;
- if the imputation methods is proper or not proper.

The first two concepts are typically dealt with in the single imputation approach as well. Commonly in the Official Statistic ignorable model are assumed, while we focus on the random hot deck method that falls in the method using implicit imputation modeling.

Here the concept of proper/not proper method is introduced. Imputation procedures that incorporate appropriate variability among the repetitions within a model (explicit or implicit, ignorable or nonignorable) are called proper (Rubin 1987). The reason for using proper imputation methods is that they properly reflect sampling variability when creating repeated imputations under a model, and as a result lead to valid inferences. For example, assume ignorable nonresponse so that respondents and nonrespondents with a common value auxiliary variable X (i.e. imputation cell) have Y values only randomly different from each other. Randomly drawing imputations for nonrespondents from matching respondents' Y values ignores some sampling variability. This variability arises from the fact that the sampled respondents' Y values at X randomly differ from the population of Y values at X. Properly reflecting this variability leads to repeated imputation inferences that are valid under the posited response mechanism. In particular Rubin and Schenker (1986) examined the hot deck procedure with MI. The imputation method assumes within the hot deck cells responses are missing randomly and the Y's are independent random variables with common mean and variance. For each unit having a missing value M values are imputed. The authors shown the standard hot deck procedure is not proper and variance with MI performs a variance underestimate. They proposed the Approximate Bayesian Bootstrap (ABB) for simple random sampling with hot deck imputation and MI method. Such technique can be viewed as a hot deck imputation method in the MI context (Kim and Fuller 2004; Kim et al. 2004).

Let us consider a collection of n units in the specific hot deck cell where there are n_r respondents and $n_{nr} = n - n_r$ nonrespondents. The ABB creates M ignorable repeated imputations as follows. For $i = 1, \ldots, M$ create n possible values of Y by first drawing n values at random with replacement from the n_r observed values of Y, and second drawing the n_{nr} missing values of Y at random with replacement from those n values. The drawing of n_{nr} missing values from a possible sample of n values rather than the observed sample of n_r values generates appropriate between imputation variability, at least assuming large simple random samples at X showing that is a proper method.

The ABB approximates the Bayesian Bootstrap by using a scaled multinomial distribution to approximate a Dirichlet distribution.

When the imputation cells cut across the sampling strata, unequal inclusion probabilities should be involved in the procedure. Nevertheless, no literature discusses the applications of ABB in this context. Some authors suggest (Brick et al. 2005) to disregard the unequal inclusion probabilities in the ABB. In the simulation below the Y values are drawn at random with equal inclusion probabilities.

3. Specialized results

A Monte Carlo simulation has been carried out for comparing the modified EDAGJK with bootstrap and MI. In the simulation a standard sampling strategy for the business survey has been implemented. The analysis of the results focuses on the statistical properties both with applicability of the methods in case of a complex survey sampling typically conducted by a National Statistical Institute.

3.1 The population and the sampling strategy

The simulation is based on the real data of the 2008 Italian enterprises belonging to the economic activity 162 according to the Statistical Classification of Economic Activities in the European Community NACE Rev.2 3-digit (number of units $N = 21,231$). This subpopulation is surveyed in the Small and Medium Enterprises (SME) survey. The SME is a yearly survey investigating the profit-and-loss account of enterprises with less than 100 employed persons, as requested by SBS EU Council Regulation n. 58/97 (Eurostat, 2003) and n. 295/2008. The Italian target population of the SME survey is about 4.5 millions active enterprises.

The following target variables have been considered: Changes of inventory of finished and semifinished products (CIF); Purchase of commodities (PUC); Operating expenses for administration (OEX); Labour cost (LCO). The values of the four variables have been taken from the balance sheets (administrative data) for the whole population.

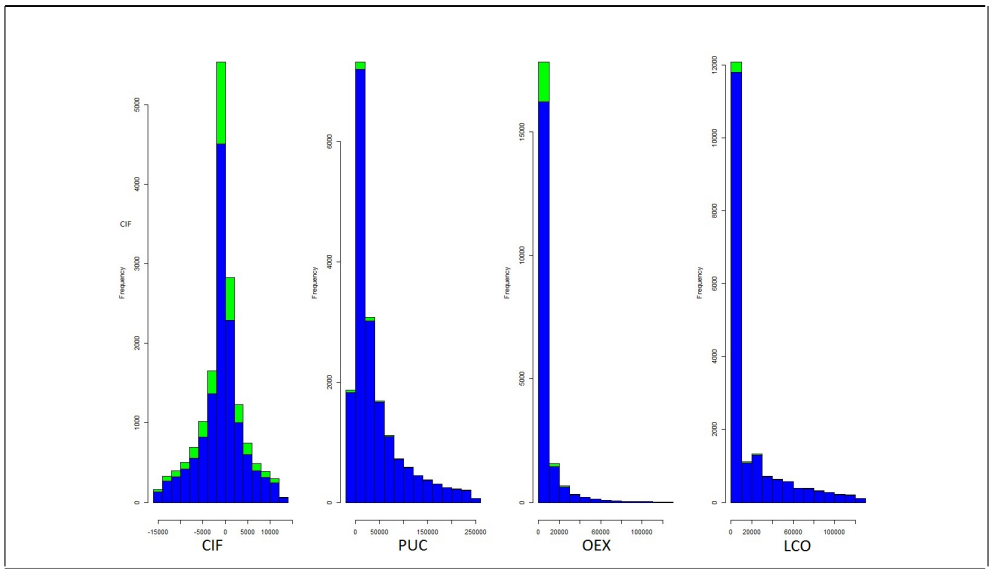
Table 1 gives some summary statistics.

Table 1 - Summary of the target variables

Variables	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
CIF	-869300	-3750	-130	-2422	1490	322700
PUC	-6631	5514	19710	41880	56230	247900
OEX	0	441	1601	6201	5294	126200
LCO	0	0	0	16700	24270	126500

Figure 1 shows that CIF variable has a symmetric distribution while especially OEX and LCO have highly skewed distributions.

Figure 1 - Distribution of the variables of interest and frequncy of non responses (light colour)



The simulation takes into account of a simplifyed version of the current sampling strategy used in the SME survey. A stratified simple random sampling design and a calibration estimator have been considered.

Strata are obtained by crossing the size classes and the regions according to the Nomenclature of territorial units for statistics NUTS 1 defined by EU. Hence, 20 strata are obtained as aggregation of the original strata of SME survey. The sample allocation in each stratum is taken from the allocation of 2008 SME survey. Table 2 shows the population and sample distribution in each stratum. The overall sample size is $n = 908$ enterprises.

The estimator calibrates the sampling weights to the number of enterprises and the number of employed persons at NACE Rev.2 4-digit, size class and NUTS 1 region.

The linear distance function (generalized regression estimator) is considered. Actually, the logit distance function is used in the SME survey, because it produces nonnegative calibrated weights. Nevertheless the logit distance has two drawbacks: the convergence is not guaranteed; it requires a time spending iterative procedure to obtain the calibrated weights. For these reasons linear distance function has been preferred in the simulation. We point out that the calibration estimators with linear and logit distance function converge asymptotically and the simulation results with the linear distance will be coherent with the ones obtained with the logit distance function.

The main task of the simulation study is to compare different methods of variance estimation for estimators of totals in a complex context, usual in the business survey, such as: stratified simple random sampling, imputation for item nonresponse and calibration estimator. The item nonresponses are imputed by means of random hot-deck procedure.

Besides this sampling strategy the simulation regards the variance of the Horvitz-Thompson (HT) estimator.

Table 2 - Population, sample distribution and missing rates (%) in the design strata

Strata (NUTS 1 region by size class)	Number of enterprises in the population	Number of enterprises allocated in the sample	Sampling rate	Missing rate by variable			
				CIF	PUC	OEX	LCO
NORTH-WEST:(size class<8)	5241	107	0.02	16.75	1.28	7.52	1.97
NORTH-WEST:(9<size class<18)	425	39	0.09	14.59	1.41	3.29	0.00
NORTH-WEST:(19<size class<28)	83	17	0.20	28.92	2.41	8.43	0.00
NORTH-WEST:(size class>29)	38	15	0.39	13.16	5.26	5.26	0.00
NORTH-EAST:(size class<8)	5087	89	0.02	18.28	2.54	9.44	3.07
NORTH-EAST:(9<size class<18)	550	48	0.09	14.18	3.45	6.00	0.18
NORTH-EAST:(19<size class<28)	129	44	0.34	15.50	4.65	7.75	0.00
NORTH-EAST:(size class>29)	55	20	0.36	19.35	3.23	3.23	0.00
CENTER:(size class<8)	3699	145	0.04	19.60	1.14	9.22	1.57
CENTER:(9<size class<18)	297	34	0.11	19.53	2.69	5.05	0.00
CENTER:(19<size class<28)	63	35	0.56	19.35	3.23	3.23	0.00
CENTER:(size class>29)	31	17	0.55	20.79	2.60	10.43	2.78
SOUTH:(size class<8)	3386	128	0.04	17.61	1.14	5.11	0.00
SOUTH:(9<size class<18)	176	45	0.26	13.51	0.00	0.00	0.00
SOUTH:(19<size class<28)	37	23	0.62	26.67	0.00	0.00	0.00
SOUTH:(size class>29)	15	8	0.53	18.64	1.16	10.21	2.83
ISLANDS:(size class<8)	1803	50	0.03	13.48	0.00	3.37	1.12
ISLANDS:(9<size class<18)	89	26	0.29	4.76	0.00	0.00	0.00
ISLANDS:(19<size class<28)	21	16	0.76	4.76	0.00	0.00	0.00
ISLANDS:(size class>29)	6	2	0.33	16.67	16.67	16.67	16.67
		Missing rate	Average	18.36	1.88	8.73	2.19

3.2 Item nonresponse model

The item nonresponses have been generated seeking to reproduce the item nonresponse pattern of the business surveys.

The SME survey suffers from item nonresponses. Actually the survey has not flags for item nonresponses and the item nonresponses are denoted by zero values.

To find out when a zero value means a real zero or a missing value, the 2008 SME data have been linked with the 2008 administrative data; the zero SME values corresponding to a non zero value in the administrative data have been identified as missing values.

A regression tree model (rpart R package) has been applied for estimating the relationship between response propensity and outcome-related auxiliary variables known for the whole population. To create missing values, response indicators were assigned to the units within nonresponse cells defined by the regression tree. Within a given response cell, units were assigned at random to be missing or nonmissing at a specified rate. The mechanism generating missingness assumes that there is a uniform response probability within each cell. This is an usual assumption for the nonresponse model even though generally a more complex unknown item nonresponse model holds. When the real nonresponse model disagrees with the working model used for the imputation, the estimates are biased. However, the simulation is focused on the variance estimate and then we define an experimental context in which the point estimates are unbiased, for not creating confounding evidences.

Table 2 shows the strata missing rates (the last four columns) for the variables of interest. Then, three type rates of item nonresponse appear: high for the CIF variable with average equal to 18.36%, medium for the OEX variable with 8.73% and low for the PUC and LCO variables with about 2% of nonresponse rate. Figure 1 underlines that for CIF variable the missing rate increases when the frequency enlarges. For the other three variables the missing rate is concentrated on the smaller values of the variable.

We checked the unbiasedness of the estimator after the hot-deck imputation computing the empirical relative bias

$$RB(\hat{\vartheta}) = \frac{1}{C}\sum_{c=1}^{C}\frac{(\hat{\vartheta}_c - \theta)}{\theta} \tag{18}$$

being $\hat{\vartheta}_c$ the estimate from the sample c drawn according to the sampling design of section 3.1 and C the number of drawn samples. To obtain a nearly zero relative bias $C = 10,000$ samples have been selected.

Table 3 shows negligible bias for all estimates: e. g. the $RB(\hat{\vartheta})\%$ are lower than 1% except for the variable CIF when calibration with imputed data is considered (1.4%).

Table 3 - Relative bias($RB(\hat{\vartheta})\%$) of the estimators

Estimators	CIF	PUC	OEX	LCO
HT with imputation	0.72	0.49	0.94	-0.09
CALIBRATION with imputation	1.40	0.51	0.97	-0.10

3.3 Results of the Monte Carlo simulation

Several methods are compared in the simulation. Furthermore, the following reference variances

$$V(\hat{\vartheta}) = \sum_{c=1}^{10,000}\frac{(\hat{\vartheta}_c - \theta)^2}{10,000}, \tag{19}$$

hereinafter denoted as empirical or Monte Carlo variances, are computed.

For the HT estimator are considered the:

- unbiased variance estimator (Wolter 2007) denoted as STANDARD method;
- EDAGJK according to Kott (Kott 2001);
- EDAGJK.I: the modified EDAGJK (section 2.2 using the replicate weights given in the (3);
- BOOTSTRAP.I: bootstrap variance methods under imputation (Shao and Sitter 1996);
- MI: using the Approximate Bayesian Bootstrap (ABB) (Kim et al. 2004, Brick et al. 2005).

For the calibration estimator have been compared the:

- TAYLOR variance estimator;
- EDAGJK.HT computed according to (4);
- EDAGJK.CAL computed according to (5);
- EDAGJK.HT.I: the modified EDAGJK (section 2.2) based on the correction factor (4);
- EDAGJK.CAL.I: the modified EDAGJK (section 2.2) based on the correction factor (5);
- BOOTSTRAP.I: bootstrap variance methods under imputation (Shao and Sitter 1996);
- MI: using the ABB.

Note that the STANDARD, EDAGJK, TAYLOR, EDAGJK.HT and EDAGJK.CAL, do not properly take into account the imputation correction.

The complete simulation has implemented twelve variance methods by four variables. The imputation procedure is performed and for seven variance estimators imputation adjustement is carried out as well. Then, computational issues led us to choose 1,000 replications in performing the variance estimates for each method.

The accuracy of the variance estimates is measured with following summary statistics:

- The Relative (percentage) Bias of Variance estimation

$$RB[v(\hat{\vartheta})]\% = 100 \times \frac{\overline{v}(\hat{\vartheta}) - V(\hat{\vartheta})}{V(\hat{\vartheta})}. \tag{20}$$

- The Relative (percentage) Root Mean Square Error of Variance estimation

$$RRMSE[v(\hat{\vartheta})]\% = 100 \times \sqrt{\frac{\frac{1}{1{,}000}\sum_{c=1}^{1{,}000}\left[v(\hat{\vartheta}_c) - V(\hat{\vartheta})\right]^2}{V(\hat{\vartheta})^2}}. \tag{21}$$

- The Coverage of the Confidence Interval (percentage), that is the percentage of intervals including θ, based on the nominal 95 % confidence intervals computed for each of 1,000 simulations. We used the normal distribution as approximation of the t distribution
  $CCI[v(\hat{\vartheta})]\% = \frac{100}{1{,}000}\sum_{c=1}^{1{,}000}\delta_c$ where $\delta_c = 1$ if $\theta \in \left(\hat{\vartheta}_c \pm 1.96\sqrt{v(\hat{\vartheta}_c)}\right)$ and $\delta_c = 0$ otherwise.
- The Lower Error Rate and Upper Error Rate
  $LER[v(\hat{\vartheta})]\% = 100 \times \frac{1}{1{,}000}$(number of samples with $\quad \theta < -1.96\sqrt{v(\hat{\vartheta}_c)} \quad$),
  $UER[v(\hat{\vartheta})]\% = 100 \times \frac{1}{1{,}000}$(number of samples with $\quad \theta > +1.96\sqrt{v(\hat{\vartheta}_c)} \quad$).

Table 4 shows that for the variables with large nonresponse rate (CIF and OEX), the methods that do not take properly into account the imputation process such as STANDARD, EDAGJK, TAYLOR, EDAGJK.HT and EDAGJK.CAL produce large downward biased variance estimates. The result was definitely expected.

Furthermore for the OEX variable we observe a very large variability with the $RRMSE[v(\hat{\vartheta})]\%$ over than 177% for all the methods. The evidence is explained by the positive skew distribution of this variable (figure 2).

The scatterplot of the 10,000 variance estimates versus the corresponding HT estimates (figure 3) shows for the OEX variable two separate clouds, being the highest one around of size 200. This is due to one extreme value within the stratum NORTH-WEST:(size class<8) with the 0.02 sampling rate. For this stratum the expected percentage contribution to the overall variance is around 85%. Furthermore, the stratum variance when the extreme value is included is about 5 times the stratum variance when the extreme value is not included in the sample.

The presence of rare extreme values is typical in the business surveys and then it is interesting to study the behaviour of the estimators in this critical context.

In the following the main comments are focused on CIF variable, because it has the highest missing rate. As concerns the HT estimator EDAGJK.I and bootstrap methods, they produce $RB[v(\hat{\vartheta})]\%$ around 7% but bootstrap has a smaller $RRMSE[v(\hat{\vartheta})]\%$ than EDAGJK.I. MI has the smallest $RB[v(\hat{\vartheta})]\%$ with the drawback to be negative.

Table 4 - Relative Bias ($RB[v(\hat{\vartheta})]$) and Relative Root Means Square Error ($RRMSE[v(\hat{\vartheta})]$) of the variance estimators with imputed data

	H-T estimator							
	$RB[v(\hat{\vartheta})]\%$				$RRMSE[v(\hat{\vartheta})]\%$			
Variance Estimator	CIF	PUC	OEX	LCO	CIF	PUC	OEX	LCO
STANDARD	-27.15	-3.37	-8.96	-6.40	53.43	38.54	204.64	15.41
EDAGJK	-21.26	3.11	-8.11	-0.30	57.94	45.90	178.76	31.26
EDAGJK.I	7.43	7.44	6.77	3.99	71.35	47.65	209.31	32.61
BOOTSTRAP.I	7.95	5.57	21.40	2.81	64.53	42.99	261.52	21.90
MI	-2.83	2.36	-1.82	-3.08	66.74	40.96	205.00	15.58
	Calibration estimator							
	$RB[v(\hat{\vartheta})]\%$				$RRMSE[v(\hat{\vartheta})]\%$			
TAYLOR	-28.48	-6.21	-9.82	-7.55	53.90	43.93	227.47	23.82
EDAGJK.HT	-19.76	3.60	-5.73	2.95	59.27	51.12	195.25	36.73
EDAGJK.CAL	-18.24	6.79	-6.56	4.98	62.15	55.04	186.05	40.89
EDAGJK.HT.I	9.48	8.01	9.15	6.61	73.30	53.02	225.78	38.60
EDAGJK.CAL.I	11.51	11.31	8.23	8.75	77.73	56.85	215.29	42.76
BOOTSTRAP.I	5.71	6.57	14.43	4.79	64.38	48.84	244.40	27.39
MI	-9.05	-4.46	-15.34	0.17	57.24	42.46	177.88	22.42

In case of calibration estimator, bootstrap outperforms the methods that consider specifically the imputation in terms of $RB[v(\hat{\vartheta})]\%$. Nevertheless the modified EDAGJK methods produce positive and not so large $RB[v(\hat{\vartheta})]\%$. MI has negative bias, but outperfoms the other estimators in terms of $RRMSE[v(\hat{\vartheta})]\%$. The table 4 shows that EDAGJK.HT is slightly better than EDAGJK.CAL at least for the CIF variable.

Table 5 shows the coverage of the confidance interval. The methods ignoring that many values are imputed have a strong reduction of the coverage rates. MI does not show good performaces, at least for the CIF variable, while a small decreasing of coverage is observed for the rest of the resampling methods. The modified EDAGJK techniques and bootstrap are essentially equivalent. For the calibration estimator, EDAGJK.CAL.I seems slightly better than EDAGJK.HT.I and bootstrap. That occurs because of a larger $RB[v(\hat{\vartheta})]\%$.

Table 5 - Coverage of the Confidence Interval($CCI[v(\hat{\vartheta})]$), the Lower and Upper Error Rate ($LER[v(\hat{\vartheta})]$, $UER[v(\hat{\vartheta})]$) with imputed data

	H-T estimator											
	$CCI[v(\hat{\vartheta})]\%$				$LER[v(\hat{\vartheta})]\%$				$UER[v(\hat{\vartheta})]\%$			
Variance Estimator	CIF	PUC	OEX	LCO	CIF	PUC	OEX	LCO	CIF	PUC	OEX	LCO
STANDARD	88.40	94.50	90.40	94.40	7.50	1.20	1.00	1.40	4.10	4.30	8.60	4.20
EDAGJK	86.20	93.00	88.20	94.00	6.80	4.70	7.70	3.40	7.00	2.30	4.10	2.60
EDAGJK.I	91.10	93.70	89.90	94.40	4.20	4.20	6.90	3.10	4.70	2.10	3.20	2.50
BOOTSTRAP.I	92.30	93.80	90.30	95.90	2.90	4.20	7.10	2.30	4.80	2.00	2.60	1.80
MI	89.90	94.40	90.60	95.50	4.50	3.90	7.20	2.50	5.60	1.70	2.20	2.00
	Calibration estimator											
	$CCI[v(\hat{\vartheta})]\%$				$LER[v(\hat{\vartheta})]\%$				$UER[v(\hat{\vartheta})]\%$			
TAYLOR	88.20	94.20	90.20	93.40	7.70	1.80	0.90	2.00	4.10	4.00	8.90	4.60
EDAGJK.HT	86.20	92.70	89.40	93.60	6.50	5.30	7.00	3.00	7.30	2.00	3.60	3.40
EDAGJK.CAL	86.76	93.88	89.87	92.98	6.32	4.61	7.12	3.51	6.92	1.50	3.01	3.51
EDAGJK.HT.I	91.80	92.90	91.20	93.70	3.50	5.20	6.30	3.00	4.70	1.90	2.50	3.30
EDAGJK.CAL.I	92.38	94.18	91.57	93.68	3.71	4.41	5.82	3.21	3.91	1.40	2.61	3.11
BOOTSTRAP.I	92.00	93.72	90.08	95.65	3.24	4.66	6.88	1.92	4.76	1.62	3.04	2.43
MI	90.50	92.20	88.20	93.50	4.00	6.00	8.40	3.30	5.50	1.80	3.40	3.20

Figure 2 - Distribution of the 10,000 HT estimates

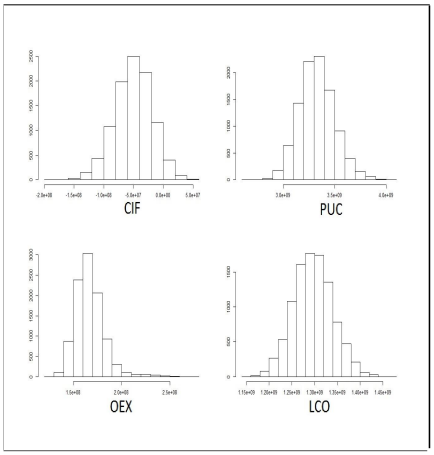
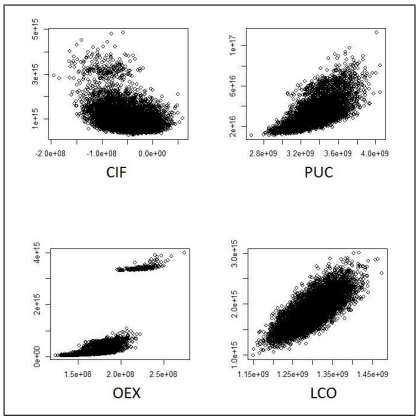


Figure 3 - Scatterplot of the HT estimates versus standard variance estimates



4. Conclusions

Many statistical surveys carried out by National Statistical Institutes are, generally, defined by a large sample drawn according to a complex sampling design. Unit and item non responses increase the trouble to make inference.

The paper investigates three variance estimators taking into account the item non responses when random hot deck imputation has been performed. Two of the three methods are the standard bootstrap and the MI, while the third one is a new variance estimator. The proposed method combines the EDAGJK technique proposed by Kott with the adjusted jackknife proposed by Rao and Shao. The reasons leading to new estimator is the good compromise among theoretical properties and practical aspects. In particular EDAGJK produces an unbiased estimator (for complete data set), it is easy to implement and not computer intensive. Furthermore, the adjusted jackknife does not require replications of the imputation procedure.

These features are quite appealing especially in a National Statistical Institute (NSI), where data production based on large data sets must be automatized as much as possible.

The three methods have been compared by means of a Monte Carlo simulation based on real business data and a sampling strategy resembling to the Small and Medium Entrerpise survey conducted by the Italian Statistical Institute. The simulation results show that the modified EDAGJK[4] with Rao and Shao adjustment produces nearly unbiased variance estimates and it works well with respect to the two benchmarking methods in terms of accuracy and coverage of coinfidance interval. Nevertheless the method is less computational demanding than bootstrap and it does not require an increasing of complexity of the data production process as for MI.

The paper show that for variable with an hight level of imputation the standard methods of variance estimation deeply under-estimates the true variance, a best practice for a NSI should be to consider the level of item non response for each variables and to performance

The empirical results shows that for variables with an high level of imputation rate, the standard methods of variance estimation deeply under-estimate the true variance. Then a

[4] An R function implementing the modified EDAGJK is available in the Deliverable 6.1 of the BLUE-Enterprise and Trade Statistics project (Blue-ETS 2013).

best practice for NSIs should be to make a screening, for the main variables of interest of each business survey, of the item non response rates and to adopt valid variance estimation methods for the variables affected by the highest (eg. >10%) item non response rates.

The information about the item non-response rates shuld be also disseminated to the external users. In fact, if the research institute releases a standard file with imputation flag variable and the replicate weights, every users can compute the variance estimates for every kind of unplanned domains of interest by a simple formula.

References

Blue-ETS (2013): Deliverable 6.1. Best practice recommendations on variance estimation and small area estimation in business surveys. Computer Codes. URL http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable6.1.pdf

Brick, J. M., Jones, M. E., Kalton, G. and Valliant, R. (2005): Variance Estimation with Hot Deck Imputation: A simulation Study of Three Methods. *Survey Methodology*, 31, pp. 151-159.

Burns, R. M. (1990): Multiple and Replicate Item Imputation in a Complex Sample Survey. *Proceedings Sixth Annual Res. Conf.*, pp. 655-665, Washington, DC: U.S. Bureau of the Census.

Chen, J. and Shao, J. (2001): Jackknife Variance Estimation for Nearest-Neighbour Imputation. *Journal of the American Statistical Association*, 95, pp. 260-269.

Di Zio, M., Falorsi, S., Guarnera, U., Luzi, O. and Righi, P. (2008): Variance Estimation in Presence of Imputation: an Application to an Istat Survey Data. *Proceedings of the Sec. on Survey Res. Meth. European Conference on Quality in Official Statistics*. URL http://q2008.istat.it/sessions/paper/17DiZio.pdf

Efron, B. (1994): Missing Data, Imputation and the Bootstrap. *Journal of the American Statistical Association*, 89, pp. 463 - 479.

Kalton, G. and Kasprzyk, D. (1986): The Treatment of Missing Survey Data. *Survey Methodology*, 12, pp. 1-16.

Kim, J. K., Brick, J. M., Fuller, W. A. and Kalton, G. (2006): On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling. *J. R. Statist. Soc. B*, 68, pp.509-521.

Kim, J. K. and Fuller, W. A. (2004): Fractional Hot Deck Imputation. *Biometrika*, 91, pp. 559-578.

Kott, P. (1998): Using the Delete-a-Group Jackknife Variance Estimator in NASS Surveys. *Nass research report 98-01* (revised 2001), NASS. URL http://www.nass.usda.gov/research/reports/RRGJ7.pdf

Kott, P. (2001): Delete-a-Group Jackknife. *Journal of Official Statistics*, 17, pp. 521-526.

Kott, P. (2006): Delete-a-Group Variance Estimation for the General Regression Estimator Under Poisson Sampling. *Journal of Official Statistics*, 22, pp. 759-67.

Little, R. J. A. and Rubin, D. B. (2002): *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics, Wiley.

Miller, D. and Kott, P. (2011): Using the DAG Jackknife to Measure the Variance of an Estimator in the Presence of Item Nonresponse. *JSM Proceedings, Statistical Computing Section*. Alexandria, VA: Am. Statist. Assoc.

Rao, J. N. K. (1996): On Variance Estimation with Imputed Survey Data. *Journal of the American Statistical Association*, 91, pp. 499-506.

Rao, J. N. K. and Shao, S. (1992): Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation. *Biometrika*, 79, pp. 811-822.

Rao, J. N. K. and Shao, S. (1999): Modified Balanced Repeated Replication for Complex Survey Data. *Biometrika*, 86, pp. 403-415.

Rao, J. N. K. and Sitter, R. R. (1995): Variance Estimation under Two-Phase Sampling with application to Imputation for Missing Data. *Biometrika*, 82, pp. 453-460.

Rao, J. N. K. and Wu, C. F. J. (1988): Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association*, 83, pp. 231-241.

Raghunathan, T. E., Lepkowski, J.M., Van Hoewyk, J. and Solenberger, P.: A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, June 2001 27, pp. 8595

Rubin, D. B. (1978): Multiple Imputation in Sample Surveys. *Proceedings Survey Res. Meth. Sec.*, Am. Statist. Assoc., pp. 20-34.

Rubin, D. B. (1987): *Multiple Imputation for Nonresponse in Surveys*. Wiley Series in Probability and Statistics, Wiley.

Rubin, D. B. and Schenker, N. (1986): Multiple Imputation for Interval Estimation for Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81, pp. 366-374.

Rust, K. (1985): Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, 1, pp. 381-397.

Rust, K. (1986): Efficient Formation of Replicates for Replicated Variance Estimation. *Proceedings Survey Res. Meth. Sec.*, Am. Statist. Assoc., pp. 81-87.

Rust, K. and Rao, J. N. K. (1996): Variance Estimation for Complex Surveys Using Replication Techniques. *Statistical Methods in Medical Research*, 5, pp. 283-310.

Saigo, H., Shao, J. and Sitter, R. R. (2001): A Repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data. *Survey Methodology*, 27, pp. 189-196.

Saigo, H. and Sitter, R. R. (2005): Jackknife Variance Estimator with Reimputation for Randomly Imputed Survey Data. *Statistics and Probability Letters*, 73, pp. 321-331.

Schafer, J. L. (1998): Multiple Imputation: a Primer. *Statistical Methods in Medical Research*,8, pp. 3-15.

Shao, J. (2003): Impact of the Bootstrap on Sample Surveys. *Statistical Science*, 18, pp. 191-198.

Shao, J. and Sitter, R. R. (1996): Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association*, 91, pp. 1278-1288.

Shao, J. and Tang, Q. (2011): Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation. *Journal of Official Statistics*, 27, pp. 507-526.

Shao, J. and Tu, D. (1995): *The Jacknife and Bootstrap*. Springer-Verlag GmbH.

Skinner, C. J. and Rao, J. N. K. (2002): Jackknife Variance Estimation for Multivariate Statistics under Hot-Deck Imputation from Common Donor. *Journal of Statistical Planning and Inference*, 102, pp. 149-167.

Särndal, C. E. (1992): Methods for Estimating the precision of Survey Estimates when Imputation has been used. *Survey Methodology*, 18, pp. 241-252.

Valliant, R., Brick, M. J. and Dever, J. (2008): Weight Adjustments for the Grouped Jackknife Variance Estimator. *Journal of Official Statistics*, 24, pp. 469-488.

Wolter, K. (2007): *Introduction to Variance Estimation*. Springer London, Limited.

Yung, W. and Rao, J. N. K. (2000): Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information. *Journal of the American Statistical Association*, 95, pp. 903-915.

L'effetto delle modificazioni longitudinali delle imprese sugli indicatori dell'indagine "Occupazione, orari di lavoro, retribuzioni e costo del lavoro nelle grandi imprese"[1]

Fabiana Rocci[2], *Laura Serbassi*[3]

Sommario

Le modificazioni longitudinali delle imprese dovute a processi di concentrazione o frammentazione, nascite e cessazioni sono molto comuni. La loro sistematica registrazione è alla base di una corretta stima dei fenomeni economici. Nelle indagini congiunturali, che devono seguire quasi in tempo reale tali eventi, essi possono rappresentare una fonte di errore non campionario che causa la distorsione delle stime negli indici di variazione. Il presente lavoro propone l'indagine mensile "Occupazione, orari di lavoro, retribuzioni e costo del lavoro nelle grandi imprese" come caso studio per un'analisi dell'efficienza delle scelte relative al disegno di indagine definito per neutralizzare tali effetti. Una simulazione su dati di indagine ha permesso di studiare in modo comparativo scelte metodologiche alternative, al fine di valutare la rappresentatività del panel di imprese e misurare gli effetti delle trasformazioni giuridiche sulle dinamiche longitudinali degli indicatori prodotti.

Parole chiave: grandi imprese, modificazioni longitudinali, panel, numeri indice

Abstract

Longitudinal changes in enterprises as concentration or fragmentation, births and deaths, are very common. These are all events causing the evolution of the business population, hence it is essential to register them properly to achieve correct estimates of economic macro-data. This is a particularly crucial issue in a short term survey context, because of the trade-off between timeliness and statistical information needed in order to register these changes properly. This paper proposes the monthly survey "Employment, worked hours, wages and labor cost in large enterprises" as a case study, making an analysis of the efficiency of the design of the survey defined in order to neutralise these effects. Alternative methodological choices of the design have been analysed through a simulation on survey data, in order to assess the representativeness of the panel and measure the effects of legal changes on the longitudinal dynamics of the indices produced.

Keywords: large enterprises, longitudinal changes, panel, index numbers

[1] Gli autori ringraziano Roberto Gismondi e Giuseppe Amato per i loro suggerimenti e il tempo dedicato. Sebbene il lavoro sia frutto dell'opera di entrambi gli autori i paragrafi 1.2, 3.3, 3.4 e 4.1 possono essere attribuiti a Fabiana Rocci e i paragrafi 1.1, 2, 3.1, 3.2 e 4.2 a Laura Serbassi. L'introduzione e le conclusioni sono a cura di entrambi gli autori.

[2] Ricercatore (Istat), e-mail: rocci@istat.it.

[3] Ricercatore (Istat), e-mail: laserbas@istat.it.

Introduzione

Un'indagine statistica ha generalmente l'obiettivo di misurare una variabile relativa a una popolazione oggetto di studio. L'interesse può essere diretto al livello che questa variabile presenta in un preciso momento o alla variazione che questa subisce in periodi successivi, sempre in relazione alla popolazione su cui lo studio è incentrato. In quest'ultimo caso, è importante sottolineare che la stima corretta della variazione è anche connessa alla conoscenza degli eventi demografici a cui la popolazione nel frattempo è stata soggetta. Le variazioni rilevate, infatti, possono essere dovute sia al cambiamenti della variabile oggetto di stima, sia alla differente struttura della popolazione in termini di numerosità delle unità o di composizione nei due momenti osservati.

Nell'ambito delle indagini sulle imprese oltre agli eventi di cessazione e di nascita di nuove imprese, sono presenti anche altri legati a una vasta gamma di possibili modificazioni longitudinali di tipo economico-strutturale e/o amministrativo. Tali aspetti, separabili da un punto di vista concettuale e formale, risultano profondamente interconnessi e difficilmente scindibili nella realtà operativa, tali da rendere l'osservazione dell'esatta popolazione oggetto di studio problematica.

Una categoria di unità statistiche particolarmente soggetta agli eventi di trasformazione è rappresentata dalle imprese di grandi dimensioni, per cui è comune osservare episodi di riorganizzazione e ristrutturazione aziendale posti in essere al fine di migliorare la *perfomance* economica. Tali modificazioni, pur mantenendo in vita l'unità, possono produrre effetti sulle sue caratteristiche economiche e quindi sui valori caratteristici della variabile sulla popolazione statistica a cui appartiene (quali per esempio la retribuzione percepita o le ore lavorate dai dipendenti).

Tale aspetto assume particolare rilievo nel caso in cui l'oggetto di studio si basa su dati longitudinali di impresa, per cui la stessa unità di rilevazione viene rilevata in periodi successivi. Un caso particolare è rappresentato dalle indagini congiunturali sulle imprese, che registrano le variazioni delle variabili economiche su base infrannuale e spesso basano la rilevazione su un panel di imprese. Tali indagini, per loro natura, devono affrontare il problema di seguire le trasformazioni e la demografia della popolazione in tempo quasi reale, ossia prima che esse siano registrate in modo definitivo e ufficiale nel registro delle imprese.

Nell'ambito della statistica ufficiale sono diverse le regole indicate per il trattamento e la registrazione delle modifiche longitudinali di impresa, in particolare per la costruzione del registro delle imprese attive (OECD 2007, EUROSTAT 2010). La tematica, invece, relativa alle modalità di registrazione nelle indagini che precedono il rilascio dei registri deve essere approfondita, in particolare per gli effetti che questi possono avere sugli indicatori di variazione congiunturali (EUROSTAT 2006), dove la conoscenza parziale e non definitiva degli eventi può causare un effetto distorsivo sugli indicatori prodotti.

Questo lavoro ha l'obiettivo di fornire una misura empirica degli effetti delle modifiche longitudinali delle grandi imprese sugli indici congiunturali dell'indagine mensile "Occupazione, orari di lavoro, retribuzioni e costo del lavoro nelle grandi imprese" (nel seguito GI). L'indagine si basa su un panel censuario delle imprese con almeno 500 dipendenti nella media dell'anno base. Il panel è chiuso, ovvero il suo aggiornamento viene effettuato solo al momento del passaggio alla nuova base. Il disegno di indagine, relativo alle regole di registrazione delle trasformazioni longitudinali e alla metodologia di calcolo degli indicatori, è stato definito coerentemente con l'adozione del panel chiuso.

L'obiettivo di questo studio è analizzare i diversi aspetti del trattamento delle modificazioni longitudinali, attraverso l'esperienza sia operativa che metodologica di questa indagine. A tal fine, le caratteristiche descritte vengono valutate attraverso una simulazione su dati di indagine, che ha permesso di misurare gli effetti di scelte metodologiche diverse.

La simulazione si basa su dati reali (rilevati dall'indagine, biennio 2003-2004) e consiste nella rielaborazione di alcuni indicatori sulla base di un panel aperto, quindi aggiornato annualmente tenendo conto di tutti gli eventi demografici intercorsi, di natimortalità, dei passaggi di soglia dimensionale e di trasformazione giuridica registrati nelle grandi imprese. I risultati finali sono stati messi a confronto con gli indicatori pubblicati dall'indagine per poter misurare le differenze negli indici dovute al diverso trattamento delle modificazioni longitudinali. I risultati dalla simulazione vengono illustrati nell'ultimo paragrafo con particolare considerazione per le dinamiche longitudinali degli indicatori prodotti sulla base dei due diversi panel.

1. Aspetti teorici: indici di variazione, natimortalità e modifiche longitudinali

Gli eventi di trasformazione longitudinale che possono modificare una popolazione di unità statistiche. Questa risulta così differente in momenti diversi, sia per la sua numerosità, sia per le caratteristiche delle unità compresenti. In generale, gli effetti delle varie modifiche possono essere ascritti a due tipologie:

- variazione nel numero di unità esistenti nella popolazione; essenzialmente si tratta di nascite, morti, sospensioni di attività, cambiamenti di strato dovuti a modifiche in uno o più caratteri dell'impresa;
- la discontinuità nel profilo delle singole unità di rilevazione in termini di confrontabilità temporale, causata essenzialmente da fusioni e scorpori, ove le unità seppur formalmente uguali a se stesse hanno in realtà diversa natura economica.

Si pone così la domanda se e quali siano gli effetti di tali modifiche sulle stime finali del parametro di interesse. Il primo caso ha un effetto prettamente numerico sugli indicatori finali, che è possibile isolare e valutare attentamente. Il secondo caso, invece, è legato alle caratteristiche economiche di un'impresa, che non sono sempre direttamente osservabili e quindi possono rappresentare un elemento di difficile analisi. Tale difficoltà è ancor più grande in presenza di fenomeni complessi, come quelli di riorganizzazione di una grande impresa, generando così un possibile errore non campionario per la rilevazione. Quindi, è importante considerare come gli eventi di modificazione longitudinale possono introdurre discontinuità nel profilo economico dell'unità, che può rappresentare un problema nella rilevazione dell'unità stessa nel caso di stime basate sui dati longitudinali di impresa. Un caso tipico è rappresentato dalla indagini congiunturali, per cui le modalità di rilevazione e di registrazione di tali eventi assumono un'importanza fondamentale. Nei paragrafi seguenti si presenta una possibile classificazione degli eventi di trasformazione societaria, in funzione del loro effetto sulla continuità temporale del profilo di un'unità statistica, nonché un'analisi di come tali eventi possono avere un affetto sugli indici a base fissa.

1.1 Le modificazioni longitudinali delle imprese

Precise convenzioni di trattamento della registrazione dei cambiamenti demografici esistono al fine di definire i registri statistici sulle imprese (EUROSTAT, 2010), che consentono alle indagini di relazionarsi ad essi in modo coerente rispetto ai propri obiettivi conoscitivi. Per quanto riguarda le indagini congiunturali, esse spesso si trovano a gestire tali eventi con notevole anticipo rispetto ai registri ufficiali spesso sulla base di informazioni parziali e frammentate, acquisite direttamente in fase di rilevazione e che devono essere elaborate in tempi molto brevi. In questi casi, è particolarmente importante esplicitare i criteri utilizzati per il trattamento statistico delle modificazioni, da cui dipendono sia le modalità di identificazione delle unità nella nuova popolazione così formatasi, sia la confrontabilità tra due periodi delle unità coinvolte negli eventi.

La possibile casistica degli eventi può essere schematizzata nel seguente modo:

- modifiche anagrafiche: si tratta sostanzialmente di modifiche legate a variabili "accessorie" e non prettamente strumentali per lo svolgimento della propria attività, come ad esempio la variazione nella ragione sociale, nel logo societario o della sede;
- modifiche di esistenza: assumono una certa importanza, oltre che per l'evidente effetto demografico che comporta un aumento od un decremento della dimensione del dominio di studio tra due tempi, anche per l'elasticità implicita nel concetto di "cessazione dell'attività" (imprese stagionali, chiusura per ristrutturazione, ecc.);
- modifiche per trasformazione: una trasformazione è un processo che permette a un'impresa di acquisire (o cedere) tutta o parte della (propria) attività di (a) un'altra impresa. Si tratta della casistica più complessa a causa dell'estrema varietà di condizioni che possono determinare la specificità della trasformazione. In generale, si tratta di eventi che nella pressoché totalità dei casi comportano modifiche nei caratteri (attività prevalente e addetti in primo luogo) e/o di esistenza (cessazione di imprese incorporate e nascite di imprese nate dalla fusione di preesistenti).

Uno dei criteri teorici usualmente seguito per classificare le modifiche longitudinali è in funzione del loro effetto sulla omogeneità temporale delle caratteristiche di un'unità statistica (Struijs e Willeboordse, 1995; OECD, 2007).

In linea generale, come descritto nel prospetto 1.1, esso conduce alla definizione di classi di modificazioni longitudinali mutuamente esclusive e che coprono tutte le casistiche correnti. Per quanto riguarda le modifiche anagrafiche, si suppone che nella grande maggioranza dei casi venga conservata l'identità dell'unità originaria. I cambiamenti nell'attività economica possono ricadere nella casistica delle modifiche di esistenza almeno dal punto di vista statistico, nel senso che in sede d'indagine l'unità coinvolta dal cambiamento di attività sarà assimilabile a una nuova nata nel dominio di destinazione (nuova attività) e a un'unità morta nel dominio di origine (vecchia attività). Un caso analogo è rappresentato nelle modifiche di esistenza dalle 'cessazioni e nascite con continuità', ossia una situazione di modifica uno a uno con perdita di identità di un'unità originaria che, in un certo qual modo, continua nell'unità nuova. Le modifiche per trasformazione coinvolgono sempre più di un'unità e possono essere di tre tipi: concentrazione (due o più unità si modificano in un'unica unità), frammentazione (un'unità si modifica in due o più unità), ristrutturazione (due o più unità si modificano in due o più unità).

Prospetto 1.1 - Classificazione sintetica delle modificazioni longitudinali delle imprese

TIPO DI MODIFICAZIONE	Numero di unità coinvolte	Continuità dell'identità
1) Modifica nei caratteri (e anagrafiche)	1:1	
2) Modifiche di esistenza		
1. Nascita	0:1	No
2. Cessazione (morte)	1:0	No
3. Cessazione e nascita con continuità	1:1	No
3) Modifiche per trasformazione		
1. Concentrazione		
1. Fusione	x:1	No
2. Fusione per incorporazione	x:1	Si (a)
2. Frammentazione		
1. Scioglimento	1:y	No
2. Scorporo	1:y	Si (a)
3. Ristrutturazione	x:y	Si o No

Fonte: Struijs e Willeboordse 1995.
(a) In questo caso c'è discontinuità nei dati economici.

Nella classificazione proposta, gli eventi di concentrazione e frammentazione sono divisi in due categorie per esplicitare i diversi effetti sulla continuità d'identità dal punto di vista statistico. L'unità che deriva da un fenomeno di concentrazione potrebbe essere o non essere la medesima rispetto a una delle unità precedenti all'evento. Unità che si fondono perdono la propria identità (fusione), mentre nel caso in cui un'unità incorpora altre unità generalmente l'unità più grande conserva la propria identità (fusione per incorporazione). Di contro, nei fenomeni di frammentazione un'unità o si smembra generando altre unità, nessuna delle quali conserva una continuità rispetto all'unità iniziale (scioglimento), oppure alcune unità si scorporano dall'unità originaria, che però rimane in vita e generalmente conserva la propria identità (scorporo). Nella classe delle ristrutturazioni finiscono le modificazioni non altrimenti classificabili che generalmente risultano le più complesse e le meno standardizzabili.

1.2 Effetti della natimortalità delle imprese sugli indici a base fissa

In termini teorici gli indicatori tipici per misurare la variazione di una variabile tra periodi successivi sono i numeri indice. I più diffusi sono quelli congiunturali che hanno l'obiettivo di rilevare, su intervalli infrannuali, l'andamento di variabili utili per l'analisi congiunturale del ciclo economico, importante nei processi decisionali del settore pubblico (governi nazionali, banche centrali ecc.) e di quello privato (imprese, mercati finanziari ecc.). In termini molto generali l'indice più semplice è quello a base fissa I_T che per ogni tempo T misura la variazione della data variabile Y rispetto al valore della stessa variabile relativo a un tempo considerato come punto di riferimento (base):

$$I_T = \frac{Y_T}{Y_B} \qquad T=1,2,\ldots n \qquad (1)$$

dove:
Y_B = ammontare complessivo della variabile Y delle unità attive al tempo base B;
Y_T = ammontare complessivo della variabile Y delle unità attive al tempo T.

È possibile evidenziare l'effetto della demografia di impresa sugli indici esprimendo la formula (1) in funzione dei valori assunti dai tre gruppi in cui l'insieme totale delle imprese può essere suddiviso: quelle presenti in entrambi i periodi su cui si rileva la variazione, quelle nate e quelle cessate nell'intervallo. Indicando con N_T il numero delle unità statistiche attive nella popolazione al tempo T, con N_{PT} l'insieme delle imprese compresenti nei due periodi della base B e il generico periodo T, con N_{GT} quello delle imprese nate nell'intervallo tra la base B e T e con N_{MT} quello delle imprese cessate durante l'intervallo, si avrà:

$N_T = N_{PT} + N_{GT} - N_{MT}$

Analogamente è possibile scomporre il valore complessivo dell'ammontare della variabile Y secondo la stessa logica nei due tempi B della base e il generico tempo T.

$\overline{Y}_B$: media della variabile Y nel periodo base B calcolato su tutte le imprese appartenenti al campo di osservazione.

Y_{PT} : ammontare complessivo della variabile Y al tempo T delle unità statistiche compresenti (P) nella popolazione sia alla base B che al tempo T;

Y_{GT} : ammontare complessivo della variabile Y al tempo T delle unità statistiche nate (G) dopo l'introduzione della base e attive al solo tempo T;

$\overline{Y}_{PB}$: media della variabile Y nel periodo base B calcolato sulle sole imprese compresenti (P) sia alla base che al tempo T;

$\overline{Y}_{MT}$: media della variabile Y sulle sole imprese presenti alla base B e cessate (M) al tempo T (valore nullo al tempo T).

Con semplici passaggi algebrici l'indice I_T può essere espresso nella forma seguente (Gismondi, 2005):

$$I_T = \frac{[I_{PT}] N_{PT} \overline{Y}_{PB} + [d_{GMT} \Lambda Y_{GMT}] N_{MT} \overline{Y}_{MT}}{N_B \overline{Y}_B} \qquad (2)$$

dove

$$I_{PT} = \frac{Y_{PT}}{\overline{Y}_{PB}} \quad \text{e} \quad \Lambda Y_{GMT} = \frac{\overline{Y}_{GT}}{\overline{Y}_{MT}} \quad \text{e} \quad d_{GMT} = \frac{N_{GT}}{N_{MT}} \qquad (3)$$

ossia in ogni tempo T l'indice di variazione della variabile Y sulla popolazione di riferimento si può esprimere come una media aritmetica ponderata di due diversi elementi. Il primo addendo è un indice semplice rispetto all'anno base I_{PT} della variabile Y sulle imprese compresenti nei due tempi B e T, ovvero delle imprese panel. Il secondo addendo invece registra l'influenza sull'indice della natimortalità delle imprese, sia in termini della loro numerosità che in termini dei valori ad esse legati. Tale effetto registra il tasso di variazione della numerosità tra nate e cessate e tra i valori della variabile Y legati rispettivamente a questi due insiemi di imprese. I pesi della ponderazione, invece, sono dati dal rapporto tra i livelli della variabile in questione registrati sui due insiemi di imprese, rispettivamente compresenti ed escluse, rispetto al livello registrato sul totale delle imprese al tempo base B. In questo modo si evidenzia l'impatto delle diverse numerosità tra imprese nuove e cessate e tra i loro valori, che possano avere dei forti effetti all'indice di variazione finale.

In tale contesto si inquadra la problematica di un'indagine congiunturale che, considerato il breve lasso di tempo tra l'evento e la produzione dei dati, potrebbe avere informazioni asimmetriche su tali cambiamenti. Questo aspetto può dare origine a una forma di errore non campionario che può causare distorsione nell'indicatore prodotto. Di conseguenza, nell'ambito delle indagini congiunturali, riveste un ruolo importante la definizione delle regole di registrazione e trattamento degli eventi di natimortalità, che devono essere coerenti con la metodologia adottata per il calcolo degli indici di variazione, al fine di evitare una forte distorsione degli indicatori e garantirne la coerenza e la confrontabilità.

2. La rilevazione sulle Grandi Imprese[4]

L'indagine mensile GI è una rilevazione finalizzata al calcolo e alla diffusione di indicatori economici congiunturali nelle imprese con almeno 500 dipendenti appartenenti al settore privato non agricolo, ad esclusione dei servizi sociali e personali (sezioni B-N della classificazione Ateco 2007). In accordo con le definizioni e le metodologie prevalenti a livello internazionale, l'unità di rilevazione dell'indagine è l'impresa, mentre l'unità di analisi è l'unità funzionale (kau). Gli indicatori diffusi sono numerosi e riguardano sia variabili sull'input di lavoro (occupazione e ore lavorate), sia variabili retributive e di costo del lavoro rappresentate in termini di valori pro capite e orari. I numeri indice prodotti sono a base fissa che viene aggiornata ogni cinque anni, come previsto dai regolamenti comunitari per le indagini congiunturali (reg. STS CE n. 1165/98 e successive modifiche). In termini schematici, il disegno metodologico dell'indagine è il seguente:

- l'indagine si basa su una rilevazione di tipo panel, che è definito in modo censuario da tutte le imprese presenti nell'archivio Asia (Archivio statistico delle imprese attive) che nella media dell'anno base hanno almeno 500 dipendenti;
- la stima delle variabili non contempla il riporto all'universo, ma avviene semplicemente per somma dei valori osservati sulle unità del panel, in quanto questo è censuario;
- è possibile distinguere gli indici in due diverse tipologie sulla base del metodo di calcolo utilizzato: gli indici concatenati dell'occupazione e gli indici pro capite e orari calcolati per le variabili ore lavorate, retribuzione lorda e costo del lavoro che registrano direttamente la variazione rispetto al valore medio dell'anno base (descritti nel prosieguo);
- il panel utilizzato è "chiuso", ossia non tiene conto della nati-mortalità d'impresa al di fuori di esso rispetto alla soglia dimensionale. Tutte le imprese presenti nel panel sono tenute a compilare il questionario mensile fino al rinnovo della base, anche se la loro dimensione occupazionale dovesse scendere sotto la soglia dimensionale dei 500 dipendenti. Coerentemente anche tutte le unità panel soggette ad eventi di trasformazione societaria, quali fusioni, scorpori ecc. restano in rilevazione con tutte le loro componenti al fine di ricostruire l'omogeneità delle unità originali.

[4] Per maggiori informazioni si veda: Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese, Istat Collana Metodi e Norme n. 29/2006.

La scelta di utilizzare un panel chiuso è stata motivata in primo luogo dalla mancata disponibilità di un archivio esaustivo degli eventi di modificazione longitudinale, che fosse aggiornato con la tempestività necessaria. Questo avrebbe comportato necessariamente la raccolta di informazioni in modo non sistematico e sulla base di fonti informative eterogenee, anche non ufficiali (giornali, comunicazioni estemporanee degli stessi rispondenti, risultanze di altre indagini, eccetera) causando aggiornamenti parziali e disomogenei. In particolare, si sarebbe corso il rischio di privilegiare la copertura nelle attività economiche per le quali maggiore è la diffusione delle informazioni sui mezzi di comunicazione di massa. In secondo luogo, la complessità di un aggiornamento continuo contrasta con gli stringenti vincoli temporali e di risorse ai quali è soggetta un'indagine mensile, dove le necessità operative condizionano inevitabilmente anche la definizione della metodologia di raccolta ed elaborazione dei dati.

La considerazione delle difficoltà di registrare, in tempi utili, la realtà di tutte le trasformazioni ha portato a valutare attentamente tali effetti sugli indicatori, che come evidenziato precedentemente, se registrati in modo incompleto o parziale, potrebbero a misurare in modo distorto il secondo elemento della formula (1). Per questo motivo è stata fatta la scelta di aggiornare il panel solo al momento del cambio base e di garantire la sua omogeneità nel periodo intermedio con specifiche regole di trattamento degli eventi.

Tuttavia, per monitorare costantemente l'evoluzione del campo di osservazione teorico e quindi poter valutare nel tempo tale scelta, è stato deciso di individuare e sottoporre a rilevazione le nuove grandi imprese registrate nell'archivio Asia (passaggi di soglia e nascite). I dati raccolti vengono inoltre utilizzati per il passaggio alla nuova base in cui il panel di indagine viene aggiornato con l'inserimento delle nuove grandi imprese e l'eliminazione di tutte le unità scese sotto la soglia dei 500 dipendenti.

3. La simulazione sui dati dell'indagine GI

La simulazione, presentata in questo lavoro, ha l'obiettivo di valutare gli effetti sugli indicatori dell'indagine GI di diverse scelte relative alle modificazioni delle imprese. E' stato reputato essere l'indagine GI un caso studio interessante sia per la caratteristiche delle grandi imprese, particolarmente soggette alle trasformazioni giuridiche, e per la natura dell'indagine, congiunturale su un panel di imprese. I dati utilizzati sono quelli provenienti dell'indagine stessa, in particolare dal cospicuo serbatoio di informazioni correntemente raccolte per monitorare l'evoluzione del campo di osservazione.

La simulazione ha consistito nel riformulare il disegno di indagine nei suoi tre punti cardine:

a) l'utilizzo di un panel aperto con aggiornamento annuale, ossia un panel che considera le imprese entrate e uscite nel/dal campo di osservazione per effetto delle modificazioni demografiche e dimensionali;
b) una diversa modalità di trattamento statistico delle modifiche longitudinali. In sostanza, mentre le convenzioni attualmente in uso hanno lo scopo di mantenere il dato statistico della singola unità soggetta a un evento confrontabile con lo stato dell'unità precedente all'evento all'interno del panel chiuso, le nuove regole si propongono di rappresentare le unità statistiche così come si trasformano nel tempo;
c) l'adattamento della metodologia di calcolo dell'indice della retribuzione lorda pro capite per mantenere, in presenza di un panel aperto, la coerenza longitudinale della

serie storica. Infatti, come accennato nel paragrafo 2 l'indagine adotta due diverse metodologie di calcolo per gli indici dell'occupazione e per quelli pro capite delle variabili ore lavorate e retribuzione. Le due metodologie di calcolo presentano caratteristiche diverse rispetto alla registrazione degli effetti della natimortalità in particolare la prima è studiata appositamente per azzerare tali effetti mentre la seconda è influenzata anche da questi ultimi. Nel prosieguo per i due tipi di indici vengono presentate singolarmente le caratteristiche che ne permettono o meno l'utilizzo anche per panel definiti aperti. In particolare, per gli indici pro capite viene presentata una proposta di adattamento funzionale al disegno adottato nella simulazione con il panel aperto.

Nella simulazione è stato effettuato il calcolo degli indici dell'occupazione e della retribuzione lorda per dipendente con i dati dell'indagine GI sul biennio 2003-2004, utilizzando il panel aggiornato annualmente sulla base della natimortalità effettiva e dei passaggi di soglia dimensionale registrati nelle grandi imprese. L'analisi dei risultati è stata effettuata attraverso il confronto tra gli indici ottenuti nella simulazione e quelli correntemente pubblicati dall'indagine per analizzarne le differenze sia in termini di livello che di dinamica mensile.

3.1 L'aggiornamento del panel Grandi imprese per la simulazione

Dal punto di vista operativo il nuovo panel aperto è stato ottenuto aggiornando il panel GI base 2000 nei mesi di gennaio per gli anni 2003 e 2004. All'inizio di ciascun anno si è provveduto a:

- escludere dal panel le imprese, che hanno avuto in media annua una dimensione inferiore alla soglia dei 500 dipendenti;
- aggiungere le imprese individuate come nuove grandi imprese. Si ricorda che tali imprese, sebbene non inserite nel panel base 2000 per il calcolo degli indici diffusi, erano comunque oggetto di rilevazione al fine di acquisire le informazioni necessarie per il passaggio alla base successiva (base anno 2005).

L'effetto complessivo sulla struttura del panel 2000 dei due aggiornamenti in termini di occupazione è un aumento del peso relativo del settore del terziario di due punti percentuali (Prospetto 3.1).

Prospetto 3.1 - Struttura del panel 2000 prima e dopo gli aggiornamenti 2003 e 2004

ATTIVITÀ ECONOMICA (Ateco 2002)	Occupati (valori assoluti in migliaia)			Occupati (valori percentuali)		
	panel 2000	panel 2003	panel 2004	panel 2000	panel 2003	panel 2004
Industria (sezioni C-F)	865	854	845	42,4	41,5	40,4
Servizi (sezioni G-K)	1.177	1.205	1.248	57,6	58,5	59,6
Totale (sezioni C-K)	**2.042**	**2.059**	**2.093**	**100,0**	**100,0**	**100,0**

Nel prospetto 3.2 sono riportate per settore di attività economica le incidenze percentuali delle unità e della corrispondente occupazione escluse o aggiunte. In generale si nota che le unità escluse sono più numerose di quelle aggiunte, sia nel 2003 che nel 2004. L'industria è particolarmente toccata da tale revisione, con l'eliminazione di quasi il 9,0 per cento delle unità presenti nel panel originario nel 2003 e del 10,5 per cento nel 2004.

Viceversa il settore dei servizi si distingue per l'aumento delle imprese e dell'occupazione (rispettivamente più 9,4 per cento e più 4,2 per cento).

Prospetto 3.2 - Unità funzionali e occupati esclusi e aggiunti nel panel 2000 - Anni 2003-2004
(incidenze percentuali sul panel 2000)

ATTIVITÀ ECONOMICA (Ateco 2002)	2003				2004			
	Imprese		Occupazione		Imprese		Occupazione	
	Escluse	Aggiunte	Escluse	Aggiunte	Escluse	Aggiunte	Escluse	Aggiunte
Industria	**8,9**	**3,0**	**2,7**	**1,6**	**10,5**	**2,4**	**3,1**	**2,1**
Estrazione minerali	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Attività manifatturiere	8,8	1,5	2,9	1,4	10,6	2,0	3,5	1,9
Energia, gas e acqua	13,3	2,2	0,9	1,7	13,0	2,2	1,1	1,7
Costruzioni	7,1	21,4	3,1	18,1	0,0	21,4	0,0	20,2
Servizi	**3,9**	**6,0**	**0,4**	**2,8**	**4,2**	**9,4**	**0,4**	**4,2**
Commercio	2,5	9,8	0,2	4,3	6,1	12,2	0,3	5,5
Alberghi e ristoranti	12,0	4,0	1,7	3,0	12,0	8,0	1,6	4,9
Trasporti, magazzinaggio e comunicazioni	1,1	5,3	0,1	1,8	2,1	6,3	0,2	2,0
Intermediazione monetaria e finanziaria	3,2	3,9	0,2	1,0	4,9	5,7	0,5	1,6
Altre attività professionali e imprenditoriali	7,2	7,1	1,4	9,7	7,4	16,0	1,2	18,4
Totale	**7,5**	**3,2**	**1,4**	**2,3**	**8,6**	**3,2**	**1,5**	**3,4**

3.2 Il trattamento delle trasformazioni giuridiche nell'indagine GI e nella simulazione

Uno degli aspetti fondamentali dell'indagine GI è la gestione degli eventi legati alle unità al fine di garantire l'omogeneità sostanziale del panel in ciascun mese e assicurarne la confrontabilità nel tempo. Quindi, è stata necessaria la definizione di modalità di trattamento convenzionale delle trasformazioni giuridiche, definendo un trattamento ad hoc di tali eventi per ridurre al minimo l'effetto sugli indici dovuto alla demografia di impresa.

Per quanto riguarda la simulazione, coerentemente con la definizione di panel aperto, si è proceduto ad un diverso trattamento degli eventi di trasformazione societaria e di natimortalità delle unità (reale o causata dai passaggi di soglia dimensionale) rispetto a quanto correntemente effettuato dall'indagine sul panel chiuso.

Utilizzando lo schema di classificazione descritto nel prospetto 1.1, vengono illustrate le scelte operative adottate per il trattamento delle modificazioni longitudinali nell'indagine GI con panel chiuso e nella simulazione con panel aperto (Prospetto 3.3). Rispetto allo schema generale sono presenti due eventi aggiuntivi che riguardano il passaggio della soglia dimensionale dei 500 dipendenti (sopra la soglia/ sotto la soglia), che per l'indagine sono concettualmente assimilabili rispettivamente alla nascita o alla cessazione. Nel prospetto sono elencate, per ogni tipologia di evento, le conseguenze sul panel a seconda che si adotti la strategia seguita correntemente in GI (panel chiuso) oppure quella sperimentata in questo contesto finalizzata ad aggiornare il panel a seconda delle modifiche avvenute sia nelle imprese già appartenenti al panel, sia in quelle esterne al panel iniziale. Dal prospetto emerge chiaramente che con il panel chiuso si verificano

diversi casi in cui si continuano a inserire nel calcolo degli indici mensili imprese non più appartenenti al dominio di interesse (frammentazione 2 e 3 e 4) poiché scese sotto la soglia, mentre non vengono considerate le nuove grandi imprese (nascita, frammentazione 6). Per quanto riguarda il panel chiuso, nei casi di passaggio di soglia le unità vengono trattati in modo particolare: tutte le imprese vengono mantenute nel panel qualunque sia la dimensione occupazionale. Alla base di tale scelta c'è l'ipotesi che anche se un'impresa scende sotto la soglia continua ad avere un comportamento economico assimilabile a quello di grande impresa.

Prospetto 3.3 - Trattamento degli eventi di modificazione longitudinale delle imprese nell'indagine GI e nella simulazione

EVENTO	Descrizione evento	Impresa a fine evento			
		Presenza indagine GI		Presenza simulazione	
		i	*j*	*i*	*j*
Modifiche anagrafiche	Impresa *i* presente nel panel cambia ragione sociale e/o codice fiscale e/o attività economica e si trasforma in una nuova impresa *j*		Si		Si
Modifiche di esistenza: nascita	Impresa *j* che nasce nell'intervallo temporale considerato con almeno 500 dipendenti	No		Si	
Modifiche di esistenza: cessazione	Impresa *i* presente nel panel che cessa l'attività	No		No	
Concentrazione 1	Impresa presente nel panel si fonde con (o acquisisce parte di) un'impresa esterna al panel; la nuova impresa costituita almeno di 500 dipendenti		Si		Si
Concentrazione 2	Impresa presente nel panel si fonde con (o acquisisce parte di) un'impresa esterna al panel GI; la nuova impresa costituita ha meno di 500 dipendenti		Si		No
Frammentazione 1	Impresa presente nel panel si scinde in due nuove imprese *i* e *j* entrambe con almeno 500 dipendenti	Si	Si	Si	Si
Frammentazione 2	Impresa presente nel panel si scinde in due nuove imprese *i* e *j* di cui una (*i*) con almeno e l'altra (*j*) con meno di 500 dipendenti	Si	Si	Si	No
Frammentazione 3	Impresa presente nel panel si scinde in due nuove imprese *i* e *j* entrambe con meno di 500 dipendenti	Si	Si (a)	No	No
Frammentazione 4	Impresa presente nel panel si scinde in due nuove imprese *i* e *j,* dopo l'evento *i* ha meno di 500 dipendenti mentre l'impresa j ha almeno di 500 dipendenti	Si	SI	No	Si
Frammentazione 5	Impresa presente nel panel si scinde in due nuove imprese *i* e *j,* dopo l'evento l'impresa *i* ha almeno 500 dipendenti e l'impresa *j* ha meno di 500 dipendenti	Si	SI	Si	No
Frammentazione 6	Impresa *i* presente nel panel cede una parte degli addetti a un'impresa *j* esterna al panel. Dopo l'evento sia l'impresa *i* sia l'impresa *j* hanno più di 500 dipendenti	Si	No (b)	Si	Si
Passaggio di soglia 1	Impresa *i* presente nel panel scende sotto la soglia di 500 dipendenti	Si		No	
Passaggio di soglia 2	Impresa *i* non presente nel panel sale sopra la soglia dei 500 dipendenti	No		Si	

(a) A meno che l'impresa *j* non sia molto piccola e difficilmente monitorabile.
(b) Solo nel caso in cui la parte ceduta rappresenta una quota prevalente dell'attività, in termini di occupati, sia della cessionaria che della cedente viene inserita nel panel anche l'impresa j.

In tutti gli altri casi la variazione di stato viene registrata con modalità finalizzate a mantenere l'omogeneità temporale del panel come definito nell'anno base. Per esempio, nel caso di una scissione in cui un'impresa dà vita a più imprese queste vengono tutte registrate

all'interno del panel a prescindere della loro dimensione occupazionale, secondo la logica che l'unione dei loro dati fornisce un dato formalmente omogeneo a quello dell'impresa precedente all'evento. Invece, nel caso di uno scorporo ovvero un'unità che cede parte dei propri occupati a un'altra azienda, se questa è interna al panel si registrerà un flusso in entrata e uno in uscita all'interno del panel stesso, se la seconda è esterna al panel si avrà soltanto un flusso in uscita. Ragionamento analogo ma opposto avviene nel caso di una fusione, qualora essa avvenga tra un'impresa appartenente al panel e una esterna, si registrerà un flusso in entrata nel panel di rilevazione.

3.3 Indice dell'occupazione

Gli indici di variazione dell'occupazione sono ottenuti con una tecnica di concatenamento frequentemente utilizzata negli indici congiunturali. Essa permette un ribasamento implicito dell'indice in modo da considerare solamente le unità attive ad ogni tempo *T* senza registrare le variazioni dovute alla natimortalità.

Partendo dalla formula dell'indice (1) è possibile scrivere:

$$I_T = \frac{Y_T}{Y_B} = \frac{Y_T}{Y_{T-1}} \cdot \frac{Y_{T-1}}{Y_{T-2}} \cdot .. \frac{Y_1}{Y_B} \cdot Y_B \qquad T=1,2,\ldots$$

Si deduce, quindi, che è possibile calcolare l'indice come prodotto degli indici mobili mensili, ovvero delle variazioni registrate rispetto al periodo precedente. Questa formulazione introduce il concetto di concatenamento, ovvero l'indice al tempo *T* è dato dal prodotto degli indicatori mobili mensili concatenati mese per mese. E' possibile quindi calcolare le variazioni del fenomeno indipendentemente ad ogni tempo *T*, per misurare le variazioni solo sulle imprese appartenenti alla popolazione così come è definita nel momento più recente di osservazione. E' necessario acquisire dalle unità statistiche il valore della data variabile *Y* sia all'inizio che alla fine del periodo di osservazione, perché per ogni periodo *T* sia possibile calcolare la variazione a livello aggregato solamente sulle unità attive sia all'inizio che alla fine del periodo osservato, anche se nella loro composizione diverse dall'insieme misurato nel tempo precedente. In termini diversi, questa proprietà permette un ribasamento implicito dell'indice.

La tecnica del concatenamento è stata adottata per l'indice dell'occupazione, che deve fornire la variazione del numero di posizioni lavorative dipendenti alla fine del mese di riferimento rispetto al valore medio registrato nell'anno base. Operativamente, ad ogni unità che fa parte del panel si chiede di fornire il numero degli occupati presenti alla fine del mese corrente e alla fine del mese precedente a quello di rilevazione (che deve coincidere con il valore all'inizio del mese). Cosicché ogni singola unità fornisce la variazione a cui è interessata l'indagine e l'acquisizione contemporanea dello stock di occupati relativi a due periodi consecutivi permette di elaborare uno stimatore della variazione occupazionale che ottimizza tale informazione.

Per semplicità espositiva definiamo per ogni unità *j*, in ogni mese *T* e per ogni aggregazione Ateco *a* che rappresenta il livello minimo a cui vengono calcolati gli indicatori (nel caso dell'indagine GI gruppo Ateco a 3 digit) le seguenti quantità:

$S = \bigcup_{a \subset S} a$ generica aggregazione *S* di ordine superiore al gruppo Ateco *a;*

$OI_{j,a,T}$ = occupazione all'inizio del mese *T;*

$OF_{j,a,T}$ = occupazione presente alla fine del mese *T*.

Quindi per ogni mese *T* il tasso di variazione congiunturale si ottiene sommando i valori di fine e inizio periodo di tutte le unità appartenenti al gruppo Ateco *a*:

$$\Phi_{a,T} = \frac{\sum_{i \in a} OF_{a,T}}{\sum_{i \in a} OI_{a,T}} \text{ : tasso di variazione occupazionale del gruppo a nel mese } T \qquad (4)$$

Il tasso di variazione è l'elemento centrale su cui si basa il concatenamento che assicura la registrazione solamente dei flussi in entrata e in uscita degli occupati sulle imprese attive nel mese di riferimento. L'indicatore finale per la generica aggregazione *S* è dato da:

$$I_{S,T} = \sum_{a \in S} \left(\Phi_{a,T} \; I_{a,(T-1)} \; \frac{\overline{OF}_{a,B}}{\overline{OF}_{S,B}} \right)$$

Inoltre, ponendo:

$$Z_{a,B} = \frac{\overline{OF}_{a,B}}{\overline{OF}_{S,B}}$$

è possibile scrivere l'indice come

$$I_{S,T} = \sum_{a \in S} \left(\Phi_{a,T} \; I_{a,(T\text{-}1)} \; Z_{a,B} \right) \qquad (5)$$

ovvero per ogni mese *T* e per ogni aggregazione *S* l'indice dell'occupazione è calcolato come media degli indici del mese *T*-1 dei gruppi *a* che formano il settore *S* (con pesi occupazionali alla base) moltiplicati per i rispettivi tassi di variazione al tempo *T*.

Quindi nella simulazione, in cui si considera di sostituire il panel chiuso con un panel aperto, è sufficiente concentrare l'attenzione sul calcolo del tasso Φ. Gli indici relativi al mese *T* nel panel aggiornato possono essere costruiti applicando agli indici calcolati con il panel chiuso nell'ultimo mese *T*-1, precedente all'aggiornamento, la variazione congiunturale registrata sul nuovo panel di rilevazione nel mese *T*.

3.4 Gli indici pro capite: disegno attuale e proposta alternativa

Gli indici pro capite sono indici semplici a base fissa dati dal rapporto tra il valore medio pro capite del mese di riferimento e il corrispondente valore pro capite medio mensile dell'anno base. Introducendo le seguenti nuove definizioni:

$Y_{S,B}$ = ammontare mensile della generica variabile Y nel settore S nell'anno base B;

$ON_{S,B}$ = occupazione media nel settore *S* nell'anno base *B;*

$Y_{S,T} = \sum_{j \in S} y_{j,S,T}$: ammontare della generica variabile *Y* nel settore *S* nel mese *T;*

$ON_{S,T} = \sum_{j \in S} ON_{j,T}$: occupazione media del settore *S* nel mese *T*;

Nella metodologia attuale per la generica variabile *Y* gli indici pro capite sono calcolati nel seguente modo:

$$I_{Y,T,S} = \frac{(Y_{S,T} / ON_{S,T})}{(\bar{Y}_{S,B} / ON_{S,B})} \quad \text{per ogni mese } T \text{ e aggregazione } S \tag{6}$$

Attraverso dei semplici passaggi algebrici, anche questo indicatore può essere espresso in funzione degli indici elementari relativi ai gruppi Ateco *a*, infatti ponendo:

$Q_{Y,a,B} = \frac{Y_{a,B}}{Y_{S,B}}$: rapporto del valore di *Y* del gruppo *a* sul totale del settore *S*;

$Z_{a,T} = \frac{ON_{a,T}}{ON_{S,T}}$: peso dell'occupazione del gruppo *a* sul totale del settore *S* al tempo *T*;

$Z_{a,B} = \frac{ON_{a,B}}{ON_{S,B}}$: peso dell'occupazione del gruppo *a* sul totale del settore *S* alla base *B*

si ha:

$$I_{Y,S,T} = \sum_{a \subset S} \left(I_{Y,a,T} \, Q_{Y,a,B} \, \mathrm{Z}_{a,T} / \mathrm{Z}_{a,B} \right) \tag{7}$$

Tale formulazione mostra che gli indici pro capite registrano sia le variazioni delle variabili oggetto di studio *Y*, sia della struttura dell'occupazione (Istat, 2010). In altri termini, si può parlare di indici di valore che tengono conto sia delle variazioni di prezzo, sia delle variazioni di quantità poiché la ponderazione dipende da come cambia la struttura occupazionale tra i diversi gruppi *a* all'interno del settore *S* rispetto a quella rilevata nell'anno base *B*. Quello che manca, a differenza dell'indice sull'occupazione, è un elemento che permette di individuare solamente il gruppo delle imprese attive nel mese di riferimento e di misurare su di esse il fenomeno al netto della variazione della struttura occupazionale.

Per questo motivo in sede di simulazione è stata proposta una metodologia alternativa che contempli un termine di concatenamento. I nuovi indici pro capite sono espressi in funzione del tasso di variazione mensile della retribuzione lorda pro capite concatenati all'indice relativo allo stesso mese dell'anno precedente, in questo caso è stata considerata variazione tendenziale per eliminare gli effetti della stagionalità.

Il tasso di variazione introdotto è dato da:

$$\Theta_{a,T} = \frac{\Delta Y_{a,T}}{\Delta ON_{a,T}} = \frac{Y_{a,T} / Y_{a,T-12}}{ON_{a,T} / ON_{a,T-12}}$$

e quindi il nuovo indice della retribuzione pro capite diventa:

$$I_{Y,S,T} = \sum_{a \ S} \left(\Theta_{a,T} \ \ I_{Y,a,T\text{-}12} \ \ Q_{Y,a,B} \ Z_{a,T} / Z_{a,B} \right) \tag{8}$$

La differenza tra i due indici espressi dalle (8) e (7) consiste nell'introduzione del fattore di concatenamento Θ che permette di calcolare il tasso di variazione sul panel successivamente alla registrazione degli eventi di modifica.

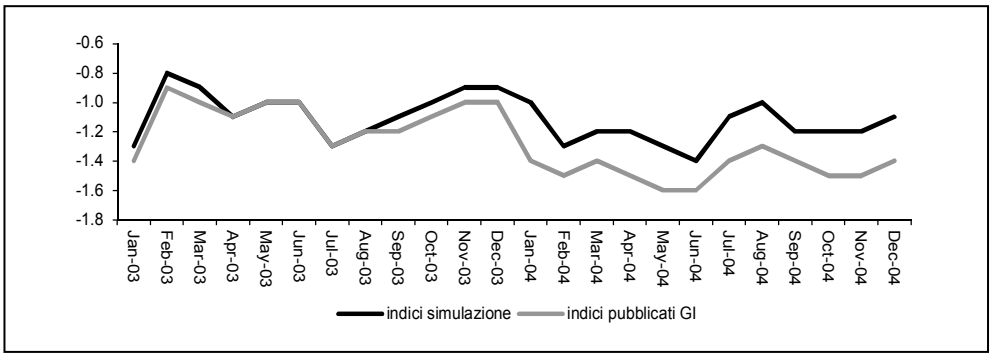
4. Analisi dei risultati

L'analisi dei risultati è stata fatta confrontando le variazioni tendenziali dei nuovi indici ottenuti per l'occupazione e per la retribuzione lorda pro capite (per il totale delle grandi imprese e separatamente per industria e servizi) con quelle degli indici attualmente diffusi dall'indagine, con l'obiettivo di verificare se il ricorso a un panel aggiornato conduca o meno a dinamiche longitudinali significativamente diverse da quelle ottenute con la tecnica attuale. Si ricorda che per la retribuzione lorda pro capite l'analisi si basa sul confronto tra i risultati ottenuti utilizzando due tipi di indice (rispettivamente espressi nella formula 7 i pubblicati e nella 8 i simulati), mentre per l'occupazione il metodo di calcolo dell'indice è sempre lo stesso (come espresso nella 5).

4.1 Variabile occupazione

Per quanto riguarda l'occupazione nel totale delle grandi imprese (Figura 4.1), l'andamento delle due serie di variazioni risulta molto simile in tutti i mesi del 2003 (con l'eccezione del mese di settembre), mentre nel 2004 le differenze tra le due serie si ampliano pur rimanendo abbastanza contenute. In particolare è importante notare che sia a gennaio 2003 che a gennaio 2004, mesi in cui è stato fatto l'aggiornamento del panel, le differenze sono minime e tali comunque da non comportare discontinuità nella serie storica. Per quanto riguarda l'entità delle variazioni tra gli indici diffusi e quelli della simulazione si riscontra una differenza media di circa 0,2 punti percentuali.

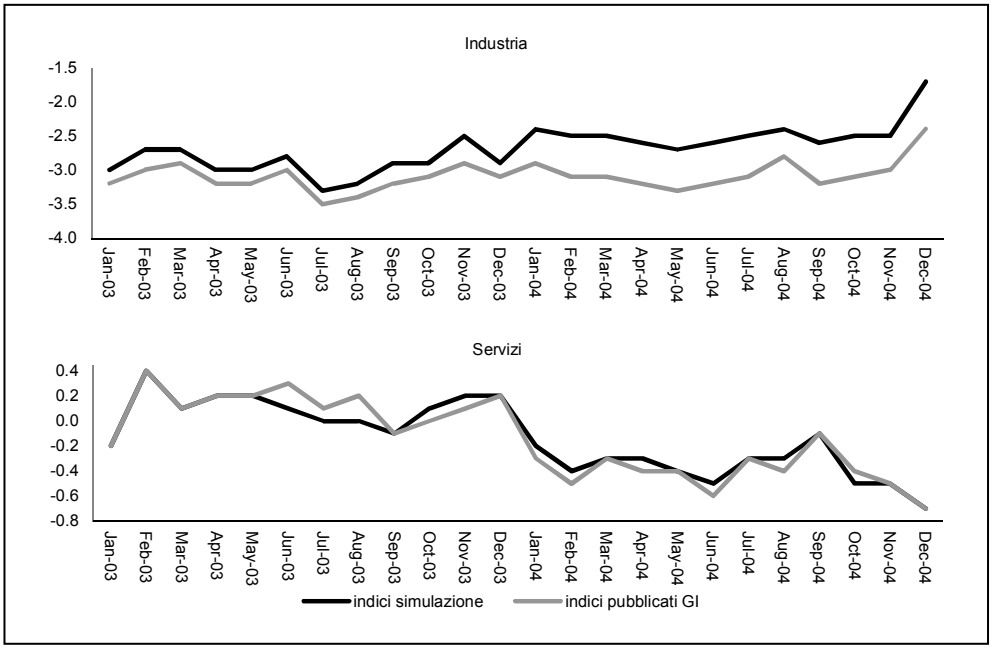
Figura 4.1 - Variazioni tendenziali degli indici dell'occupazione nel totale delle grandi imprese – Anni 2003-2004 *(variazioni percentuali di indici in base 2000=100)*



Osservando separatamente i due settori dell'industria e dei servizi (Figura 4.2), si evidenzia che tale differenza può essere ricondotta prevalentemente a quello industriale, nel quale la nuova serie si colloca in media circa 0,4 punti percentuali sopra la serie pubblicata, mentre nei servizi le due serie presentano livelli e andamento delle variazione analoghi.

Facendo riferimento alle modalità di aggiornamento del panel, si ricorda che mentre nel settore dei servizi sono state soprattutto aggiunte nuove imprese il settore dell'industria è stato aggiornato prevalentemente attraverso l'eliminazione di numerose unità scese ormai ampiamente sotto la soglia dimensionale dei 500 dipendenti (paragrafo 3.1). In sostanza per l'occupazione la differenza nelle due serie di indici viene determinata soprattutto dal diverso andamento congiunturale delle unità escluse mentre l'operazione di aggiunta di nuove imprese ha effetti molto più contenuti.

Figura 4.2 - Variazioni tendenziali degli indici dell'occupazione nell'industria e nei servizi – Anni 2003-2004 *(variazioni percentuali di indici in base 2000=100)*

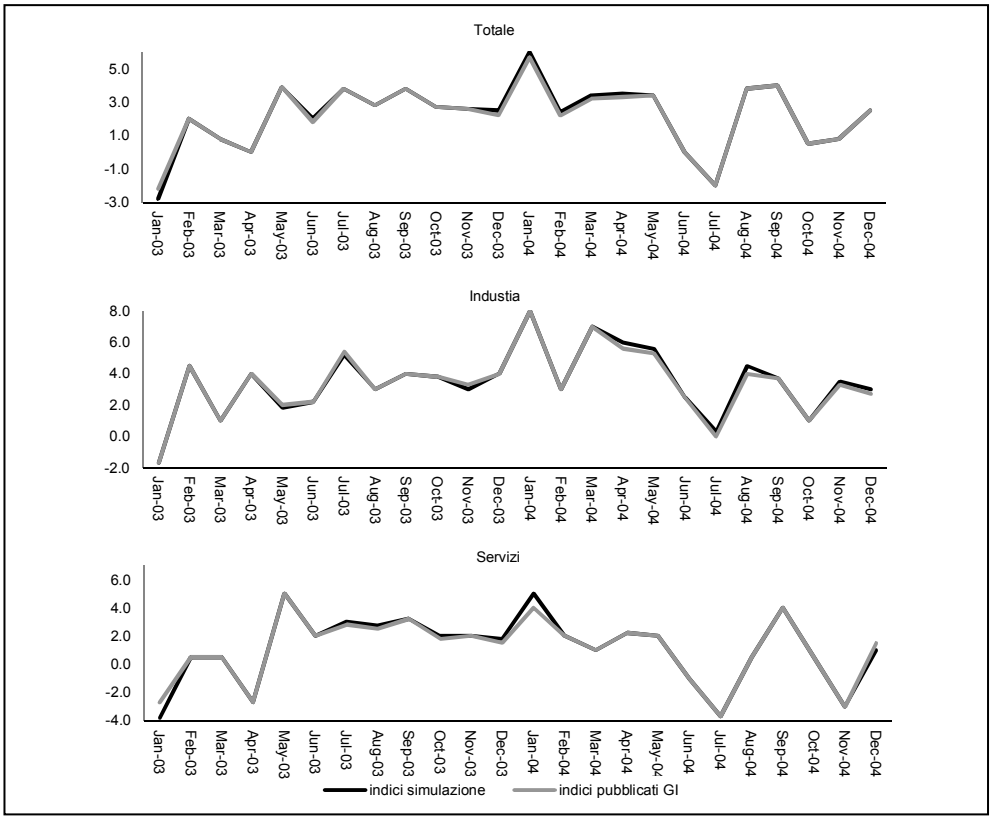


4.2 Variabile retribuzione lorda pro capite

Per quanto riguarda gli indici della retribuzione lorda pro capite, gli effetti dell'aggiornamento del panel sono diversi da quelli registrati sugli indici dell'occupazione. Nel totale delle grandi imprese i valori delle differenze della variabile retributiva sono frequentemente nulli o molto contenuti, soprattutto in considerazione del fatto che si tratta di un retribuzione mensile di cassa che ha un andamento molto più variabile rispetto all'occupazione (come evidenziato dalla scala delle variazioni sulle ordinate utilizzata nelle rispettive figure). Anche l'analisi settoriale riportata nella figura 4.3 rileva differenze tra le due serie non sostanziali con spezzate quasi sovrapposte.

Nello specifico in entrambi i settori sono presenti alcune differenze comprese tra più e meno 0,5 punti percentuali, con numerosi casi di differenze nulle o prossime allo zero. L'unica differenza significativa tra le due serie riguarda il settore dei servizi sul mese di gennaio 2003 dove la variazione tendenziale della serie simulata sul panel aperto è 1,1 punti percentuali inferiore rispetto a quella diffusa calcolata sul panel chiuso (che ha un effetto nel totale di meno 0,6 punti percentuali). Il motivo è riconducibile all'aggiunta delle nuove imprese, che in quel mese hanno avuto una retribuzione media pro capite più bassa di quella registrata sulle permanenti. Il salto registrato a gennaio 2003 non è quindi dovuto al diverso livello retributivo delle imprese entrate o uscite, ma ad un evento isolato che ha influenzato solo la dinamica di quel mese. Una conferma delle ultime osservazioni si ha nei mesi successivi, in cui le serie sono sempre molto simili e anche la differenza che si evidenzia a gennaio 2004, dove i due indici assumono lo stesso valore in termini assoluti (102,8), è un effetto di rimbalzo riconducibile a quanto accaduto nello stesso mese dell'anno precedente. Con riferimento alle retribuzioni lorde pro capite si può quindi concludere che le variazioni tendenziali dell'insieme delle unità aggiunte risultano molto vicine a quelle dell'insieme delle permanenti (e tali, quindi, da non modificare in modo significativo la serie degli indici finali), a differenza di quanto riscontrato per gli indici dell'occupazione.

Figura 4.3 - Variazioni tendenziali degli indici della retribuzione lorda pro-capite nel totale delle Grandi imprese, nell'industria e nei servizi- Anni 2003-2004 *(variazioni percentuali indici in base 2000=100)*



5. Conclusioni

Le riflessioni proposte nel documento sono relative all'effetto del trattamento delle modifiche longitudinali delle imprese nell'indagine GI sulla misurazione delle variazioni dell'occupazione e della retribuzione lorda pro capite. Esse si basano su una simulazione empirica effettuata su dati dell'indagine, utilizzando un panel aperto in sostituzione dell'attuale panel chiuso. Nella simulazione è stata anche introdotta una variante metodologica per il calcolo degli indici della retribuzione lorda pro capite, necessaria per rendere coerente la struttura metodologica a seguito della diversa tipologia di panel utilizzato. Lo scopo finale è stato quello di misurare gli effetti che la natimortalità delle imprese hanno sugli indici di variazione prodotti dall'indagine GI osservando il loro profilo longitudinale.

I risultati hanno confermato che l'attuale esclusione dai calcoli degli effetti della natimortalità delle imprese rilevate, comprese quelle nel frattempo nate nell'universo di riferimento, conduce a stime differenti rispetto a quelle ottenibili con un panel aggiornato annualmente. Tuttavia, gli effetti registrati sono molto contenuti e praticamente irrilevanti per la retribuzione pro capite, mentre per l'occupazione si evidenzia la tendenza del panel chiuso a produrre una stima sistematicamente più negativa dell'andamento della variabile, anche se di entità contenuta. I problemi maggiori si registrano nel settore dell'industria. In questo settore, ci sono numerosi casi di imprese la cui dimensione è scesa sotto la soglia dei 500 dipendenti, registrando una costante decrescita nella variabile occupazionale. Invece, per la variabile retributiva i risultati ottenuti dimostrano che sia nel settore dell'industria, sia in quello dei servizi la struttura del panel chiuso è tale da cogliere l'andamento delle variabili retributive dell'universo corrente, senza che gli eventi demografici comportino particolari differenze tra le due serie storiche.

L'analisi di queste differenze porta a concludere che l'utilizzo del panel chiuso non garantisce l'assenza di effetti demografici sul calcolo delle variazioni degli indici prodotti, ma va sottolineato che tali effetti sono differenti a seconda della variabile e del settore oggetto di studio. In generale, si può affermare che le caratteristiche del profilo longitudinale (quindi delle variazioni mensili degli indicatori) sono manutenute pressoché uguali nei due panel. Questo porta ad affermare che anche rilevare solo le imprese del panel chiuso possa permettere di cogliere ciò che avviene nell'universo di riferimento, anche se nel frattempo questo è variato nella sua composizione e numerosità. L'unica eccezione importante riguarda le unità che da un punto di vista definitorio escono dell'universo teorico di riferimento (quelle scese sotto la soglia), la cui inclusione può provocare distorsioni nei calcoli degli indicatori in particolar modo della variabile occupazionale. Ciò fa supporre che tali imprese abbiano un diverso pattern economico rispetto a quello delle imprese che rimangono sopra la soglia dimensionale dei 500 dipendenti. Sebbene i risultati ottenuti in questo contesto sperimentale dovrebbero essere confermati da ulteriori verifiche empiriche, le soluzioni mostrate rappresentano delle proposte da valutare per rispondere a quesiti relativi al reale impatto delle trasformazioni longitudinali sugli indicatori congiunturali dell'indagine GI.

Più in generale, le procedure adottate per la simulazione possono essere inquadrate nel più ampio contesto della valutazione della cosiddetta "qualità globale" di un'indagine nel caso specifico di indagini non campionarie, ossia basate su un meccanismo di selezione del panel di tipo deterministico, in cui l'effetto dei principali fattori distorsivi (autoselezione dei rispondenti effettivi, modifiche dell'universo di riferimento) può crescere sensibilmente allontanandosi dal periodo scelto come base.

Riferimenti Bibliografici

Eurostat 2006, *Methodology of Short-term Business Statistics. Interpretation and Guidelines*, March 2006, Eurostat, Lussemburgo.

Eurostat – OECD 2007, *Manual on Business Demography Statistics.* OECD Publications, Paris 2008.

Eurostat 2010, *EU Business registers reccomandations manual.* Methodologies and working papers Eurostat, Lussemburgo 2010.

Gismondi R. 2001, "Nascite e cessazioni delle imprese: gli effetti sul calcolo di numeri indici", *Rivista di statistica ufficiale*, 3, 11-54, Istat, Roma.

Gismondi R. 2005, "Le modificazioni longitudinali delle imprese: definizioni, conseguenze sulle tecniche di stima di una variazione e il caso dell'indagine mensile sulla produzione industriale", rapporto tecnico, Istat, Roma.

Istat 1997, *Le modificazioni temporali delle unità in ASIA: definizioni, classificazioni e convenzioni*, Istat, Roma.

Istat 2006, "Rilevazione mensile sull'occupazione, gli orari di lavoro e le retribuzioni nelle grandi imprese", collana Metodi e norme n°29/2006, Istat, Roma.

Istat 2008, "Controllo e correzione nell'indagine mensile sulle grandi imprese: metodi e prime evidenze da un'analisi retrospettiva sulla qualità", collana Contributi n° 13/2008, Istat, Roma in corso di pubblicazione.

Istat 2010, "Retribuzione pro capite nelle grandi imprese: effetti di composizione dell'occupazione", collana Approfondimenti 30-03-2010.

Mol J. 1999, "Treatment of Register Changes", in *Short-term Statistics Methodological Manual*, 59-65, Eurostat, Lussemburgo.

ONU 1993, *System of National Accounts*, disponibile sul sito web http://unstats.un.org/unsd/sna1993.

Srinath K.P. 1987, "Methodological Problems in Designing Continuous Business Surveys: Some Canadian Experiences", *Journal of Official Statistics*, 3, 283-288.

Struijs P., Willeboordse A. 1995, "Changes in Populations of Statistical Units", *Business Survey Methods*, 65-84, John Wiley & Sons, New York.

Un archivio longitudinale amministrativo per la stima della povertà a livello locale[1]

Maria Elena Comune[2]

Sommario

In questo studio[3]*, è stato realizzato un archivio longitudinale amministrativo di redditi familiari e individuali, con la finalità di stimare i redditi e la povertà delle famiglie nel comune di Brescia per gli anni 2005-2008. A tal fine, sono stati utilizzati dati di fonte anagrafica e fiscali, opportunamente linkati. Come definizione di povertà, è stato adottato un approccio unidimensionale basato su un unico indicatore (il reddito "dichiarato") e una soglia di povertà relativa "locale", calcolata con riferimento al comune di Brescia. In questa sede, sono presentate la metodologia utilizzata, mettendo in evidenza i problemi incontrati e le soluzioni adottate, e le principali stime di redditi e povertà secondo una prospettiva trasversale e longitudinale.*

Parole chiave: stime trasversali e longitudinali, logorio del campione, regole di inseguimento, dati amministrativi, povertà relativa, soglia di povertà relativa locale, reddito.

Abstract

In this study, a longitudinal income database based on administrative data was set up aiming at estimating income and poverty of households and individuals living in the municipality of Brescia in the period 2005-2008. This was achieved by linking data sourced from the Population Register and fiscal authorities. The definition of poverty was obtained through a one - dimensional approach, based on a single indicator ("declared" income) and a local relative poverty line, evaluated with reference to the municipality of Brescia. The paper includes a description of both the used methodology, highlighting the problems faced and the adopted solutions, and of the main estimates of poverty rates from a cross-sectional and longitudinal perspective.

Keywords: cross-sectional and longitudinal estimates, attrition, tracing rules, administrative data, relative poverty, local relative poverty line, income.

[1] Il lavoro è stato presentato alle "*Giornate della Ricerca metodologica*", organizzate dall'Istat nei giorni 20-21 marzo 2013.

[2] Istat - Sede per la Lombardia (comune@istat.it). L'articolo pubblicato impegna esclusivamente l'Autore, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

[3] Realizzata dall'autore come tesi di dottorato in "Sociologia e Metodologia della Ricerca Sociale" presso l'Università Cattolica del Sacro Cuore di Milano. L'autore ringrazia i propri relatori Prof. Giancarlo Rovati, Università Cattolica del Sacro Cuore di Milano, il Prof. Gian Carlo Blangiardo, Università Bicocca di Milano e il Dr. Trentini del Comune di Brescia – Staff di Statistica.

1. Premessa

La ricerca, nata da una convenzione tra l'Università Cattolica del Sacro Cuore di Milano e il Comune di Brescia, si è posta l'obiettivo di stimare le disuguaglianze dei redditi e la povertà presenti a livello comunale. L'esigenza di conoscere lo stato d'indigenza e la povertà delle famiglie bresciane e i redditi da loro percepiti a livello locale è stata motivata da una totale assenza di statistiche ufficiali sui redditi e povertà a un livello territoriale così fine. E' noto, infatti, che le uniche indagini prodotte dall'Istituto Nazionale di Statistica - Istat che stimano in modo ufficiale la povertà (Reddito e condizioni di vita – EU-SILC e Consumi delle famiglie) sono di natura campionaria e forniscono dati rappresentativi solo a livello nazionale, per ripartizione e, al massimo, regionale.

L'esigenza a livello locale, pertanto, era quello di stimare i redditi e la povertà per il comune di Brescia non solo con riferimento ad un preciso anno, realizzando quindi analisi trasversali o *cross-section*, ma per più anni consecutivi dando origine ad analisi di tipo longitudinale ed evidenziando in questo modo l'andamento temporale del fenomeno oggetto di studio, soprattutto in considerazione dell'avvento della crisi economica nel 2008.

Per raggiungere questo obiettivo, è stato realizzato un archivio trasversale e longitudinale amministrativo sui redditi familiari e individuali (che da qui in poi chiameremo ALAR), basato sull'utilizzo di dati anagrafici provenienti dall'Anagrafe comunale e di dati fiscali provenienti dal *Sistema di Interscambio Anagrafe Tributaria Enti Locali dell'Agenzia delle Entrate* (SIATEL)[4], opportunamente agganciati.

Attraverso le stime dei redditi è stato possibile stimare la povertà relativa delle famiglie e degli individui per gli anni indicati, secondo un approccio di povertà basata sul reddito dichiarato in sede di denuncia dei redditi e riferita al contesto locale, calcolando una linea di povertà relativa "*locale"* per il comune di Brescia.

La realizzazione dell'archivio longitudinale ha permesso, inoltre, di analizzare le variazioni delle condizioni economiche familiari intervenute nel periodo di riferimento, di valutare la natura della povertà, se transitoria o permanente, e le transizioni da uno stato di povertà a uno di benessere e viceversa.

Questo contributo presenta la metodologia statistica utilizzata per costruire l'archivio trasversale e longitudinale, i principali risultati ottenuti e i limiti delle fonti amministrative utilizzate.

2. L'uso di dati amministrativi nella ricerca longitudinale

L'importanza di avere stime longitudinali nella ricerca statistica sociale ed economica è ben nota. E' risaputo, infatti, che le analisi dei dati longitudinali possono permettere di indagare, in una prospettiva dinamica diversi aspetti della popolazione oggetto di studio in campo sociale: dal monitoraggio dei cambiamenti sociali intervenuti nel tempo, alla valutazione dell'impatto sociale delle politiche, alla mobilità della popolazione sul territorio con le loro caratteristiche e motivazioni sociali annesse e così via (Golini 2001).

[4] E' un sistema di collegamento telematico che consente lo scambio attivo di informazioni anagrafiche e tributarie fra Amministrazione pubblica centrale e locale. Attraverso questo sistema, i Comuni, le Province, le Regioni, i Consorzi di bonifica e le Comunità montane possono consultare i dati posseduti dalla banca dati dell'Amministrazione Finanziaria per l'espletamento delle loro funzioni, ai sensi della legge n. 675 del 31/12/1996 (e successive modificazioni e integrazioni).

La metodologia più utilizzata per ottenere informazioni di tipo longitudinale, oltre all'inserimento di quesiti retrospettivi nei questionari, è quella di seguire in un arco temporale definito le stesse unità di rilevazione, andando a costituire un panel di individui o famiglie.

Purtroppo, l'Italia non vanta una lunga tradizione nella ricerca longitudinale[5]. La scarsa presenza di indagini panel sulle famiglie nella statistica ufficiale e, soprattutto, nelle discipline sociali, in passato, è stata infatti denunciata da molti ricercatori e studiosi (Trivellato 1995; Ruspini 2004).

Nel 1996, al fine di rispondere alle crescenti esigenze metodologiche e richieste di ricercatori, fu insediata dal Presidente dell'Istat una Commissione di studio cui fu affidato il compito di formulare proposte in merito alla progettazione di campioni per le indagini longitudinali sulle imprese e famiglie, che potevano essere condotte dall'Istat (Leti 2001)[6].

E' solo di recente che in Italia, grazie alle richieste dell'Unione europea, è stata introdotta nella statistica ufficiale l'indagine panel *European Community Household Panel - ECHP* (Istat 2002), sostituita nel 2004 dall'indagine sui redditi e condizioni di vita *Statistics on Income and Living Conditions - EU-SILC* (Istat 2008a).

I principali motivi di questa scarsa presenza di indagini longitudinali in Italia sono di certo legati alle difficoltà di realizzazione degli stessi panel, considerando gli elevati costi, la complessità organizzativa, metodologica e di elaborazione dei dati che presentano.

Per realizzare un'indagine longitudinale, è richiesto infatti un salto di qualità rispetto alla progettazione e conduzione delle indagini tradizionali, in relazione a diversi aspetti della rilevazione, da quello metodologico, a quello organizzativo, a quello finanziario.

A fronte di un così forte impegno della fase di progettazione, i vantaggi analitici offerti dai panel rispetto agli studi trasversali ripetuti o a una singola indagine longitudinale retrospettiva sono ben numerosi ed evidenziati nella letteratura scientifica (Lalla 2003).

Anziché ricorrere alle tradizionali indagini panel, basate sulla somministrazione di un questionario ad un campione di soggetti, ai fini di analizzare un fenomeno sociale o economico in una prospettiva longitudinale è possibile utilizzare informazioni amministrative già esistenti, dando origine ad un archivio o database longitudinale.

Nello specifico, costruire un archivio longitudinale utilizzando dati amministrativi significa mettere in relazione (o agganciare), mediante una chiave, le informazioni raccolte dalle istituzioni pubbliche e amministrative relative alle stesse unità di rilevazione e ad anni diversi. Di solito, le informazioni sono rilevate per fini amministrativi dell'Ente che li raccoglie (ad esempio per motivi fiscali, pensionistici, anagrafici o giuridici) e sono utilizzate per il controllo o l'intervento nei confronti di singoli individui, imprese o entità di tipo diverso (come le persone giuridiche) cui si riferiscono (Fortini 2000).[7]

Questa alternativa presenta numerosi vantaggi soprattutto se paragonata alla possibilità di rilevarle tramite intervista diretta e somministrazione di un questionario.

Le informazioni contenute in un archivio longitudinale, basate sulla raccolta e utilizzo di dati amministrativi, hanno il vantaggio di non presentare le tradizionali difficoltà e problematiche di costruzione legate alla natura campionaria del dato e alla somministrazione di un questionario, come mezzo di rilevazione del dato. Né presentano le

[5] Per avere un elenco di indagini longitudinali realizzate in Italia, Europa e nel Nord-America consultare Ruspini 2004.

[6] Gli studi avviati in quella sede, sono stati oggetto di discussione e pubblicazione nel 2001 (Rivista di Statistica Ufficiale, Istat, Quaderni di ricerca).

[7] Da diversi anni, in diversi Paesi del Nord Europa (tra cui ad esempio la Finlandia) è possibile ottenere in via continuativa informazioni statistiche dai dati amministrativi di origine fiscale (Fattore, Mezzanzanica 2007).

difficoltà legate ai problemi organizzativi come la gestione dei rilevatori e la reperibilità delle unità rispondenti, né quelli finanziari, poiché i costi dell'uso dei dati amministrativi sono contenuti. Inoltre, solitamente, i tempi di realizzazione di un archivio longitudinale amministrativo sono brevi, se confrontati con analoghe indagini condotte mediante intervista, poiché il dato amministrativo, se è disponibile, lo è in modo continuativo per tutte le unità di rilevazione, previa concessione, autorizzazione o vendita dell'ente pubblico.

Un grosso vantaggio dell'utilizzo del dato amministrativo è costituito dal poter indagare e analizzare il fenomeno oggetto di studio fino a un dettaglio territoriale piuttosto fine, come quello comunale, per il quale difficilmente esistono fonti di dati e informazioni statistiche.

Infine, si deve tenere conto che, nella realizzazione di un archivio longitudinale, i dati amministrativi non sono influenzati dai problemi legati alla natura ripetitiva che si verificano nelle indagini statistiche basate sulla somministrazione di un questionario, quali il condizionamento delle risposte, l'aumento della reattività all'intervista e il problema delle mancate risposte per alcune variabili sensibili (come la stima del reddito). L'impiego di dati amministrativi presenta anche il vantaggio di evitare la molestia statistica dei rispondenti che si verifica nelle indagini campionarie per troppe e, a volte continue, richieste da parte dei rilevatori di rispondere ai quesiti dell'indagine con i conseguenti rifiuti nel fornire le risposte, minando così l'efficienza delle stime.

I principali aspetti negativi dell'utilizzo di dati amministrativi risiedono nella qualità del dato fornito dagli Enti. Le informazioni desunte dagli archivi amministrativi potrebbero essere state rilevate dalle Amministrazioni Pubbliche in modo non adeguato ed essere inutilizzabili a fini statistici. Il problema nasce dal fatto che i dati amministrativi derivano da processi organizzativi che non sono progettati a fini statistici e quindi non sono sottoposti a processi di gestione della qualità del dato (da un punto di vista statistico), paragonabili a quelli messi in atto nell'ambito di indagini campionarie o censuarie progettate per fini conoscitivi (Fattore, Mezzanzanica 2007). Gli archivi possono, infatti, presentare problemi d'incompletezza, inaccuratezza, inadeguatezza, obsolescenza delle informazioni o essere soggetti a duplicazioni delle unità contenute.

Pertanto, le informazioni amministrative prima di essere utilizzate devono essere sottoposte a processi e tecniche di normalizzazione del dato per ripulirle dagli eventuali errori e rendere le basi di dati coerenti e utilizzabili a fini statistici (Fortini 2000).

Un altro limite, che non può essere corretto a posteriori, è rappresentato dall'eventuale limitata varietà di informazioni raccolte negli archivi amministrativi che può dimostrarsi non sufficiente per gli obiettivi di indagine statistica.

Nel caso della costruzione dell'archivio ALAR, le fonti amministrative utilizzate hanno presentato il vantaggio di fornire dati e informazioni in altro modo difficilmente reperibili a livello comunale e di notevole importanza per lo studio sui redditi e povertà, come la stima dei redditi familiari. E', infatti, risaputo che nelle indagini con somministrazione di un questionario, è alta la probabilità di avere mancate risposte nella rilevazione di dati sensibili, come il reddito percepito, a causa di resistenze e timori di controlli che inducono l'intervistato a non rispondere, con un conseguente effetto distorsivo sulle stime. Dall'altro lato, le fonti amministrative hanno presentato alcuni problemi di copertura e sottostima del reddito.

Entrando nello specifico delle fonti utilizzate, l'Anagrafe comunale bresciana si è dimostrata una base dati di buona qualità e piuttosto affidabile, essendo continuamente aggiornata e controllata per fini amministrativi e demografici dall'Ufficio Anagrafe del Comune di Brescia (Palamenghi, Riva, Trentini 2005). In particolare, le informazioni

identificative della persona e della famiglia, come la matricola individuale e familiare e il codice fiscale, che sono informazioni a uso interno, sono altamente affidabili dal momento che costituiscono le chiavi del database anagrafico. Normalmente hanno un buon grado di affidabilità anche le informazioni certificabili come il cognome e nome, il sesso, la data e luogo di nascita, relazione di parentela con l'intestatario della scheda di famiglia, stato civile, essendo aggiornate e verificate periodicamente (come al momento del rilascio di certificati o del rinnovo della carta d'identità). Invece, le informazioni non certificabili, come il titolo di studio e la professione, presentano una qualità variabile non tanto perché dipendono dalla dichiarazione del cittadino, ma soprattutto perché si tratta di informazioni non sottoposte a periodiche verifiche e aggiornamenti e soggette a mutamenti nel tempo (Palamenghi, Riva, Trentini 2005). Si è deciso pertanto di non utilizzare queste ultime nella costruzione dell'archivio ALAR, essendo informazioni poco affidabili e presentando un ampio margine di errore.

Per quanto riguarda la qualità dell'archivio SIATEL, costituito da un sottoinsieme di dati estratti a livello comunale e controllati dall'Agenzia delle Entrate, i modelli fiscali forniscono gli importi dei redditi (da lavoro dipendente e autonomo, le pensioni di vecchiaia, superstiti e invalidità), ma non forniscono alcuna informazione riguardo ad altre forme di reddito. Rimangono infatti esclusi (non dichiarati) quei redditi soggetti a imposta sostitutiva (redditi da capitale finanziario), quelli esenti dal pagamento Irpef (come le pensioni sociali, le pensioni di guerra, le pensioni di invalidità, le indennità di accompagnamento, le rendite per infortunio permanente), altre prestazioni sociali in denaro di natura non pensionistica, gli assegni al nucleo familiare dei dipendenti pubblici e dei pensionati, i redditi da lavoro autonomo di alcune figure professionali (agenti e rappresentanti di commercio).

Un'altra limitazione degli archivi fiscali è costituita dal fatto che i redditi denunciati nelle dichiarazioni dei redditi possono sottostimare il reddito reale effettivamente percepito dai dichiaranti, soprattutto per comportamenti illeciti di chi evade il fisco non dichiarando per intero il proprio guadagno o per il fatto di non denunciare la propria attività lavorativa (lavoro irregolare). E' ovvio che ciò può accadere soprattutto per i lavoratori autonomi che non ricevono una busta paga dal datore di lavoro e che quindi non sono controllabili, mentre i redditi da lavoratori dipendenti e pensionati sono più certi e affidabili.

Dalle analisi dei dati anagrafici e amministrativi linkati a livello familiare e provenienti dagli archivi SIATEL, è emerso, infine, un problema di copertura dei redditi dichiarati. In altre parole si è verificato che circa un 9,7% delle famiglie (corrispondente a 6,4% di individui) presenti nel registro anagrafico non possiede un reddito familiare, ossia in altri termini per queste famiglie non è presente la corrispondente dichiarazione dei redditi negli archivi fiscali (730 o UnicoPF), né il relativo CUD (770-S) rilasciato per i pensionati che non presentano la dichiarazione. Una delle principali cause di questa situazione poco realistica è da ricercarsi soprattutto nelle caratteristiche degli archivi amministrativi utilizzati. Da analisi approfondite sui dati si è appurata l'esistenza di un problema di sfasamento e aggiornamento temporale delle variabili "residenza anagrafica" e "residenza fiscale" contenute nei modelli. Si è potuto constatare, infatti, che circa un 3% delle famiglie registrate in anagrafe che appare "senza reddito dichiarato", a distanza di tre anni non è più presente nell'anagrafe comunale, a dimostrazione del fatto che queste famiglie sono emigrate in altro comune e non tempestivamente cancellate dall'anagrafe. D'altra parte, un altro 3% delle famiglie residenti a distanza di tre anni percepisce un reddito non nullo. La

spiegazione di quest'ultimo caso può essere costituita da un tardivo aggiornamento degli archivi fiscali, dall'emersione di lavoro irregolare o da eventuali correzioni di dichiarazioni errate (ad esempio si è visto che i dati presenti negli archivi 770 possono essere errati, poiché sono compilati dai datori di lavoro e non dai diretti interessati). Infine, il rimanente 3% di famiglie "senza reddito dichiarato" può essere spiegato da altri fattori, come la presenza di false denunce dei redditi e problemi di copertura degli archivi. Gli archivi fiscali, infatti, come già indicato, non riportano i redditi non dichiarati (per evasione fiscale o lavoro irregolare) e quelli esenti dal pagamento dell'Irpef (come le pensioni di anzianità, di guerra, sociali, ecc).

Purtroppo, non essendo facilmente risolvibile il problema della copertura e della mancanza di informazioni certe sui redditi che non sono stati presentati o denunciati, nelle analisi dei dati trasversali e longitudinali che seguiranno nei prossimi paragrafi, si è fatto riferimento solo alle famiglie che presentano al loro interno almeno un dichiarante, tralasciando quindi le famiglie "senza reddito dichiarato"[8].

3. La costruzione dell'archivio longitudinale amministrativo

Nel presente lavoro, come si è accennato, è stato costruito un archivio trasversale e longitudinale amministrativo sui redditi (ALAR), per gli anni 2005-2008, facendo riferimento ai dati amministrativi provenienti dall'Anagrafe del Comune di Brescia e a quelli fiscali SIATEL dell'Agenzia delle Entrate, opportunamente linkati al fine di ricostruire i redditi familiari e la povertà relativa locale.

Nel dettaglio, dall'Anagrafe comunale sono state estratte informazioni identificative sugli individui residenti e dagli archivi fiscali, invece, informazioni relative ai redditi percepiti e dichiarati dagli individui. In particolare, dal sistema SIATEL sono stati estratti e elaborati i modelli 730, 770-Semplificato[9], UNICO Persone Fisiche-UPF, contenenti informazioni relative ai redditi da lavoro dipendente, autonomo e da trasferimenti sociali.

Come punto di partenza per questo progetto, si è condotta un'analisi delle precedenti esperienze note nel campo e realizzate fino al momento dello studio (Comune 2012).

In letteratura, la maggior parte degli studi volti a stimare i redditi individuali e familiari e la povertà delle famiglie a livello locale è costituita da indagini campionarie basate sulla somministrazione di un questionario[10].

Le principali esperienze che hanno utilizzato invece dati amministrativi di origine fiscale per la stima dei redditi e povertà sono state il Progetto AMeRIcA realizzato dal Comune di Milano (Fattore, Mezzanzanica 2007), lo studio sui redditi del Comune di Brescia nell'anno 2000 (Palamenghi, Riva, Trentini 2005).

[8] Anche nell'esperienza del Progetto AMeRIcA del Comune di Milano, le analisi dei dati si sono concentrate solo su quelle famiglie che presentavano almeno un dichiarante tra i componenti la famiglia. In questo caso, le famiglie "senza reddito dichiarato" erano pari al 15% del totale. Nel Progetto è stata adottata una metodologia simile a quella adottata in questo studio. Si basava anch'essa sull'integrazione dell'Anagrafe con i dati fiscali dell'Agenzia delle Entrate (Fattore Mezzanzanica 2007).

[9] E' da distinguere il modello 770 – S presente in SIATEL da quello 770 dell'Agenzia delle Entrate. Il primo infatti è definito semplificato perché non contiene tutte le voci presenti nel 770, ma solo le principali (per approfondimenti www.agenziaentrate.gov.it).

[10] Tra questi: *Stima della povertà a livello locale: i casi della regione Toscana e delle province di Modena e Trento*, (Ballini, Betti, Lemmi, Marzadro, Morciano, Neri e Salvati 2007) e *Povertà e vulnerabilità delle coppie nel Canavese* (Negri, Solera 2007).

La scelta della metodologia da adottare per la costruzione dell'indagine ALAR ha tenuto conto dei vantaggi e degli svantaggi di ciascuna esperienza citata.

In particolare, in questo caso, si è scelto di estrarre un campione della popolazione anagrafica (metodologia adottata anche dal Comune di Brescia) al fine di snellire le procedure gestionali dei dati e di elaborazione degli stessi.

Tenendo presente che l'adozione di un campione di famiglie estratte dall'Anagrafe poteva non garantire la rappresentatività statistica delle principali variabili di interesse per le nostre analisi, si è scelto di adottare un campione di famiglie piuttosto ampio. Pertanto, in base allo studio degli errori campionari e al dettaglio di analisi che si voleva realizzare è stato estratto un campione casuale semplice[11] di 8 mila famiglie e 17 mila individui[12] per l'anno 2005.

3.1 Stima dell'*attrition* o logorio del campione

Un primo problema di carattere metodologico che è stato affrontato durante la realizzazione dell'archivio, riguarda la presenza dell'*attrition* o logorio del campione.

Ricordando che il campione estratto dall'anagrafe comunale si riferisce all'anno 2005, può verificarsi infatti il problema della perdita di rappresentatività del campione nei tempi successivi a quello di estrazione.

Come indicato, la costruzione di un archivio longitudinale presuppone un salto di qualità nell'approccio metodologico, dovuto soprattutto alla presenza di concetti del tutto nuovi.

Primo fra tutti, si deve tenere conto che la popolazione di riferimento in un'analisi di tipo longitudinale ha la caratteristica di essere *dinamica*, intendendo con questo termine una popolazione di individui soggetta a mutare nel tempo in termini di numerosità, composizione per età, sesso, cittadinanza e così via, a causa dei movimenti naturali degli individui (nascite, morti, migrazioni). L'altro concetto, simile, ma lievemente più complesso a cui si fa riferimento è quello di *famiglia longitudinale*, intendendo anche in questo caso che la famiglia nel tempo può nascere, morire, trasferirsi in un altro comune, cambiare nel tempo la sua struttura interna con l'entrata e l'uscita di componenti, originati dalla nascita, morte, trasferimenti, matrimoni, separazioni, divorzi degli individui che la compongono. Pertanto, le dinamiche temporali che avvengono a livello familiare e/o individuale, costituiscono un aspetto naturale della dinamica della popolazione, determinando flussi di entrata e di uscita di famiglie e individui *da* e *nella* popolazione di riferimento. Legato a questi concetti è il logorio del campione (*attrition*), ossia la perdita delle unità di rilevazione, come conseguenza naturale dei movimenti descritti. In altre parole, estratto un campione di famiglie al tempo t, è naturale che questo non sia più rappresentativo delle famiglie nei tempi successivi t+K, a causa dei mutamenti indicati.

Per questi motivi, prima della realizzazione dell'archivio ALAR, si è voluto verificare e quantificare la presenza di *attrition* in un caso concreto.

A titolo esemplificativo, si è fatto riferimento ai risultati dello studio sui redditi delle famiglie che il Comune di Brescia ha realizzato nell'anno 2000. In questo caso, per stimare l'*attrition* del campione iniziale, ossia per valutare la perdita di unità di rilevazione e conoscere quante sono le famiglie che sopravvivono negli anni successivi all'estrazione,

[11] Per le stime trasversali, è stato applicato il coefficiente di riporto all'universo calcolato per ciascun anno; per le stime longitudinali, si è fatto invece riferimento al coefficiente di riporto dell'anno di partenza (2005).

[12] Il campione nel lavoro del Comune di Brescia era costituito da 995 famiglie e 2.209 individui.

per ciascun anno disponibile sono state eliminate dal campione iniziale tutte quelle famiglie che sono decedute (per decesso di tutti i suoi componenti) o emigrate fuori città nell'intervallo di tempo considerato[13].

Dalle analisi, emerge che con il passar degli anni il campione iniziale perde gradualmente le unità di cui è inizialmente composto: da un contingente di 995 unità nel 2000 passa, infatti, a 735 famiglie nel 2008, perdendo circa un quarto del campione iniziale (*attrition* = 26,1%). In altre parole, è possibile ritrovare nella popolazione anagrafica di otto anni dopo solo il 73,9% delle famiglie estratte dalla popolazione anagrafica del 2000. La stessa analisi è stata ripetuta per gli individui mediante il confronto tra il campione iniziale e l'universo di riferimento[14]. Anche in questo caso, l'*attrition* del campione si attesta intorno al 25,8% (Tavola1).

Tavola 1 - Stima dell'*attrition* nel campione dello Studio del Comune di Brescia, per famiglie e individui - Anni 2000 - 2008

ANNI	Valori Assoluti	Valori percentuali
	FAMIGLIE	
2000	995	100,0
2001	962	96,7
2002	910	91,5
2003	871	87,5
2004	833	83,7
2005	792	79,6
2006	772	77,6
2007	754	75,8
2008	735	**73,9**
	INDIVIDUI	
2000	2.209	100,0
2001	2.140	96,9
2002	2.025	91,7
2003	1.932	87,5
2004	1.866	84,5
2005	1.794	81,2
2006	1.728	78,2
2007	1.680	76,1
2008	1.639	**74,2**

Fonte: Elaborazioni dell'autore su dati del Comune di Brescia

Nel passo successivo, per verificare la rappresentatività trasversale del campione, sono state analizzate le principali caratteristiche delle famiglie (tipologia familiare, numero dei componenti) e anche, degli individui (età, sesso, cittadinanza) appartenenti al campione e messe a confronto con quelle dell'intera popolazione anagrafica. In linea teorica, in assenza di *attrition*, le distribuzioni del 2000 e del 2008 dovrebbero coincidere.

Conducendo le analisi a livello familiare e, in particolare, esaminando i dati per tipologia familiare e numero di componenti, emerge che il campione non è affatto

[13] Questa analisi è stata possibile grazie alla disponibilità dei dati anagrafici del Comune di Brescia con cui confrontare il campione.

[14] Nonostante che gli individui non costituiscano le unità campione, l'elaborazione è stata utile per fornirci un'idea su quanti sono coloro che si perdono negli anni, a conclusione del quadro descritto.

rappresentativo della popolazione di riferimento. In particolare, nel campione iniziale compare una sottostima piuttosto forte per le persone sole e più lieve per i monogenitori con figli e per le altre tipologie familiari. Al contrario, risultano gonfiate le tipologie delle coppie con o senza figli. Per quanto riguarda la dimensione familiare, sono sottostimate le famiglie costituite da un solo componente, a conferma di quanto visto per la tipologia familiare, mentre le famiglie con due, tre o quattro componenti si attestano a valori superiori a quelli della popolazione di riferimento; rimangono più coerenti con la realtà anagrafica del 2008 le percentuali di famiglie con cinque o più componenti (Tavola 2).

Tavola 2 - Distribuzioni assolute e percentuali delle famiglie appartenenti al campione iniziale dello Studio del Comune di Brescia (anno 2000) e della popolazione anagrafica nel 2008, per tipologia familiare e numero dei componenti

VARIABILE	Campione Iniziale (2000)		Popolazione anagrafica 2008	
	V. assoluto	V. percentuale	V. assoluto	V. percentuale
	TIPOLOGIA FAMILIARE			
Persona sola	358	36,0	38.651	42,2
Coppia senza figli	212	21,3	15.612	16,6
Coppia con figli	289	29,0	22.069	23,5
Monogenitore con figli	100	10,0	11.348	12,1
Altro	36	3,6	3.964	4,3
	NUMERO COMPONENTI			
Uno	358	36,0	38.651	42,2
Due	288	28,9	24.345	26,6
Tre	183	18,4	14.815	16,2
Quattro	125	12,5	10.130	11,1
Cinque e più	41	4,1	3.703	4,0
Totale	**995**	**100,0**	**91.644**	**100,0**

Fonte: Elaborazioni dell'autore su dati del Comune di Brescia

Dall'analisi per individui[15], emerge che la distribuzione per età del campione iniziale è sottostimata per le fasce giovanili e sovrastimata per gli anziani (Tavola 3). Questo risultato è facilmente spiegabile considerando che il campione iniziale per costruzione non include i nati e chi è immigrato in periodi successivi all'estrazione del campione. D'altra parte, molti degli anziani presenti nel campione iniziale sono nel frattempo deceduti, determinando una forte differenza nelle frequenze relative associate all'ultima classe di età.

Il campione iniziale non presenta grosse differenze con la popolazione anagrafica del 2008 per quanto riguarda il genere, per il quale quindi rimane rappresentativo. Più grave è il confronto tra le distribuzioni degli individui per cittadinanza: la componente italiana nel campione è, infatti, fortemente sovrastimata, al contrario di quella straniera che viene sottostimata (6,6% contro il 15,6%), a dimostrazione del fatto che il campione iniziale non riflette la dinamica migratoria degli stranieri che in quegli anni è stata piuttosto

[15] Si ricorda che essendo l'unità campionaria la famiglia, si compie una piccola forzatura metodologica analizzando la rappresentatività degli individui. Essendo piuttosto marcate le differenze tra la distribuzione degli individui campione e quella della popolazione di riferimento, i risultati a cui si perviene sono da considerare piuttosto significativi.

intensa (dal 2002 al 2008 si è assistito ad un incremento della percentuale degli stranieri dal 6,0% al 15,6)[16].

In conclusione, le diverse e molteplici analisi hanno permesso di dimostrare l'inadeguatezza di un campione iniziale fermo al tempo dell'estrazione, di valutarne l'*attrition*, ossia la perdita di unità e la rappresentatività campionaria e dare anche indicazioni sulle distorsioni introdotte.

Tavola 3 - Distribuzioni assolute e percentuali degli individui appartenenti al campione iniziale dello Studio del Comune di Brescia (anno 2000) e alla popolazione anagrafica del 2008, per età, sesso e cittadinanza

VARIABILE	Campione Iniziale (2000)		Popolazione anagrafica 2008	
	V. assoluto	V. percentuale	V. assoluto	V. percentuale
		CLASSI DI ETÀ		
0-14	135	6,1	25.645	13,3
15-29	272	12,3	27.232	14,2
30-44	510	23,1	44.565	23,2
45-59	449	20,3	38.752	20,1
60-74	431	19,5	34.730	18,0
75-89	341	15,4	19.823	10,3
90 e più	71	3,2	1.695	0,9
		SESSO		
Maschio	1.065	48,2	91.595	47,6
Femmina	1.144	51,8	100.847	52,4
		CITTADINANZA		
Italiana	2.064	93,4	162.206	84,3
Straniera	145	6,6	30.236	15,6
Totale	**2.209**	**100,0**	**192.442**	**100,0**

Fonte: Elaborazioni dell'autore su dati del Comune di Brescia

(a) Si intende il campione iniziale di individui con età aumentata di 8 anni per poter fare un corretto confronto con la popolazione del 2008.

3.2 L'applicazione delle regole di inseguimento

Al fine di costruire l'archivio ALAR e per superare il problema dell'*attrition*, si è fatto riferimento ad alcune soluzioni adottate in indagini note a livello nazionale ed europeo[17], consistenti nell'applicazione di alcune regole di inseguimento. Queste presentano la caratteristica di restituire la rappresentatività trasversale del campione.

Pertanto, nelle *waves* successive all'anno di estrazione del campione (dal 2005 fino al 2008), sono stati ricostruiti i campioni trasversali di famiglie e individui applicando le seguenti regole: sono state incluse nel campione tutte le famiglie in cui figura almeno un individuo appartenente al campione originario, includendo anche i coabitanti non appartenenti al campione originario. Questi ultimi fanno parte del campione trasversale (e non longitudinale) finché conviventi con gli individui campione. Non sono state considerate le eventuali famiglie dei coabitanti nate per scissione della famiglia di origine (ad esempio,

[16] Fonte: sito http://demo.istat.it/archivio.html. E' stato considerato il 2002, poiché era il primo disponibile, dopo il Censimento 2001.

[17] Per approfondimenti si veda Comune 2012.

la famiglia nata dopo la separazione del coabitante dall'individuo campione). Sono stati inclusi i nuovi nati e eliminati i deceduti nel periodo di riferimento.

Oltre a queste, basandosi sulle indicazioni derivate dalle analisi dell'*attrition* effettuate, si è deciso di aggiungere un'altra regola. In particolare, al fine di includere nel campione le famiglie straniere immigrate nel periodo di riferimento, è stato estratto e aggiunto al campione originario un piccolo campione rappresentativo solo delle nuove famiglie entrate a far parte della popolazione tra il tempo t e t+k, ossia dopo il 2005.

Avendo a disposizione i dati sulla popolazione iniziale, ossia l'intera anagrafe comunale, è stato possibile in seguito verificare la rappresentatività del campione aggiornato con le regole precedenti.

Da queste analisi, si può notare che le regole d'inseguimento permettono complessivamente di aggiustare il campione e di renderlo rappresentativo rispetto alle principali caratteristiche individuali e familiari della popolazione, ad eccezione di qualche caso.

A livello familiare, si nota che l'applicazione di queste regole permette di rendere più simili le distribuzioni per numero di componenti delle famiglie del 2005 alla rispettiva distribuzione dell'universo delle famiglie presente in anagrafe nel 2008 (Tavola 4). Anche la tipologia familiare aggiornata si avvicina di più a quella del 2008 con qualche eccezione in cui si mostra un lieve peggioramento (modalità di risposta: altro).

Ripetendo la stessa analisi anche per gli individui, si perviene agli stessi risultati: le stime dei giovani e degli anziani e soprattutto per cittadinanza aggiornate e corrette al 2008 sono più precise e vicine al valore "vero" della popolazione di riferimento. Fanno eccezione le stime per sesso che non presentano nessun miglioramento, anche se aggiornate, e la stima della classe di età centrale 45-59 che peggiora lievemente (Tavola 5).

Tavola 4 - Distribuzioni assolute e percentuali delle famiglie del campione iniziale, aggiornato e della popolazione nel 2008, per tipologia familiare e numero dei componenti

VARIABILE	Campione Iniziale (anno 2005)	Campione Aggiornato al 2008	Popolazione anagrafica al 2008
	TIPOLOGIA FAMILIARE		
Persona sola	41,3	41,2	42,2
Coppia senza figli	17,3	16,9	16,6
Coppia con figli	25,1	24,0	23,5
Monogenitore con figli	11,7	12,3	12,1
Altro	4,6	5,0	4,3
	NUMERO COMPONENTI		
Uno	41,3	41,8	42,2
Due	27,0	26,6	26,6
Tre	16,6	16,4	16,2
Quattro	11,4	11,1	11,0
Cinque e più	3,8	4,1	4,0
Totale	**100,0**	**100,0**	**100,0**

Fonte: Elaborazioni dell'autore su dati anagrafici

Alla fine di tutto il processo di aggiornamento dei campioni trasversali con le regole di inseguimento, si nota che il numero delle famiglie campionate cresce al passare degli anni, così come gli individui campione. Dal 2006 in poi, è quasi costante la quota di nuove

famiglie che entrano a far parte dei campioni, rappresentanti delle famiglie immigrate nel comune di Brescia nel periodo corrispondente (Tavola 6).

Tavola 5 - Distribuzioni assolute e percentuali delle famiglie del campione aggiornato e della popolazione anagrafica del 2008, per età, sesso e cittadinanza

VARIABILE	Campione Iniziale (anno 2005)	Campione Aggiornato al 2008	Popolazione anagrafica al 2008
	CLASSI DI ETÀ (a)		
0-14	10,2	13,2	13,3
15-29	13,6	14,2	14,1
30-44	23,6	23,6	23,2
45-59	20,2	19,8	20,1
60-74	19,3	18,3	18,0
75-89	11,5	10,0	10,3
90 e più	1,5	0,9	0,9
	SESSO		
Maschio	47,8	47,9	47,6
Femmina	52,2	52,1	52,4
	CITTADINANZA		
Italiana	87,2	84,4	84,3
Straniera	12,8	15,6	15,6
Totale	**100,0**	**100,0**	**100,0**

Fonte: Elaborazioni dell'autore su dati anagrafici

(a) Si intende il campione iniziale di individui dell'anno 2005 con età aumentata di 3 anni per poter fare un corretto confronto con la popolazione del 2008

Tavola 6 - Schema riassuntivo dell'applicazione delle "regole di inseguimento" al campione di famiglie iniziale - Anni 2005-2008

ANNO	Numero famiglie	Numero individui campione	Individui non campioni (Coabitanti)	Nuove famiglie campionate	Nuovi individui famiglie campionate
2005	8.000	16.899	0	0	0
2006	8.096	17.088	484	329	451
2007	8.181	17.239	907	374	525
2008	8.310	17.531	1.357	365	515

Fonte: Elaborazioni dell'autore su dati anagrafici

3.3 L'abbinamento dei dati anagrafici e fiscali

Dopo aver dato origine ai quattro campioni trasversali (uno per ogni anno) aggiornati e rappresentativi della popolazione di riferimento, nella fase successiva è stato realizzato il *record-linkage* tra i dati anagrafici campionari della popolazione residente nel comune di Brescia e i dati fiscali tratti dagli archivi delle dichiarazioni dei redditi (730, UNICO Persone Fisiche-Upf e 770-S) presenti nel SIATEL, per ciascun anno (2005-2008), ottenendo il collegamento tra le informazioni anagrafiche individuali e quelle sui redditi.

L'impianto metodologico sui cui poggia il processo d'integrazione tra i dati anagrafici e quelli amministrativi prende spunto dall'esperienza dell'Istat di abbinamento tra i dati campionari dell'indagine EU-SILC e quelli fiscali provenienti dall'Agenzia delle Entrate

(Istat 2009a), in cui l'integrazione tra i dati campionari e quelli fiscali è stata utilizzata al fine di migliorare le stime sui redditi percepiti dagli intervistati.

Le operazioni compiute possono essere riassunte nel modo seguente.

Come primo passo, sono state verificate e controllate le principali variabili presenti nel campione estratto dalla popolazione anagrafica, in particolar modo i codici fiscali[18] degli individui che costituiscono la chiave di aggancio tra gli archivi anagrafici e quelli fiscali; le matricole individuali che identificano l'individuo e costituiscono la chiave di aggancio tra archivi anagrafici del Comune appartenenti ad anni diversi, utili per costruire la componente longitudinale dell'archivio; la parentela dell'individuo con il capofamiglia, utile per identificare la tipologia familiare; la data di nascita e il sesso.

Un lavoro più lungo e meticoloso è stato rappresentato dalla conversione degli archivi fiscali presenti in SIATEL da un formato compresso ("codici a serrare") a un formato direttamente leggibile da un qualunque pacchetto statistico o programma di Microsoft Office e che presenti variabili di lunghezza e posizione prefissata. Questa operazione è stata necessaria perché gli archivi fiscali sono caratterizzati da record aventi una struttura formata da una parte fissa e una variabile, dove nella parte fissa sono registrate le informazioni identificative del contribuente (nome e cognome, sesso, indirizzo data di nascita, ecc.), nella parte variabile le varie voci riguardanti i redditi dichiarati nel corrispondente modello di dichiarazione dei redditi.

Dopo aver riportato i modelli fiscali su un database relazionale idoneo per realizzare le opportune elaborazioni dei dati, si è aperta una lunga fase di verifica, controllo e preparazione dei dati per renderli idonei all'abbinamento tra gli archivi. In particolare, si dovevano trasformare gli archivi in modo tale che le informazioni di ciascun individuo fossero sintetizzate in un solo record. A tal fine, si è verificato se le informazioni individuali, soprattutto quelle relative alle principali variabili di reddito, presenti in ciascun modello fiscale, fossero in parte o in tutto replicate. In seguito, nel caso di presenza di più informazioni relative allo stesso individuo, queste dovevano essere riportate a livello individuale.

Si è scelto di procedere nel seguente modo, distintamente per ciascun tipo di modello:

1. Nei modelli 770-S, dopo aver eliminato i records doppioni (i percipienti del reddito che presentavano le stesse informazioni relative sia al sostituto d'imposta sia all'ammontare del reddito percepito), nei casi di presenza di più di una dichiarazione rilasciata dai sostituti d'imposta relative ai diversi lavori/pensioni svolti dallo stesso individuo, le informazioni sono state riportate a livello individuale aggregando, per somma, i redditi percepiti;
2. Nei modelli 730 o UNICO Persone Fisiche (UPF), dopo aver eliminato i doppioni (che presentavano gli stessi valori del reddito lordo percepito e le imposte pagate), nel caso di contemporanea presenza di due o più dichiarazioni dei redditi presentati dallo stesso individuo, si è provveduto ad accertare quale dichiarazione fosse la più aggiornata, in base ad alcuni campi presenti negli archivi che indicano lo stato di lavorazione e aggiornamento della dichiarazione (flag di integrazione e correzione della dichiarazione). Nei pochi casi residui, in assenza di indicazioni sullo stato di aggiornamento o su quale versione fosse la più veritiera, si è ricorso a effettuare

[18] I codici fiscali degli individui presenti nell'Anagrafe comunale di Brescia hanno una buona qualità, poiché l'ufficio competente provvede a verificarne la loro esattezza.

una media dei redditi percepiti e dichiarati nelle duplici dichiarazioni, consolidando le informazioni relative allo stesso individuo su un solo record;

3. Nei casi in cui il contribuente aveva presentato la denuncia dei redditi sia attraverso il modello fiscale 730, sia attraverso il modello Unico Persone Fisiche, (una decina di casi per ciascun anno considerato), si è proceduto a eliminare le dichiarazioni presentate attraverso il modello 730, dato che per motivi temporali (la scadenza del 730 è precedente a quella dell'Unico) la dichiarazione più aggiornata è quella fornita con l'UnicoPF.

Alla fine di questa procedura, ossia dopo aver controllato e sistemato gli archivi, si è proceduto all'abbinamento tra il campione anagrafico e gli archivi dei redditi percepiti da ciascun individuo rilevabili presso le varie fonti (730, 770-S e UPF), mediante la tecnica "exact matching" che permette di abbinare informazioni relative alla stessa unità statistica da distinte fonti del dato. L'abbinamento usato è del tipo uno a uno, in base al quale a ogni unità statistica corrisponde un solo record in ognuna delle fonti da integrare. L'abbinamento è stato realizzato tramite l'utilizzo di una "chiave di abbinamento" che identifica univocamente l'unità statistica comune ai diversi archivi. Nel nostro caso la chiave è rappresentata dal codice fiscale dell'individuo[19].

Tavola 7 - Individui del campione per modello di dichiarazione e abbinamento con dati fiscali - Anno 2005 *(valori assoluti e percentuali)*

INDIVIDUI PER MODELLO DI DICHIARAZIONE	Link campione con dati fiscali	
	Valori assoluti	Valori percentuali
Modello 730	4.641	40,8
Modello 770 - S (ridotto)	3.973	35,0
Modello UNICO Persone Fisiche	2.572	22,6
Totale dichiaranti linkati	11.366	100,0
Dichiaranti linkati con età>=15 anni	11.366	77,1
Individui non linkati con età >=15 anni	3.373	22,9
Totale individui campione con età>=15 anni	14.739	100,0
Individui con età>=15 anni	14.739	87,2
Individui con età<15 anni	2.160	12,8
Totale individui campione	**16.899**	**100,0**

Fonte: Elaborazioni dell'autore su dati anagrafici e fiscali

In sintesi, tramite la chiave del codice fiscale sono stati abbinati il 77,1% degli individui con età maggiore di 15 anni, rimanendo un 22,9% di individui che non hanno presentato dichiarazione dei redditi perché non percipienti alcun reddito (Tavola 7). Rientrano in questa percentuale sia chi non percepisce alcun reddito, come gli studenti, le casalinghe, i disoccupati, sia chi percepisce un reddito esente dal pagamento dell'Irpef (come le pensioni sociali, pensioni di guerra, pensioni di invalidità, indennità di accompagnamento, rendite

[19] Al fine di includere anche gli individui che presentavano codici fiscali diversi, ma che corrispondevano alle stesse persone, sono stati abbinati i record degli archivi cambiando la chiave di aggancio, ossia prendendo, contemporaneamente, come riferimento il cognome e nome, il sesso e la data di nascita. Così facendo, si sono recuperati solo pochi casi (rispettivamente per ciascun anno: 17, 61, 99 e 123 unità). Dopo aver effettuato l'aggancio con i dati amministrativi, il codice fiscale e altre informazioni identificative (quali il nome e cognome) sono stati eliminati dalle basi di dati integrate, in virtù delle disposizioni in materia di protezione dei dati personali (d.lgs. n. 196/2003).

per infortunio permanente). E' possibile che all'interno di questo gruppo ci sia una quota non definibile di persone che evadono il fisco o che svolgono attività lavorative non dichiarate ("irregolari").

3.4 La struttura dell'archivio ALAR

Alla fine del processo di costruzione dell'archivio (estrazione del campione, aggiornamento del campione per ciascun anno con le regole di inseguimento e *record-linkage* con i dati fiscali), per ciascun anno si è ottenuto un file contenente l'elenco di tutti gli individui (uno per ogni riga) con le informazioni anagrafiche e fiscali.

In particolare, nell'archivio sono presenti le informazioni identificative relative agli individui registrati nell'Anagrafe comunale (nome e cognome, codice fiscale, sesso, data di nascita, indirizzo, matricola familiare, stato civile, parentela con il capofamiglia) e quelle relative ai redditi percepiti e dichiarati dagli individui provenienti dai modelli fiscali utilizzati (730, 770-Semplificato, Unico Persone Fisiche-UPF).

Scendendo nel dettaglio delle voci fiscali estratte dai modelli fiscali e presenti nell'archivio, provengono dai modelli 730 le caratteristiche dei dichiaranti (e del coniuge convivente, se la dichiarazione è congiunta), tra cui nome e cognome, codice fiscale, sesso, data di nascita, città natale, di residenza e fiscale; i vari tipi di reddito, da quello dei terreni (agrario e dominicale), a quello dei fabbricati, da lavoro dipendente e assimilato, al reddito imponibile per il calcolo dell'Irpef, all'imposta Irpef netta, al reddito imponibile per il calcolo dell'addizionale comunale, all'addizionale comunale all'Irpef dovuta e certificata (Agenzia delle Entrate 2006b).

Del modello Upf, figurano invece i redditi da lavoro autonomo composti dai più voci: redditi derivanti dall'esercizio delle arti e professioni; redditi dei titolari d'impresa in contabilità ordinaria e semplificata; redditi di partecipazione in società di persone e assimilate; redditi di allevamenti, compensi derivanti da attività di lavoro autonomo, anche se svolte all'estero, non esercitate abitualmente (Agenzia delle Entrate 2006c). La somma di questi redditi al netto delle ritenute d'acconto versate fornisce la stima del reddito netto da lavoro autonomo (Istat 2009a).

Tra le informazioni contenute nei modelli 770 - S presenti in SIATEL, sono state estratte le informazioni identificative del sostituto d'imposta (la matricola e la denominazione), quelle del contribuente cui sono riferite le dichiarazioni rilasciate dal datore di lavoro, il reddito percepito (al netto delle ritenute contributive), il reddito imponibile, le ritenute Irpef, le deduzioni di progressività d'imposta, le deduzioni per coniuge e familiari a carico, le detrazioni per oneri, gli oneri sostenuti, le addizionali comunali Irpef, i redditi conguagliati da altri sostituti d'imposta (Agenzia delle Entrate 2006a).

In sintesi, i modelli 730 e Unico Persone Fisiche (UPF) hanno fornito informazioni a livello individuale sui redditi da lavoro, redditi da pensione, redditi da fabbricati, da terreni e i redditi da lavoro autonomo non derivanti da attività professionale; dai modelli UPF, oltre alle voci precedenti valide per il modello 730, sono stati recuperati anche i redditi derivanti da attività autonoma e professionale. Dalle certificazioni CUD presenti nei modelli 770-S sono state ricavate invece le informazioni, rilasciate dai sostituti d'imposta, relative ai redditi da lavoro e da pensione di chi non ha presentato la dichiarazione dei redditi con il modello 730 o Upf, perché esentati a farlo.

Nell'archivio così costruito, sono state poi aggiunte ed elaborate altre variabili come le fasce di età dell'individuo, i coefficienti di riporto, la stima dell'addizionale Irpef

regionale[20], il reddito lordo e netto[21], a prezzi correnti e costanti, percepito da ciascun individuo (Tavola 8).

Tavola 8 - Struttura del file individuale dell'archivio ALAR

DATI IDENTIFICATIVI INDIVIDUALI (MATRICOLA INDIVIDUALE, MATRICOLA FAMILIARE)	Sezione dati anagrafici (età, data e luogo di nascita, sesso, relazione di parentela capofamiglia)	Variabili di utilità (coefficienti di riporto all'universo trasversale e longitudinale, coefficiente di equivalenza Ocse)	Variabili reddito (fonte 730)	Variabili reddito (fonte UnicoPF)	Variabili reddito (fonte 770)	Calcolo reddito lordo	Stima Irpef regionale	Calcolo reddito netto, a prezzi correnti e a prezzi costanti
00001								
00002								
00003								
……..								

In base alla matricola familiare degli individui, è stato possibile ricostruire la composizione di ciascuna famiglia e ottenere i files familiari, uno per ciascun anno, contenenti le informazioni della famiglia di appartenenza. Per ciascuno di essi, sono state costruite le seguenti variabili: la tipologia familiare (persona sola, coppia sola, coppia con figli, monogenitore con figli, altro), il numero di componenti, il numero di anziani in famiglia, il numero di percettori di reddito, il numero di disoccupati e i relativi redditi familiari dati per somma dei redditi degli individui appartenenti alla stessa famiglia (lordo, netto, equivalente, a prezzi correnti e costanti -Tavola 9).

In sintesi, si sono ottenuti quattro file individuali e quattro familiari, uno per ciascun anno, utilizzati per ottenere le stime trasversali.

Tavola 9 - Struttura del file familiare dell'archivio ALAR

DATI IDENTIFICATIVI FAMIGLIA (MATRICOLA FAMILIARE)	Variabili ricostruite (tipologia familiare, numero componenti, numero figli, numero anziani in famiglia, numero di percettori in famiglia, numero di disoccupati)	Variabili di utilità (coefficienti di riporto all'universo trasversale e longitudinale, coefficiente di equivalenza Ocse)	Reddito lordo, netto, a prezzi correnti e costanti	Reddito lordo e netto equivalente familiare, a prezzi correnti e costanti
00001				
00002				
00003				
……..				

[20] Si è stimata l'addizionale regionale all'Irpef in base alla conoscenza delle aliquote applicate dalla Regione Lombardia ai redditi imponibili per gli anni di riferimento, non essendo questa voce disponibile nell'archivio SIATEL. L'imposta addizionale comunale all'Irpef, anch'essa mancante negli archivi fiscali, invece, non è stata calcolata, poiché fino all'anno 2011 nel comune di Brescia non era dovuta.

[21] Il reddito netto è stato ottenuto come differenza tra quello percepito in busta paga (definito "lordo al netto dei contributi previdenziali") e le imposte pagate (Irpef, addizionali regionali e comunali all'Irpef).

Tavola 10 - Struttura del file longitudinale e individuale dell'archivio ALAR

DATI IDENTIFICATIVI INDIVIDUALI (MATRICOLA INDIVIDUALE, MATRICOLA FAMILIARE)	Sezione dati anagrafici (età, data e luogo di nascita, sesso, relazione di parentela capofamiglia)	Sezione dati anagrafici (età, data e luogo di nascita, sesso, relazione di parentela capofamiglia)	Sezione dati anagrafici (età, data e luogo di nascita, sesso, relazione di parentela capofamiglia)	Sezione dati anagrafici (età, data e luogo di nascita, sesso, relazione di parentela capofamiglia)	Variabili reddito	Variabili reddito	Variabili reddito	Variabili reddito
	Anno 2005	Anno 2006	Anno 2007	Anno 2008	Anno 2005	Anno 2006	Anno 2007	Anno 2008
00001								
00002								
00003								
…….								

Per ottenere le stime longitudinali è stato creato un file longitudinale individuale contenente tutti i 17 mila individui iniziali e quelli che successivamente sono entrati a far parte di uno dei campioni trasversali (coabitanti e nuovi estratti). La stessa cosa è stata fatta a livello familiare (le 8 mila famiglie iniziali più quelle nuove). In entrambi i casi sono state riportate le principali variabili di reddito relative a ciascun individuo o famiglia (Tavola 10 e 11). E' da sottolineare che per le analisi longitudinali si è scelto di considerare facenti parte del campione longitudinale, solo le famiglie e gli individui presenti in tutte e quattro le *waves*, rispettivamente pari a circa 6.900 e a 14.900 unità (dal 2005 al 2008), escludendo quindi tutti coloro che sono usciti dal campione negli anni successivi al 2005, emigrando in altro comune. Ciò è stato fatto per rendere più omogenee le analisi longitudinali.

Nei due file longitudinali (uno degli individui e uno delle famiglie) sono state inserite in modo sequenziale per ciascun anno le informazioni dei soggetti campioni ritenute più utili e interessanti ai fini delle analisi dei dati (dalle caratteristiche individuali, familiari a quelle sui redditi).

Tavola 11 - Struttura del file longitudinale familiare dell'archivio ALAR

DATI IDENTIFICATIVI FAMIGLIA (MATRICOLA FAMILIARE)	Variabili ricostruite (tipologia familiare, numero componenti, numero figli, ecc.)	Variabili ricostruite (tipologia familiare, numero componenti, numero figli, ecc.)	Variabili ricostruite (tipologia familiare, numero componenti, numero figli, ecc.)	Variabili ricostruite (tipologia familiare, numero componenti, numero figli, ecc.)	Variabili reddito	Variabili reddito	Variabili reddito	Variabili reddito
	Anno 2005	Anno 2006	Anno 2007	Anno 2008	Anno 2005	Anno 2006	Anno 2007	Anno 2008
00001								
00002								
00003								
……..								

4. Definizioni utilizzate per le stime del reddito e della povertà

Per le analisi trasversali e longitudinali che sono presentate in questo contributo, sono state scelte in modo accurato e appropriato le definizioni cui fare riferimento per la stima del reddito e della povertà.

Partendo dalle varie voci disponibili negli archivi fiscali (reddito lordo, imponibile, netto, con o senza fabbricati[22]) è stato scelto di utilizzare il reddito netto, tenendo conto dei risultati di precedenti esperienze note in letteratura (Palamenghi, Riva, Trentini 2005; Fattore, Mezzanzanica 2007) e degli obiettivi di questo studio. Volendo stimare l'effettiva disponibilità monetaria degli individui e delle famiglie con le relative soglie di povertà, il reddito che meglio soddisfa questa esigenza è quello netto, stimato come differenza tra la retribuzione lorda percepita da ciascun contribuente, i contributi sociali versati all'ente previdenziale e le imposte dovute dal dichiarante (le ritenute Irpef, l'addizionale regionale all'Irpef e l'addizionale comunale all'Irpef).

Infine, come si è già precisato, essendo la fonte dei dati fiscali basata sulle dichiarazioni dei redditi presentate dai contribuenti all'Agenzia delle Entrate, il reddito stimato disponibile corrisponde concettualmente a un reddito "dichiarato", con tutti i limiti di affidabilità che ne conseguono.

Per quanto riguarda invece le definizioni adottate per le stime di povertà, l'approccio teorico di riferimento adottato in questo studio è stato quello unidimensionale, basato su un unico indicatore, il reddito. In altre parole, si è fatto riferimento alla povertà economica, ossia ci si è concentrati solo sull'aspetto economico della povertà, non tenendo conto di altri aspetti che possono incidere sulla qualità della vita e sul benessere, generalmente compresi negli approcci di povertà multidimensionale (Chiappero Martinetti 2006). Inoltre, ci si è basati su una misura di povertà relativa, in cui si tiene conto dello standard di benessere raggiunto dalla popolazione nel suo complesso.

La definizione ufficiale di povertà, utilizzata dall'Istat, si basa sull'uso di una linea di povertà nota come *International Standard of Poverty Line* (Ispl) che definisce povera una famiglia di due componenti con una spesa per consumi inferiore o pari alla spesa media per consumi pro-capite (Istat 2010a). La linea di povertà relativa individua pertanto il livello di spesa per consumi che rappresenta il limite di demarcazione tra famiglie povere e non povere ed è calcolata a livello nazionale.

In questo studio, invece, è stata applicata la definizione utilizzata in ambito europeo, "*at risk of poverty rate*", in cui è "povero" chi ha un reddito familiare *equivalente* inferiore a una soglia di povertà pari al 60% del reddito familiare mediano *equivalente nazionale*. Di solito, si ricorre all'utilizzo del reddito *equivalente* per effettuare confronti corretti tra famiglie con diversa numerosità familiare. In termini pratici, si applicano al reddito familiare delle appropriate scale di equivalenza, costituite da un insieme di coefficienti correttivi, in modo da ottenere un reddito che tenga conto delle economie di scala realizzabili all'aumentare dell'ampiezza familiare. Il concetto di scala di equivalenza è strettamente connesso con il concetto di povertà adottato e con la variabile di riferimento:

[22] Il reddito da fabbricato non è stato considerato poiché questa informazione non è disponibile per tutti i contribuenti, ma solo per quelli che hanno presentato la dichiarazione dei redditi tramite il modello 730 o Unico Persone Fisiche. Le certificazioni rilasciate dai datori di lavoro (modelli 770), infatti, non riportano le rendite da fabbricato poiché non è tra i dati in loro possesso. Sarebbe stato utile tenerne conto al fine di valutare l'incidenza del possesso dell'abitazione nella formazione del reddito e della ricchezza familiare, supponendo che gli immobili costituiscano una fonte di reddito, se affittati, o comunque una fonte di risparmio se utilizzati dagli stessi proprietari.

solitamente per la stima della povertà attraverso la spesa per consumi familiare è utilizzata la scala Carbonaro, mentre per quella tramite il reddito, come in questo studio, si usa quella Ocse modificata[23].

Inoltre, in questo lavoro, per stimare la povertà delle famiglie residenti nel comune di Brescia è stata adottata, anziché una soglia di povertà nazionale, una linea di povertà "*locale"*, basata cioè sul reddito delle famiglie residenti nel comune e riferita pertanto al livello di benessere della popolazione residente all'interno dei confini comunali. Questo approccio è stato adottato in virtù dei molteplici studi che recentemente hanno messo in evidenza i limiti dell'adozione di un'unica soglia nazionale per stimare la povertà relativa a livelli sub-nazionali. Fino ad oggi, infatti, la statistica ufficiale ha utilizzato la soglia di povertà relativa nazionale per fornire stime di povertà nazionali, ripartizionali o al massimo regionali (mai comunali). Negli ultimi anni, diversi studi in letteratura hanno dimostrato che la misura della povertà necessita di strumenti di misurazione diversificati territorialmente. Il problema nasce dal fatto che la povertà relativa, per definizione, è riferita al contesto in cui si trova, di cui è espressione, mentre la soglia di povertà utilizzata per la stima locale è di solito costruita in base al benessere nazionale e non a quello locale. E' stato dimostrato che l'utilizzo della soglia nazionale per la stima di povertà relative a livelli sub-nazionali tende a sovrastimare i livelli di povertà relativi delle regioni del Mezzogiorno e a sottostimare quelli del Centro Nord[24].

L'utilizzo di linee di povertà differenziate per livello territoriale, in realtà non è una novità, poiché queste sono già da tempo utilizzate in ambito europeo: la diffusione della povertà dei singoli Stati Membri è infatti stimata attraverso l'impiego di tante linee di povertà nazionali quanti sono i Paesi dell'UE (Chiappero Martinetti, Bottiroli Civardi 2002). Anche nella stima della povertà assoluta ufficiale si utilizzano soglie di povertà che variano per ripartizione geografica, tipologia familiare e dimensione del comune (Istat 2009b).

Come per ogni altra metodologia statistica, nella sua applicazione ci si deve ricordare delle modalità nel suo utilizzo: l'adozione di un'unica linea di povertà nazionale ha il vantaggio di permettere i raffronti tra territori, ma lo svantaggio di non far emergere le disparità presenti in ciascuno di essi; d'altra parte, con l'adozione di una linea di povertà relativa al contesto locale si misura la diseguaglianza specifica presente all'interno di quel territorio, ma si ottengono risultati ovviamente non comparabili con le altre realtà locali (Freguja, Pannuzi 2007).

[23] La scala di equivalenza Ocse modificata prevede per ciascuna famiglia un coefficiente correttivo pari alla somma di più coefficienti individuali: 1 per il primo componente adulto, 0,5 per ogni altro adulto e 0,3 per ogni minore di 14 anni.

[24] La ricerca di Accolla (2009) dimostra che per la regione Lombardia nel 2007 l'incidenza di povertà relativa basata sui consumi delle famiglie calcolata ufficialmente dall'ISTAT con la soglia nazionale è pari al 4,8%, mentre adottando una soglia a livello regionale sale al 10,0%. Agli stessi risultati si perviene anche stimando la povertà mediante i redditi: lo studio di Mori (2007), dimostra che la povertà relativa, calcolata utilizzando dati Banca d'Italia e una soglia nazionale pari al 50% del reddito mediano, al Nord nel 2004 si attesta attorno al 4,6%, mentre con una linea regionale sale al 7,7%. Anche nello studio di Chiappero Martinetti-Vega Pansini (2009), calcolando il tasso di rischio di povertà mediante il reddito e in base ai dati dell'indagine EU-SILC, è stato dimostrato che l'utilizzo di una linea sub-nazionale (pari al 60% del reddito mediano) determina un aumento della stima dell'incidenza di povertà nei territori considerati (per la Lombardia 16,0% contro 11,4% della linea nazionale).

5. Analisi preliminari sui dati dell'archivio ALAR

Le prime analisi condotte sui dati dell'archivio ALAR realizzato hanno fornito stime sui redditi percepiti dalle famiglie bresciane, in una prospettiva trasversale, delineando un quadro di riferimento su quanto guadagnano le famiglie bresciane in ciascun anno considerato e quanto sia variato il loro potere d'acquisto nell'arco di tempo osservato.

Nel 2005, il reddito netto medio familiare si è attestato a 25,3 mila euro (pari a circa 2 mila euro al mese), mentre quello mediano a 19,3 mila euro (pari a circa 1.600 euro al mese), con una diminuzione nel periodo considerato e in termini reali dell'0,2% (Tavola 12).

Le analisi sui redditi familiari condotte sulle stesse famiglie seguite nel tempo, in una prospettiva longitudinale, ci indicano una situazione lievemente peggiore di quella emersa con l'analisi *cross-section*, dato che nell'arco temporale di riferimento (dal 2005 al 2008) si è verificata una perdita del potere d'acquisto delle famiglie, pari allo 0,9% per il reddito netto familiare mediano. Emerge, inoltre, che le famiglie hanno registrato forti diminuzioni del loro reddito in valori reali soprattutto nel biennio 2007-2008, a seguito della crisi finanziaria (Tavola 13).

Tavola 12 - Valori medi e mediani annui di reddito netto delle famiglie dichiaranti e residenti nel comune di Brescia *(migliaia di euro)* **- Analisi trasversale - Anni 2005 - 2008**

ANNI	2005	2006	2007	2008	Var. perc. 2005/2008
			REDDITO FAMILIARE NETTO		
Medio	25,3	26,4	27,0	27,2	7,6
Mediano	19,3	19,8	20,3	20,5	6,3
			REDDITO FAMILIARE NETTO A PREZZI COSTANTI (2005)		
Medio	25,3	25,9	26,3	25,5	1,0
Mediano	19,3	19,4	19,8	19,2	-0,2

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

Tavola 13 - Valori medi e mediani annui di reddito netto delle famiglie sempre presenti (a) **e residenti nel comune di Brescia nel periodo indicato - Analisi longitudinale - Anni 2005 - 2008**

ANNI	2005	2006	2007	2008	Var. perc. 2005/2008
			REDDITO FAMILIARE NETTO		
Medio	24,8	26,0	26,4	26,7	7,8
Mediano	19,1	19,7	20,1	20,2	5,5
			REDDITO FAMILIARE NETTO A PREZZI COSTANTI (2005)		
Medio	24,8	25,5	25,5	25,1	1,2
Mediano	19,1	19,4	19,4	19,0	-0,9

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

(a) Per omogeneità del dato, nelle analisi longitudinali, si è fatto riferimento al sottocampione di famiglie sempre presenti nei quattro anni considerati (6.900 unità)

Distinguendo per cittadinanza, emerge una forte differenza tra i redditi percepiti dagli italiani e quelli percepiti dagli stranieri, essendo il reddito netto equivalente dei primi quasi il doppio di quello dei secondi (rispettivamente 15,7 mila euro e 8,7 mila euro).

Lo studio sui redditi differenziati per le principali caratteristiche familiari ha consentito di individuare le categorie più vulnerabili, ossia quelle che hanno un maggiore rischio di cadere in povertà, a causa dei bassi redditi percepiti. Dai risultati è emerso che, tra le famiglie italiane, quelle che soffrono un evidente disagio economico sono costituite principalmente da quelle in cui sono presenti anziani, disoccupati o un solo percettore di reddito; i monogenitori con figli e le persone sole percepiscono redditi lievemente sotto la media, ma non rappresentano casi preoccupanti. Gravi, invece, le situazioni delle famiglie straniere numerose, con tre o più figli, soprattutto se minori, e quelle con due disoccupati. Migliore è la situazione delle famiglie straniere, costituite da coppie senza figli o persone sole o con almeno due percettori di reddito.

In sostanza, per le famiglie italiane i fattori critici sono rappresentati dall'essere anziano e dalla mancanza di lavoro, mentre per le famiglie straniere dalla presenza di famiglie numerose, dall'avere figli, soprattutto minori e, ovviamente dall'essere disoccupate.

Passando a una misura dell'incidenza di povertà relativa delle famiglie[25], è emerso che circa oltre un quinto delle famiglie con almeno un dichiarante (21,2% per l'anno 2005[26] e 21,6% per il 2008) si trova in difficoltà economica relativamente al livello standard di benessere della popolazione locale (Tavola 14).

Tavola 14 - Linee di povertà e incidenza di povertà delle famiglie dichiaranti e residenti nel comune di Brescia - Anni 2005-2008

ANNI	2005	2006	2007	2008
	FAMIGLIE			
Reddito netto mediano familiare equivalente	13.842	14.267	14.638	14.863
Linea di povertà standard	8.305	8.560	8.782	8.917
Non povere	*78,8*	*79,3*	*78,9*	*78,4*
Sicuramente non povere	70,5	70,7	70,9	69,6
Quasi povere	8,3	8,6	8,0	8,8
Povere	***21,2***	***20,7***	***21,1***	***21,6***
Appena povere	7,2	7,5	7,4	7,7
Sicuramente povere	14,0	13,3	13,6	13,9

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

[25] L'indicatore incidenza di povertà è definito come il rapporto tra il numero di famiglie (individui) in condizione di povertà e il numero di famiglie (individui) residenti.

[26] Si fa presente che l'errore relativo campionario associato all'incidenza di povertà per l'anno 2005 è del 2,1%. Pertanto la stima ha un range che va dal 20,0% al 22,3%.

Tavola 15 - Linee di povertà e incidenza di povertà degli individui dichiaranti e residenti nel comune di Brescia - Anni 2005-2008

ANNI	2005	2006	2007	2008
	INDIVIDUI			
Reddito netto mediano familiare equivalente	14.327	14.780	15.145	15.188
Linea di povertà standard	8.596	8.868	9.087	9.113
Non poveri	*78,6*	*79,1*	*78,5*	*77,8*
Sicuramente non poveri	70,2	70,1	70,0	69,2
Quasi poveri	8,4	9,0	8,5	8,6
Poveri	***21,4***	***20,9***	***21,5***	***22,2***
Appena poveri	7,7	7,4	7,8	7,4
Sicuramente poveri	13,7	13,5	13,7	14,8

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

Complessivamente, per il comune di Brescia, emerge la seguente situazione economica: nel 2005, più del 70% delle famiglie con almeno un dichiarante è collocato tra i "sicuramente non poveri", cui si aggiunge un 8,3% di "quasi poveri", che corrispondono alle situazioni familiari lievemente sopra la soglia di povertà e che facilmente potrebbero scivolare nella categoria inferiore. Quelle "povere" sono pari al 21,2%, distinte in "appena povere" (7,2%) e "sicuramente povere" (14,0%)[27].

Si perviene a risultati lievemente peggiori per l'incidenza di povertà relativa sugli individui, anche se con lo stesso andamento temporale (Tavola 15). In entrambi i casi è nettamente evidente il peggioramento delle condizioni economiche con l'avvento della crisi nel 2008.

Da un'analisi delle principali caratteristiche socio-demografiche, emerge che lo stato di povertà/benessere degli individui bresciani è legato alla cittadinanza: la presenza straniera è più elevata tra i "sicuramente poveri" (33%), mentre è minima tra i "sicuramente non poveri" (5%). Un'altra lieve differenza tra le classi più povere e quelle meno povere è costituita da una maggiore presenza femminile tra le famiglie più povere: le donne sono più presenti tra le fasce a rischio povertà o "appena povere" (rispettivamente 55,3%, 55,7%) e meno presenti tra i "sicuramente non poveri" (51%). Per quanto riguarda l'età media, pur non essendoci grosse differenze, si evidenzia che i più giovani sono anche i più poveri, spiegabile con il fatto che chi ha un'età oltre la media ha raggiunto una maggiore stabilità economica e che gli stranieri che sono i più poveri sono anche più giovani degli italiani. Le analisi, in ultimo, confermano che le condizioni di benessere/povertà delle famiglie e individui, sono legate alla presenza e alla numerosità di percettori di reddito in famiglia: all'aumentare dei percettori di reddito, infatti, migliorano le condizioni economiche delle famiglie andando a collocarsi nelle categorie superiori e viceversa.

L'aspetto più interessante delle analisi longitudinali dei redditi percepiti è rappresentato dai flussi di famiglie e individui tra condizioni di benessere e povertà

[27] Qui si fa riferimento alla classificazione ampliata delle categorie di povertà inserite negli ultimi anni nelle pubblicazioni ufficiali dell'Istat, utili a superare la netta dicotomizzazione tra poveri e non poveri (due soglie di povertà pari al 120% e all'80% di quella standard, Istat 2011).

riportati nella matrice di transizione (Tavola 16). Da queste analisi emerge che, fatta eccezione per le "sicuramente non povere" che presentano una situazione molto stabile di benessere (più del 92% a distanza di tre anni è ancora collocato in quella classe!), per il resto, le altre categorie presentano situazioni fluttuanti nel tempo, dimostrando una natura piuttosto transitoria delle classi più prossime alla linea di povertà.

Per le famiglie "quasi povere", ossia per quelle che si collocano lievemente sopra la linea di povertà relativa, la metà rimane nella stessa classe, un quarto peggiora le proprie condizioni economiche scivolando nelle classi sotto la linea di povertà e un altro quarto migliora andando a far parte delle "sicuramente non povere" (26,7%).

Tavola 16 - Matrice di transizione tra condizioni di povertà e benessere delle famiglie sempre presenti (a) **e residenti tra gli anni 2005 e 2008 - Analisi longitudinale** *(Percentuali di riga)*

		Anno 2008				*Valori assoluti 2005*
	Condizione di povertà/benessere	Sicuramente Non poveri	Quasi poveri	Appena poveri	Sicuramente poveri	
Anno 2005	Sicuramente Non poveri	**92,4**	3,6	1,7	2,4	4.591
	Quasi poveri	26,7	**50,8**	13,4	9,1	528
	Appena poveri	20,4	8,8	**52,4**	18,3	431
	Sicuramente poveri	14,3	8,1	9,1	**68,5**	718
Valori assoluti 2008		4.572	527	440	729	6.268

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

(a) Per omogeneità del dato, nelle analisi longitudinali, si è fatto riferimento al sottocampione di famiglie sempre presenti nei quattro anni considerati e con dati validi (escluse le famiglie "senza reddito dichiarato").

Per quanto riguarda le famiglie "appena povere", il 52,4% di loro rimane nella stessa classe, meno di un terzo riesce a collocarsi nella classe di ordine superiore, ossia migliora nettamente le condizioni economiche e la rimanente parte (18,3%) peggiora notevolmente, finendo nella classe delle famiglie "sicuramente povere".

Infine, più grave la situazione delle famiglie "sicuramente povere", poiché più del 68% delle famiglie classificate in questa classe è ancora sicuramente povero a distanza di tre anni. E' comunque sorprendente che del restante 32% delle famiglie più povere, circa il 15% riesce a fare il salto di qualità e andare a collocarsi nella classe dei "sicuramente non poveri".

Per gli individui appartenenti ai "sicuramente non poveri", i risultati si presentano simili a quelli presentati per le famiglie (Tavola 17). Risultano più facili le transizioni verso le condizioni migliori: sembra che per gli individui sia più facile rispetto alle famiglie uscire dalla condizione di povertà, poiché a distanza di tre anni il 36,1% ha migliorato le proprie condizioni contro il 31,5% familiare. Più ampie anche le transizioni dalle condizioni vicine alla povertà (appena o quasi poveri) verso le altre categorie (in entrambi i casi solo poco più del 45% rimane nella stessa categoria dopo tre anni).

Tavola 17 - Matrice di transizione tra condizioni di povertà e benessere degli individui sempre presenti (a) **e residenti tra gli anni 2005 e 2008 - Analisi longitudinale** *(Percentuali di riga)*

		Anno 2008				*Valori assoluti 2005*
	Condizione di povertà/benessere	Sicuramente Non poveri	Quasi poveri	Appena poveri	Sicuramente poveri	
Anno 2005	Sicuramente Non poveri	**92,5**	3,3	1,8	2,4	10.035
	Quasi poveri	31,2	**45,8**	12,1	10,8	1.194
	Appena poveri	21,4	12,4	**46,9**	19,4	1.016
	Sicuramente poveri	15,7	10,2	10,2	**63,9**	1.642
Valori assoluti 2008		10.126	1.170	971	1.620	13.885

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

(a) Per omogeneità del dato, nelle analisi longitudinali, si è fatto riferimento al sottocampione di famiglie sempre presenti nei quattro anni considerati e con dati validi (esclusi gli individui che appartengono a famiglie "senza reddito dichiarato").

Approfondendo le analisi sui flussi mediante le matrici di transizione annuali, si evidenzia che tra un anno e l'altro è presente una continua diminuzione delle percentuali delle famiglie "sicuramente non povere" che rimangono in quella stessa classe (dal 2005 al 2006 tale percentuale è del 95,8%, tra il 2006 e il 2007 scende lievemente al 95,4% fino ad arrivare tra il 2007 e il 2008 al 94,8%). Ancora più preoccupante è la quota delle famiglie "sicuramente povere" che soprattutto nell'ultimo biennio è rimasta intrappolata in questa condizione (dal 2005 al 2006 tale percentuale è del 75,4%, tra il 2006 e il 2007 è salita al 79,5% fino a toccare tra il 2007 e il 2008 l'81,5%). In entrambi i casi, i risultati mettono in evidenzia un progressivo e continuo impoverimento delle famiglie. Alle stesse conclusioni si arriva analizzando le matrici di transizione annuali degli individui.

Esaminando i dati sul numero di anni di povertà registrata dalle famiglie nel periodo considerato, è interessante scoprire che nel periodo 2005-2008 ben oltre il 71% delle famiglie dichiaranti e residenti a Brescia non è mai scesa sotto la soglia della povertà. D'altro canto, è piuttosto alta la percentuale delle famiglie che hanno sperimentato la condizione di povertà in tutti e quattro gli anni di osservazione (11,3%), seguita da quelle di un anno su quattro (7,7%). Nettamente inferiori le percentuali delle situazioni intermedie di due o tre anni di povertà (Tavola 18)[28]. Per gli individui si ottengono risultati simili a quelli delle famiglie.

[28] In queste elaborazioni come nelle precedenti, sono state escluse le sequenze delle famiglie in cui figuravano quattro anni di "senza reddito dichiarato". A differenza delle analisi precedenti, nelle tavole 18 e 19, si è scelto di includere le sequenze delle famiglie in cui figuravano 1, 2 o 3 casi di "senza reddito dichiarato". Questo è stato fatto al fine di non perdere l'informazione della sequenza posseduta dei casi validi.

Tavola 18 - Percentuale di famiglie e individui in povertà per numero di anni tra il 2005 e il 2008 - Analisi longitudinale

Anni di povertà									
Nessun anno		Un anno		Due anni		Tre anni		Quattro anni	
V.A.	%	V.A.	%	V.A.	%	V.A.	%	V.A.	%
				FAMIGLIE					
4.669	71,5	504	7,7	317	4,9	302	4,6	739	11,3
				INDIVIDUI					
10.201	71,0	1.122	7,8	733	5,1	719	5,0	1.587	11,0

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

Analizzando le sequenze degli stati di povertà ("sicuramente poveri" e/o "appena poveri"), emerge che quasi il 40% delle famiglie povere può essere definito di "lunga durata" (con tutte le limitazioni del caso, dovute al ristretto campo di osservazione), poiché presenta questa condizione in modo stabile in tutti i quattro anni osservati (Tavola 19). Piuttosto elevata è anche la quota delle famiglie che presenta questo stato in un anno solo (27%). Minori invece sono le percentuali degli altri due casi (due o tre anni di povertà): si tratta di famiglie border-line, che entrano e escono dalla condizione di povertà facilmente. Anche in questo caso, l'analisi per individui porta a risultati simili a quelli ottenuti per le famiglie (Tavola 19).

E' interessante notare il maggior coinvolgimento delle famiglie nello stato di povertà per l'anno 2008, pari al 69,4% delle famiglie povere in almeno un anno, contro il 66,9% del 2005, il 65,0% del 2006 e il 67,2% del 2007. Inoltre, si può asserire che gli effetti negativi della crisi erano già presenti nell'anno precedente al 2008: le percentuali di famiglie in povertà nel 2007 e nel 2008 sono infatti di certo più elevate di quelle degli anni precedenti.

I risultati dell'analisi delle sequenze degli stati di povertà e del numero di anni in povertà presentate però potrebbero essere migliorate. Queste infatti risentono della mancanza di informazione su quelle famiglie che pur essendo presenti in anagrafe non hanno dichiarato redditi. Come scelta metodologica, è sembrato opportuno includere nelle elaborazioni precedenti anche quelle sequenze in cui compariva almeno un dato incerto ("senza reddito dichiarato") ed escludere solo le sequenze composte da quattro anni "senza reddito", in modo da non perdere l'informazione di almeno un caso di povertà nella sequenza.

E' probabile che la presenza delle famiglie "senza reddito dichiarato" nei dati dell'archivio ALAR, non faccia emergere la povertà nella sua gravità, poiché molti di questi casi non dichiarati potrebbero corrispondere ad un effettivo stato di povertà (si pensi ad esempio a tutti coloro che percepiscono le pensioni sociali o di guerra ecc, che hanno bassi importi ai limiti della sussistenza). Inoltre, in circa un terzo dei casi in cui manca la dichiarazione, la causa è dovuta ad uno sfasamento temporale tra gli archivi anagrafici e fiscali, ossia a una mancata registrazione del reddito percepito nell'anno corretto.

Tavola 19 - Sequenze di stati di povertà nel periodo 2005-2008 e percentuale di famiglie in povertà per numero di anni tra il 2005 e il 2008 (a) **- Analisi longitudinale**

2005	2006	2007	2008	Famiglie		Individui	
				V. A	%	V. A.	%
			UN ANNO DI POVERTÀ				
0	0	0	1	180	9,7	394	9,5
0	0	1	0	90	4,8	198	4,8
0	1	0	0	68	3,7	139	3,3
1	0	0	0	166	8,9	391	9,4
	Totale un anno			**504**	**27,1**	**1.122**	**27,0**
			DUE ANNI DI POVERTÀ				
0	0	1	1	106	5,7	233	5,6
1	1	0	0	89	4,8	224	5,4
1	0	1	0	33	1,8	75	1,8
1	0	0	1	37	2,0	85	2,0
0	1	1	0	28	1,5	67	1,6
0	1	0	1	24	1,3	49	1,2
	Totale due anni			**317**	**17,1**	**733**	**17,6**
			TRE ANNI DI POVERTÀ				
0	1	1	1	121	6,5	261	6,3
1	1	1	0	95	5,1	243	5,8
1	1	0	1	46	2,5	122	2,9
1	0	1	1	40	2,1	93	2,2
	Totale tre anni			**302**	**16,2**	**719**	**17,2**
			QUATTRO ANNI DI POVERTÀ				
1	1	1	1	739	39,7	1.587	38,1
	Totale famiglie e individui			**1.862**	**100,0**	**4.161**	**100,0**

Fonte: Elaborazioni dell'autore su dati dell'archivio ALAR

Pertanto, al fine di migliorare le analisi delle sequenze longitudinali, si potrebbero fare delle ipotesi e attribuzioni di stato. Tenendo presente che difficilmente la condizione di povertà cambia stato da un anno all'altro e ricordando lo sfasamento tra archivi accennato, ad esempio, agli individui appartenenti a famiglie "senza reddito dichiarato" si potrebbe imputare la condizione registrata negli anni precedenti o successivi. Nel complesso, le situazioni da sistemare non sono molte, poiché rappresentano circa 646 casi.

I risultati ottenuti dopo queste correzioni, delineano una situazione della povertà peggiore della precedente: in questo caso il numero delle sequenze di quattro anni di povertà aumenta fino a toccare il 50,5% (anziché 38,1% precedente), a scapito ovviamente degli altri casi le cui percentuali diminuiscono.

E' ovvio che apportando questi correttivi non emergono i possibili cambiamenti nella condizione di povertà dovuti a shock tra un anno e l'altro. D'altro lato, però, si cerca di trovare una soluzione alla mancanza di informazione dei "senza reddito dichiarato".

6. Conclusioni e confronti con risultati di altri studi

In questo contributo, è stata descritta la metodologia utilizzata per costruire un archivio longitudinale al fine di stimare il reddito e la povertà delle famiglie e individui residenti del comune di Brescia.

Le principali caratteristiche di questo lavoro risiedono nell'utilizzare dati di fonte amministrativa, sopperendo alla carenza di informazioni di natura statistica a livello comunale (o *small-area*) e fornendo in questo modo stime in altro modo difficilmente recuperabili. Utilizzando dati di tipo amministrativo, già disponibili, si è potuto beneficiare dei brevi tempi di realizzazione, dei costi contenuti, della mancanza di molestia statistica e di tutti quei vantaggi di cui si è fatto cenno nel paragrafo 2.

Nel processo di realizzazione dell'archivio, non sono però mancate da affrontare problematiche metodologiche, legate alla presenza dell'*attrition* del campione, di elaborazione dei dati e di definizione del reddito.

Un'altra novità di questo studio è rappresentata dall'aver adottato per la stima di povertà una linea di povertà "*locale",* facendo riferimento ad un approccio di povertà monetaria, basato sui redditi *dichiarati* dagli individui in sede di denuncia fiscale.

L'elevata incidenza di povertà registrata nel comune di Brescia ed emersa in questo lavoro tramite la realizzazione dell'archivio ALER segue la direzione di altri lavori[29], in cui la povertà stimata con una soglia locale è più elevata di quella stimata con una soglia di povertà calcolata a livello nazionale. Prendendo come riferimento il rischio di povertà calcolato sui redditi degli individui emerge, infatti, che l'incidenza di povertà del comune di Brescia ottenuta con l'archivio ALAR è molto più elevata di quella ottenuta con l'indagine EU-SILC per la Lombardia: per il 2005 rispettivamente il 21,4% e il 10,5% (Istat 2010b). Il risultato ottenuto è più elevato anche di quello misurato dalla statistica ufficiale in Lombardia nel suo complesso (4,7% nel 2006[30]) e dell'Italia in generale (11,1%), basato sulle spese per i consumi delle famiglie[31].

La metodologia adottata per la realizzazione dell'archivio amministrativo sui redditi familiari ALER (dall'adozione delle regole di inseguimento, al *record-linkage* tra archivi, all'utilizzo di una linea di povertà relativa locale ecc.) ha portato a risultati delle stime di povertà coerenti con quelle di altri studi locali.

Le stime qui ottenute sono state confrontate con le stime di povertà dell'archivio DISREL[32] in cui sono riportati e armonizzati i risultati di diverse indagini realizzate a livello di *small-area*, utilizzando soglie di povertà locali basate sui redditi. Il valore

[29] Vedi nota 24.

[30] Per consultazioni si rimanda a http://dati.istat.it.

[31] Si fa notare che quest'ultima differenza è dovuta non solo all'utilizzo di una linea locale di povertà, ma anche al fatto che le stime ufficiali dell'Istat sulla povertà relativa calcolate sui consumi delle famiglie e non sul reddito familiare forniscono sempre stime più basse delle prime (Freguja, Pannuzi 2007). Le stime ufficiali sul tasso di povertà calcolate dall'Eurostat sui redditi sono anch'esse più elevate di quelle nazionali basate sui consumi (Istat): ad esempio per il 2006 e per l'Italia sono 19,6% quelle sui redditi, contro 11,1% di quelle sui consumi (http://epp.eurostat.ec.europa.eu/ e www.istat.it).

[32] L'archivio sulle Disuguaglianze di reddito a livello locale (Benassi, Colombini, 2007) contiene le stime delle seguenti indagini: Indagine sulle condizioni di vita delle famiglie toscane (17,3% per l'anno 2000), Indagine sulle condizioni economiche e sociali della provincia di Modena (12,4% nell'anno 2002), Indagine sulle condizioni economiche e sociali delle famiglie trentine (16,3%, per l'anno 2006), Stima del reddito delle famiglie bresciane del Comune di Brescia (18,3% per l'anno 2000), Indagine sui consumi delle famiglie milanesi (19,4, per l'anno 2006), da non confondersi con il Progetto AMeRIcA.

dell'incidenza di povertà ottenuta con l'archivio ALAR è lievemente superiore alla stima bresciana dell'anno 2000 (Palamenghi, Riva, Trentini 2005)[33] e in linea con i risultati delle altre rilevazioni, tenendo conto che queste indagini si riferiscono a periodi e a territori diversi e che la maggior parte di esse è basata sulla somministrazione di un questionario. E' bene sottolineare che tutte quante assumono valori nettamente superiori a quelli ufficiali noti a livello regionale e basati su una soglia di povertà nazionale.

Rispetto al Progetto AMeRIcA, realizzato con una metodologia simile[34] a quella qui adottata, l'incidenza di povertà delle famiglie ALAR è leggermente più bassa (21,4% nel 2005 contro il 25% per il 2004). Probabilmente su questa differenza di valore ha inciso il diverso approccio di povertà adottato nei due casi (sui redditi per ALAR e sui consumi per AMeRIcA).

Di recente (maggio 2011), anche il Centro Studi Sintesi di Venezia[35] ha stimato l'incidenza di povertà relativa nel comune di Brescia utilizzando dati fiscali dichiarati dagli individui in sede di presentazione dei redditi all'Agenzia delle Entrate e una soglia di povertà locale. Quest'ultima è basata sui livelli di spesa per consumi delle famiglie, tenendo conto anche della dimensione del nucleo familiare medio e del numero medio di percettori di reddito per ciascun nucleo familiare. Il risultato di questa ricerca è molto vicino a quello ottenuto con la realizzazione dell'archivio ALAR: rispettivamente per l'anno 2008, 20,0% e 21,6%.

I limiti di questa metodologia risiedono nella qualità degli archivi fiscali utilizzati (problemi di copertura e della presenza di famiglie "senza reddito dichiarato", di cui si è parlato a lungo) e nella possibile sottostima del reddito.

Come è risaputo la stima di variabili sensibili come il reddito non è facile: se rilevato attraverso la somministrazione di un questionario si possono incontrare le resistenze degli intervistati a rilasciare questa informazione con conseguenti effetti distorsivi sulle stime, se rilevata a fini amministrativi, in sede di denuncia dei redditi all'Agenzia delle Entrate, si può andare incontro ad una sottostima del reddito per motivi derivanti dalle caratteristiche degli archivi utilizzati o per dichiarazioni dei contribuenti non realistiche (ad esempio, per evasione fiscale). In effetti, le stime dei redditi familiari ottenuti tramite l'archivio ALAR sono inferiori a quelle ottenute con l'indagine EU-SILC (Istat 2008b), in cui è stata adottata una metodologia di confronto e correzione dei dati fiscali con quelli campionari da questionario al fine di migliorare la qualità del dato stimato (per l'anno 2005, il reddito familiare netto mediano è di 19,3 mila per il comune di Brescia (fonte ALAR) e di 25,8 mila euro, esclusi i fitti imputati, per la Lombardia, fonte EU-SILC).

Le stime complessive sui redditi potrebbero essere in parte migliorate se fosse possibile recuperare le informazioni delle famiglie "senza reddito dichiarato". Un miglioramento potrebbe essere ottenuto consultando direttamente i dati dell'Agenzia delle Entrate, anziché l'archivio SIATEL, in modo da recuperare sul territorio quelle situazioni di sfasamento tra

[33] Per approfondimenti si veda Comune 2012.

[34] Anche il Progetto AMeRIcA (Fattore Mezzanzanica 2007), si basa sull'integrazione di dati anagrafici con quelli fiscali, però fa riferimento all'universo della popolazione e non ad un campione come nel caso ALAR. Inoltre, per la stima della povertà utilizza la soglia di povertà regionale calcolata sui consumi delle famiglie (elaborazioni IReR), anziché una soglia di povertà locale basata sui redditi (ALAR).

[35] www.centrostudisintesi.com.

residenza anagrafica e fiscale che non sono riportate in SIATEL perché escono fuori dal campo di osservazione[36].

Per una migliore stima del reddito "disponibile" a livello individuale e familiare, inoltre, sarebbe interessante avere a disposizione dati su altre forme di tassazione, come le imposte pagate sugli immobili (nel 2008 ICI - Imposta comunale degli immobili), che rappresenta una diminuzione del reddito disponibile delle famiglie.

Tra gli altri possibili sviluppi futuri di questo studio, sarebbe molto utile costruire un archivio longitudinale osservando più anni, allungando quindi il periodo di riferimento.

Inoltre, sarebbe auspicabile un aggiornamento continuo dell'archivio, anno per anno, man mano che si rendano disponibili i dati fiscali, previo coinvolgimento degli enti produttori e fornitori di dati. Ciò consentirebbe di approfondire gli studi longitudinali andando a analizzare i processi di entrata e uscita delle famiglie e degli individui dallo stato di povertà, con le relative durate e permanenze. Molti recenti lavori stanno andando nella direzione di studiare modelli di durata per verificare e individuare quali siano le variabili individuali o familiari legate alle probabilità di entrata e uscita nello stato di povertà e soprattutto individuare se ci sono gruppi particolari di popolazione con determinate caratteristiche che presentano una povertà di lunga durata (Devicienti, Gualtieri 2006, 2011).

Infine, sarebbe estremamente interessante replicare lo studio su tutto il territorio nazionale con l'obiettivo di stimare una matrice di transizione tra condizioni di povertà e benessere a livello nazionale. In questo caso, potrebbe essere adottata una metodologia simile a questo lavoro, scegliendo un campione anagrafico di famiglie a cui associare i dati fiscali, con il vantaggio di ridurre la quantità di dati da trattare a livello nazionale.

[36] Si ricorda che SIATEL è un'estrazione dei dati fiscali posseduti dall'Agenzia delle Entrate. I dati sono stati estratti e selezionati per domicilio fiscale corrispondente al comune di Brescia.

Riferimenti bibliografici

Accolla G. 2009. *Povertà economica in Lombardia: dalla povertà relativa a quella assoluta* in ORES. L'esclusione sociale in Lombardia. Rapporto 2008. Milano: Guerini Associati, pp. 55 – 70.

Agenzia delle Entrate. 2006a. *770 - Semplificato 2006. Redditi 2005. Istruzioni per la compilazione*. www.agenziadelleentrate.gov.it

Agenzia delle Entrate. 2006b. *Modello 730 / 2006. Redditi 2005. Istruzioni per la compilazione* . www.agenziadelleentrate.gov.it

Agenzia delle Entrate. 2006c. *UNICO Persone fisiche 2006. Fascicolo 1. Dichiarazione delle persone fisiche. Periodo di imposta 2005. Istruzioni per la compilazione*. www.agenziadelleentrate.gov.it

Ballini F., Betti G., Lemmi A., Marzadro S., Morciano M., Neri L., Salvati N. 2007. *Stima della povertà a livello locale: i casi della regione Toscana e delle province di Modena e Trento*, in Brandolini A., Saraceno C. (a cura di), Povertà e Benessere: una geografia delle disuguaglianze in Italia. Bologna: Il Mulino.

Chiappero Martinetti E., Bottiroli Civardi M. 2002. *Povertà nei gruppi e povertà tra i gruppi*, in G. Carbonaro (a cura di), Studi sulla Povertà. Problemi di misura e analisi comparative. Franco Angeli, pp. 65- 86.

Chiappero Martinetti E. 2006. *Povertà multidimensionale, povertà come mancanza di capacità ed esclusione sociale: un'analisi critica e un tentativo di integrazione*, in G. Rovati (a cura di), Le dimensioni della povertà: strumenti di misure e povertà. Carocci, pp. 41-78.

Chiappero Martinetti E., Vega Pansini R. 2009. *Condizione economica, indebitamento e condizione abitativa delle famiglie italiane tra benessere e povertà*, in ORES, L'esclusione sociale in Lombardia. Rapporto 2008. Milano: Guerini Associati, pp. 71 – 94.

Comune M.E. 2012. *Diseguaglianze dei redditi e povertà delle famiglie attraverso dati amministrativi. Un'indagine longitudinale nel comune di Brescia*. Tesi di dottorato in Sociologia e metodologia della ricerca sociale. Università Cattolica del Sacro Cuore. http://tesionline.unicatt.it/

Devicienti F., Gualtieri V. 2006. *Dinamiche e persistenza della povertà in Italia: un'analisi con microdati panel di fonte ECHP*, in G. Rovati (a cura di), Le dimensioni della povertà: strumenti di misure e povertà. Carocci, pp. 179-208.

Devicienti F., Gualtieri V. 2011. *The persistence of income poverty and lyfe-style deprivation: Evidence from Italy*. ECINEQ. Working Paper, n° 229.

Fattore M., Mezzanzanica M. 2007. *Distribuzione dei redditi e disagio sociale: il caso di Milano*, in IReR, L'esclusione sociale in Lombardia. Milano: Guerini Associati, pp. 87-98.

Fortini M. 2000. *Linee guida metodologiche per rilevazioni statistiche, nozioni metodologiche di base e pratiche consigliate per rilevazioni statistiche dirette o basate su fonti amministrative*. Istat. Roma.

Freguja C., Pannuzi N. 2007. *La povertà in Italia: che cosa sappiamo dalle varie fonti?* in Brandolini A., Saraceno C. (a cura di), Povertà e Benessere. Una geografia delle diseguaglianze in Italia. Bologna: Il Mulino, pp. 23-60.

Golini A. 2001. *Alcune rapide riflessioni in tema di indagini longitudinali nel demografico e nel sociale*. Istat. Quaderni di ricerca. Rivista di Statistica Ufficiale. n.1, pp.75-77.

Istat. 2002. *Panel europeo sulle famiglie.* Metodi e norme n.15

Istat. 2008a. *L'indagine europea sui redditi e le condizioni di vita delle famiglie (EU-SILC).* Metodi e Norme, n. 37.

Istat. 2008b. *Distribuzione del reddito e condizioni di famiglia*. Statistiche in breve, pag.12.

Istat. 2009a. *Integrazione di dati campionari EU-SILC con dati di fonte amministrativa*. Metodi e Norme, n. 38.

Istat. 2009b. *La stima della povertà assoluta.* Metodi e Norme. n. 39, pp. 24-26.

Istat. 2010a. *La povertà in Italia. Anno 2010*. Statistiche Report. 15 luglio

Istat. 2010b. *La distribuzione del reddito in Italia - Indagine europea sui redditi e condizioni di vita delle famiglie (EU-SILC)* . Anno 2006. Argomenti n. 38, pp. 82-83.

Lalla M. 2003. *Il disegno della seconda indagine sulle condizioni economiche e sociali delle famiglie nella Provincia di Modena*. Università degli Studi di Modena e Reggio Emilia Dipartimento di Economia Politica. CAPP. Centro di Analisi delle Politiche Pubbliche. Materiali di discussione, n. 512.

Leti G. 2001. *La commissione Istat per la progettazione di indagini longitudinali sulle imprese e sulle famiglie*. Istat. Quaderni di ricerca. Rivista di Statistica Ufficiale, n. 1, pp.59-66.

Marcia Freed Taylor M.F., Brice J., Buck N., Prentice-Lane E. 2009. *British Household Panel Survey*. User manual volume A. Introduction, technical report and appendices, February.

McGonagle K., Schoeni R. 2006. *The panel Study of Income Dynamics: Overview and Summary of Scientific Contributions After Nearly 40 Years*. Technical Series Paper 06-01.

Mori A. 2007. *Povertà economica: stime mediante i redditi*, in IReR, L'esclusione sociale in Lombardia. Milano: Guerini Associati, pp. 59-70.

Negri N., Solera C. 2007. *Povertà e vulnerabilità delle coppie nel Canavese* in Brandolini A., Saraceno C. (a cura di), Povertà e Benessere. Una geografia delle disuguaglianze in Italia. Bologna: Il Mulino.

Palamenghi M., Riva L., Trentini M. 2005. *Criteri e metodi di stima del reddito delle famiglie bresciane*. Università degli Studi di Brescia. Rapporti di Ricerca del Dipartimento Metodi Quantitativi. Quaderno n. 247.

Ruspini E. 2004. *La ricerca longitudinale.* Metodologia delle Scienze umane. Franco Angeli.

Trivellato U. et al. 1995. *Prospettive per possibili analisi longitudinali nella statistica ufficiale italiana.* Rapporto di ricerca per la Commissione per la garanzia per l'informazione statistica.

Evaluating administrative data quality as input of the statistical production process[1]

Fulvia Cerroni[2], Grazia Di Bella[3], Lorena Galiè[4]

Sommario

Valutare e analizzare la qualità dei dati da fonte amministrativa è un'esigenza crescente negli Istituti nazionali di statistica (INS), poichè i processi di produzione sempre più utilizzano questo tipo di dati. Monitorare la qualità delle forniture di dati amministrativi che entrano nel processo di produzione statistica, valutare il loro possibile utilizzo a fini statistici, supportare l'acquisizione dei dati amministrativi sono azioni che devono essere eseguite nelle prime fasi del processo di produzione. L'articolo riporta l'esperienza dell'Istat nell'ambito del progetto europeo BLUE-ETS volto a sviluppare un quadro concettuale della qualità dei dati amministrativi sulla base di un approccio multidimensionale di indicatori di qualità e a definire un nuovo strumento, la Quality Report Card, che possa essere associato ai dati amministrativi e utilizzato in generale dagli INS.

Parole chiave: dati amministrativi, qualità dei dati, indicatori di qualità dell'input.

Abstract

Evaluating and reporting data quality of administrative sources is a growing need for National Statistical Institutes (NSIs) as more and more production processes are using this type of data source. To monitor the quality of the administrative data supply that enter the statistical production process, to evaluate its possible use for statistical purposes and to support administrative data acquisition are tasks that should be performed in the early stages of the production process. The paper reports Istat experience within the European project BLUE-ETS in developing a conceptual framework of administrative data quality based on a multidimensional approach of quality indicators and in defining a new comprehensive instrument, the Quality Report Card, that can be associated to administrative data and generally used by NSIs.

Keywords: administrative data, data quality, input quality indicators.

---

[1] The authors thank Piet Daas and the other BLUE ETS WP4 members. Although the paper is the result of the combined work of the authors, the Sections are to be awarded as it follows: Sections 1, 2.1, 4 and 5 to Grazia Di Bella; Sections 2.1 and 2.2 to Cerroni Fulvia; Section 3 to Lorena Galiè. The text published exclusively engages the authors, the views expressed do not imply any liability by Istat.

[2] Technical assistant (Istat), e-mail: cerroni@istat.it.

[3] Senior researcher (Istat), e-mail: dibella@istat.it.

[4] Technical assistant (Istat), e-mail: galie@istat.it.

The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat.

1. The need to define the statistical quality of administrative data

Administrative data (AD) were added in the last few years as a further source next to the data collected from sample and census surveys, for the production of official statistics in National Statistical Institutes (NSIs).

This synergy enables to make the statistical production process more efficient: it is possible to expand the available statistical information, to reduce the so-called "statistical burden" among economic agents, to maximize available resources (Unece, 2007). But, if the use of administrative sources in the statistical process allows to reduce the task of data capturing, new tools need to be addressed: a) for the acquisition of AD; b) for their use in the statistics production process. From the management point of view, AD acquisition requires to define new production functions and new organizational structures capable of maintaining relations with AD holders (ADH) and improving the collaboration process. From the methodological point of view it is necessary to define shared and standardized procedures to meet *a posteriori* the quality standards imposed by official statistics (Wallgren and Wallgren, 2007).

International organizations, involved in producing statistics, are favoring the development of standardized methodologies or the sharing of best practices for the use of AD and there are many initiatives on this subject[5]. Considering the wide range that characterizes the types of AD and the different ways in which they are currently used in the statistical production process, the question arises to what extent it is possible to define generalized methods for their statistical use.

Meanwhile, the NSI's production processes are deeply evolving, not only in the field of business statistics for which the availability of AD is more extensive.

An interesting overview on actual uses of AD for producing business statistics (SBS, STS, Prodcom and Business Register Regulations) in all Member States and EFTA Countries has been conducted by the ESSnet AdminData – Workpackage 1 "Overview of existing practices in the use of AD for producing business statistics over Europe". All information collected is made available to users (internal NSI users) in a Database, which can be browsed by topic, by domain (regulation) and by country. It is possible to get information on the combination of sources used for producing statistics (among survey data, admin data and registers) and on how AD are used: directly for producing statistics, or indirectly for the sampling frame, in editing & validation, in imputation of missing values, in estimation procedures[6].

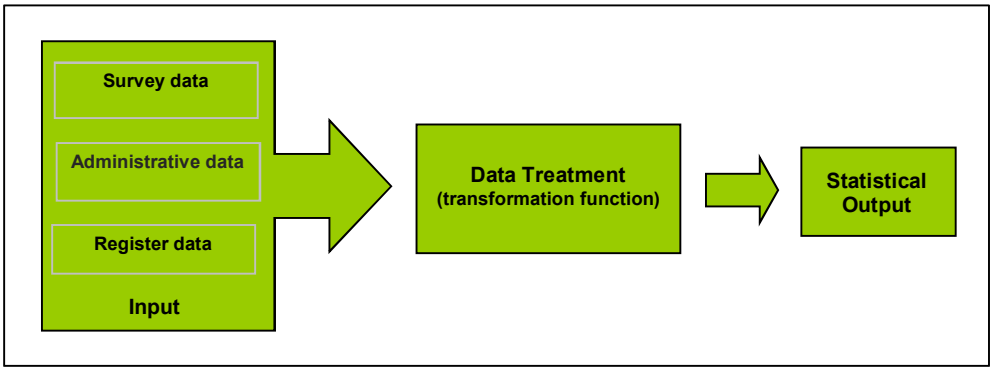
To better focus on the paper objectives, in Figure 1 the statistical process that uses AD is shown. Sometimes the input to the process may consist of a combination of data from different sources and it is also common to use multiple Administrative Sources within the

[5] Nordic countries produced comprehensive documentation of their best practices based on their long experience in using administrative data also for producing censuses data (Unece, 2007). The European Commission programme called MEETS - Modernisation of European Enterprise and Trade Statistics - has funded several activities on this issues in these last years (http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/MEETS_programme); more information for projects that have addressed the AD issue are available at: (a) http://cros-portal.eu/ for ESSnet AdminData (European Statistical System project on The Use Of Administrative And Accounts Data For Business Statistics) had the scope to find common ways for use of dministrative data for business statistics; (b) http://www.blue-ets.istat.it/ for BLUE-ETS project, WP4 that investigated the possibility to increase the use of AD for statistical purposes.

[6] Costanzo et al. (2011), ESSnet AdminData (2013a), Database available on the web site: http://essnet.admindata.eu/.

same process, then data are integrated and treated to produce the statistical output. These are the cases: (a) when an administrative source fails to meet the requirements of quality and multiple sources need to be combined in order to adjust the shortcomings of the separate sources; (b) when data integration produces a much richer information content, such as Linked Employer Employee data or Educational data combined with employment data and so on; (c) for producing longitudinal data (Wallgren and Wallgren, 2007).

Figure 1 - The statistical production process that uses administrative data



Sometimes AD can be used as they are to produce statistical data: this is the limit case where the transformation function in Figure 1 is the identity function. Generally, AD enter the production process directly for producing statistics or registers after the Treatment procedure, or they support the survey process: for producing the sampling frame, in editing & validation process, in imputation of missing values, for unit non response treatment, in estimation procedures[7].

In this paper the problem of defining the quality of AD which enter the statistical production process will be treated. Regardless of the manner in which the AD are used, their quality evaluation in terms of input process is a useful information that has to be associated to AD. In general the lower the quality of the input data and the greater the effort to bring the output data to acceptable quality levels.

The work here presented is based on the results obtained within the BLUE-ETS project, ended in 2013 and specifically Work package 4 (WP4) "Improve the use of administrative sources" aimed to develop an instrument able to determine the statistical quality of AD, a Quality Report Card for Administrative data (QRCA), generally applicable to AD sources in different European countries (Daas et al., 2011a, 2011b, 2013).

The conceptual framework of the QRCA will be presented and relevant quality indicators selected and classified in order to evaluate the statistical usability of AD will be described. In the development of quality indicators associated with the use of AD for statistical purposes, it is useful to distinguish three types of indicators:

[7] It has to be considered that it is not so rare the case in which public administrations directly produce statistical data from their own administrative data as a result of an agreement within the National Statistical System (Sistan in Italy).

1. Input quality indicators: to define the quality of AD used as input in the statistical production process;
2. Process quality indicators: to measure the quality associated with the production process that uses AD to produce statistics;
3. Output quality indicators: to measure the output quality of statistics involving AD, taking input and process quality into account.

In this paper indicators of the type 1 evaluating AD are analyzed. A further useful specification of Input quality indicators, developed within BLUE-ETS WP4 work (Daas et al., 2011), considers the value of the additional information brought to the specific statistical production process by the Administrative Source. A general quality assessment not considering the specific additional information to a statistical process is referred to as Data Source Quality (DSQ), otherwise it is called Input Output-oriented Quality (IOQ).

Quality aspects related to the output production will not be considered here. An interesting analysis of the overall quality of register-based statistics is developed in Statistics Sweden (Wallgren, Wallgren, 2007; Laitila et al, 2011). Within the ESSnet AdminData project quality indicators of type 3 have been studied: starting from the state of play in terms of the use of quality indicators for business statistics involving AD across NSIs[8], output quality indicators for statistics involving AD have been produced (ESSnet AdminData, 2013a)[9].

In Section 2 the conceptual framework of the AD quality is described. To test the quality framework, an application to a case-study is reported in Section 3. The Social Security Data source actually in use in more statistical processes and under analysis for other potential uses in Istat has been evaluated as case study. Some general issues on the QRCA usability in Istat are also presented in Section 4.

Before proceeding, it is useful to firstly focus on some definitions used in this paper and some issues related to the context.

The concept of AD quality, as it is here considered, concerns the quality in terms of statistical usability in the production process. For instance an administrative data set with a very good quality for its original purpose may have a poor statistical quality that can affect its statistical usability. Regarding the definitions, we rely on the ESSnet Admin Data Glossary[10] so *Administrative source* and *Administrative dataset* are defined as follows:

Administrative Source

> A data holding containing information collected and maintained for the purpose of implementing one or more administrative regulations. In a wider sense, any data source containing information that is not primarily collected for statistical purposes.

Administrative Dataset

> Any organised set of data extracted from one or more Administrative Sources, before any processing or validation by the NSIs.

[8] This work is already in place on the production of Eurostat Quality Report Framework for Business Statistics under Regulation (CE) no. 295/2008 and user test carried out within EU and EFTA countries NSIs.

[9] For more information on ESSnet AdminData - Work Package 6 results see the website http://essnet.admindata.eu/.

[10] ESSnet AdminData Glossary http://essnet.admindata.eu/.

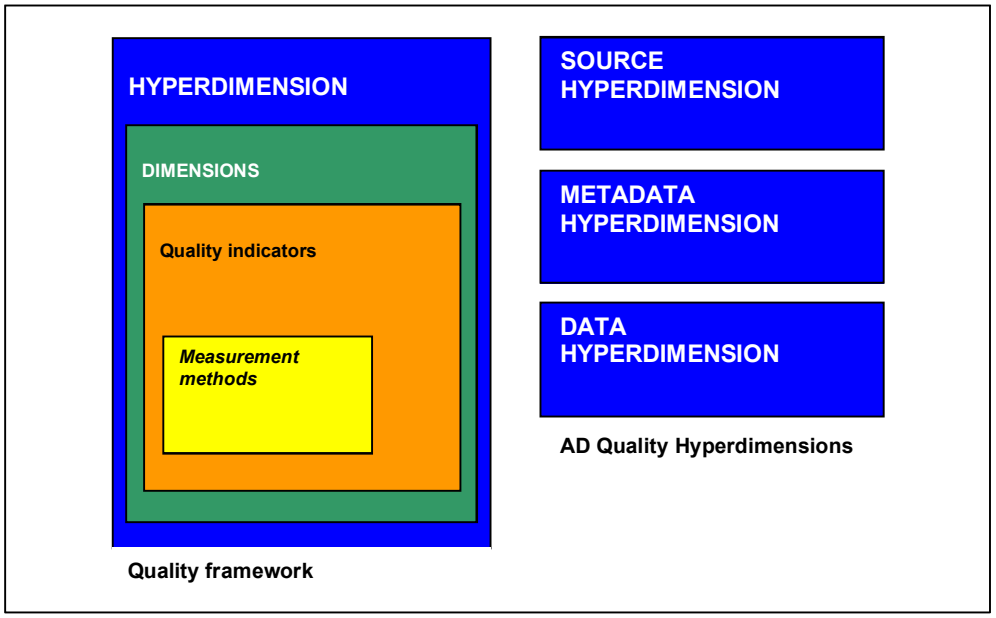
Administrative Data are considered as data derived from an Administrative Source. The term AD quality refers to the statistical quality of the Administrative Dataset provided by AD source holder and received by Istat.

2. The conceptual framework

2.1 A generalized and flexible framework

Following Daas and Ossen (2011), it is useful to define a framework for the statistical quality of AD using a hierarchical and multidimensional approach whose graphical representation is shown in Figure 2. Here the hierarchical aspect is made by the levels in which quality breaks down. Four levels are considered: i) the Hyperdimension (or category) level, ii) the Dimension level that breaks down each Hyperdimension; iii) quality indicators within each Dimension and iv) several measurement methods included in each quality indicator.

Figure 2 - Multidimensional and hierarchical quality framework



Considering secondary data source quality[11], three relevant Hyperdimensions has been selected: (a) the Source itself by considering the data acquisition procedure, the legal basis for its use and a description of the data source use effects on the NSI; (b) the Metadata focused on conceptual aspects such as the units and variables definition, their comparability

[11] "Secondary data source" as defined by Hox and Boeije, 2005: "Data originally collected for a different purpose and reused for another research question". This is a more general definition including also other kind of data such as: commercial data and big data.

with NSI's definitions and the presence of unique identification keys; (c) the Data Hyperdimension related to the data quality (facts).

The multidimensionality is given by the fact that the approach looks at the quality of the AD source as a whole and not only with reference to the Data: that is the reason why the concept of Hyperdimension is introduced.

While the Hyperdimensions, the Dimensions and the Quality indicators are fixed, the Measurement methods are flexible as it is possible to choose within more suggested methods which are the most suitable to measure each indicator. This approach allows for a flexible AD quality evaluation structure capable of defining a generalized tool regardless of:

- the types of AD (tax data, social security data, educational data,…);
- the statistical processes involving AD (business statistics, population statistics, social statistics,…);
- the way AD are used (to directly produce statistics, to form a frame for sample surveys, to provide auxiliary information for sampling design or estimation purpose, for edit and imputation procedures).

As far as the Source and the Metadata Hyperdimensions are concerned, Dimensions and indicators are well described in Daas and Ossen (2011) where a Checklist was developed to assist the preliminary evaluation of AD and support the decision to use or not the source in the statistical production process before acquiring data. The Source and the Metadata quality Dimensions are reported below in Table 1.

Table 1 - AD quality Dimensions and indicators for Source and Metadata Hyperdimension

HYPERDIMENSIONS	Dimensions
Source	1. AD source holder (information for the acquisition data procedure). 2. Relevance of the AD source. 3. Privacy and security. 4. Delivery. 5. Relationships and feedback with the AD source holder.
Metadata	1. Clarity and interpretability . 2. Comparability at the metadata level. 3. Unique keys. 4. Data treatment (by data source keeper).

In the next Section the Data Hyperdimension components, as defined in the BLUE-ETS WP4 project, will be presented. Results from the application of quality indicators on AD should be collected and shown in a quality report, the QRCA, which is an instrument for the input quality evaluation in statistical production processes based on AD.

Obviously the indicators presented are producer statistics-oriented and not user statistics-oriented since their aim is to support the statistical production within NSIs. In paragraph 4 the objectives of the QRCA are described considering the different type of users.

2.2 Dimensions, indicators and measurement methods

The topic of this section is to firstly define the Dimensions selected to evaluate the quality related to the Data Hyperdimension and secondly to describe the selected indicators within the Dimensions.

Measuring AD statistical quality is not the same thing as measuring statistical survey data quality. It is well known that the quality of statistics (i.e. the statistical output) is defined by Eurostat with reference to the following six quality Dimensions (Eurostat 2003a, 2003b, 2005):

- relevance;
- accuracy;
- timeliness and punctuality;
- accessibility and clarity;
- comparability;
- coherence.

However these criteria do not apply to AD as they are not able to correctly or exhaustively evaluate all the quality aspects of AD considered as the input of the statistics production process. Therefore a further effort has to be made to revise the existing statistical survey quality framework and to formalize a new one useful for the AD quality evaluation.

The five quality Dimensions selected for AD by the BLUE-ETS WP4 project are described and reported below in Table 2. From a general point of view they focus on the AD quality from the moment they arrive in the NSI, till the moment they are available as input for the statistical production process. Having said this, Eurostat components formerly defined to report on the quality of survey statistics (users statistics-oriented) should be totally reviewed into an input quality perspective (producer statistics-oriented). Hence this means a change of the viewpoint for some existent Dimensions, like Accuracy and Completeness, and the introduction of new Dimensions aimed to evaluate some very crucial aspects of the AD quality, like Technical checks and Integrability, that are totally new with respect to the survey's world as described below (Table 2). In particular the new Integrability Dimension is of primary importance since AD concepts, rules and classification criteria for objects and variables are different from the NSI's ones and to make AD usable for statistical purposes often a reconciliation has to be made (Wallgren and Wallgren, 2007). Integrability Dimension measures the extent to which AD can be integrated in the statistical production process.

Table 2 - AD quality Dimensions in the Data Hyperdimension

DIMENSION	Description
Technical checks	Technical usability of the file and data in the file.
Integrability	Extent to which the data source is capable of undergoing integration or of being integrated.
Accuracy	The extent to which data are correct, reliable and certified.
Completeness	Degree to which a data source includes data describing the corresponding set of real-world objects and variables.
Time-related Dimension	Indicators that are time or stability related.

The AD quality Dimensions may be applied to different levels: the entire dataset, the objects[12] and the variables. Actually this classification allows maintaining simple, comprehensive and compact the theoretical structure. Quality indicators by Dimension and level of application are shown in Table 3.

Table 3 - Quality indicators by Dimension

INDICATORS BY DIMENSION	Level	Description
1. Technical checks		
Readability	Dataset	Accessibility of the file and data in the file.
Convertibility	Objects	Conversion of the file to the NSI-standard format.
File declaration compliance	Variables	Compliance of the data in the file to the metadata agreements.
2. Integrability		
Comparability of objects	Objects	Similarity of objects in source with the objects used by NSI.
Alignment of objects	Objects	Linking-ability (align-ability) of objects in source with those of NSI.
Linking variable	Variables	Usefulness of linking variables (keys) in source.
Comparability of variables	Variables	Proximity (closeness) of variables.
3. Accuracy		
Authenticity	Objects	Legitimacy of objects.
Inconsistent objects	Objects	Extent of erroneous objects.
Dubious objects	Objects	Presence of untrustworthy objects.
Measurement error	Variables	Correctness of a value with respect to the measurement process.
Inconsistent values	Variables	Extent of inconsistent (out of range) variable's values or combinations of values for variables.
Dubious values	Variables	Presence of implausible values or combinations of values for variables.
4. Completeness		
Under-coverage	Objects	Absence of target objects (missing objects) in the dataset.
Over-coverage	Objects	Presence of non-target objects in the dataset.
Selectivity	Objects	Statistical coverage and representativeness of objects.
Redundancy	Objects	Presence of multiple registrations of objects.
Missing values	Variables	Absence of values for (key) variables.
Imputed values	Variables	Presence of values resulting from imputation actions by data source holder.
5. Time-related Dimension		
Timeliness	Dataset	Lapse of time between the end of the reference period and the moment of receipt of the dataset.
Punctuality	Dataset	Possible time lag between the actual delivery date of the dataset and the date it should have been delivered.
Overall time lag	Dataset	Overall time difference between the end of the reference period in the dataset and the moment the NSI has concluded that it can definitely be used
Delay	Dataset	Extent of delays in registration.
Dynamics of objects	Objects	Changes in the population of objects (new and dead objects) over time.
Stability of variables	Variables	Changes of variables or values over time.

[12] The term "object" is used to generalize and to include events registered or units of a target set (administrative population); from objects should be possible to derive statistical units (Zhang, 2011).

For example, within the Accuracy Dimension inconsistency is evaluated for objects as well as for variables in order to have a complete view of the extent to which data are correct, reliable and certified. Same symmetry occurs for Completeness too as the assessment can take place in terms of objects (under and over coverage indicators) and in terms of variables (indicators for missing values).

To draw a parallel and alight the differences with the survey data quality assessment (Eurostat, 2003b), it has to be said that the Dimensions of Accuracy and Completeness are fully reviewed. In surveys under and over-coverage errors are indicators included into the Accuracy Dimension as well as the missing values that are "item non response". In AD quality the coverage is, instead, evaluated only *a posterior* with respect to the statistical target population and therefore it is not connected to the Accuracy Dimension but concerns the data completeness. This approach is typical of AD because the statistical population is not defined *a priori* as in the case of statistical surveys.

Missing values are included into the Completeness Dimension too as they may arise from inaccuracy of the source, but also by the fact that most of the times some variables are not mandatory for subpopulations while they are considered relevant for statistical purposes: so it is a problem of completeness rather than accuracy.

It is important to note that the quality Dimensions are not mutually orthogonal and, as a matter of fact, some trade - offs are present. For instance it often happens that, to get timely data, a lower quality in the Completeness Dimension is considered acceptable, for example for the production of short-term statistics.

From the operational point of view, several R scripts have been developed that allow to automatically calculate some suggested indicators and provide attractive graphical displays of data that can help the application of the quality framework within the NISs (Tennekes et al., 2011).

After a general overview on the AD quality Dimensions, in the following paragraphs indicators and measurement methods are described.

2.2.1 Technical checks and indicators

The Technical checks Dimension includes the IT-related indicators for the data in a source through which the technical usability of the file and data in the file are verified (Daas et al., 2011b). This Dimension includes three indicators:

1. the Readability;
2. the Convertibility;
3. the File declaration compliance.

All these indicators are associated to the dataset and check respectively whether (1) it is impossible to physically access the data as the file cannot be opened or it is damaged, or (2) data are not correctly convertible into NSI-standard formats, or (3) the data delivered are not conform to the definitions included in the metadata, if any are provided, or data do not comply with the request of the NSI.

Since this Dimension looks at the technical usability of files and data, it is useful to remark that the corresponding quality indicators are important:

a) to support the data loading and decide whether carrying on using the data source or going back to the ADH because of errors in the delivery;
b) for monitoring data deliveries when the source becomes in use fully operative;

c) to provide assistance when a new data source is being studied and its technical usability is explored for the first time.

To have an overall look of the data provided, with the aim to identify obvious errors in supply, it is helpful to use the Tableplot function available in the R package (Tennekes et al., 2011).

Some measurement methods for the Technical Checks indicators are proposed in the following Table 4.

Table 4 - Suggested measurement methods by indicator for Technical checks

INDICATORS BY DIMENSION	Measurement methods
Readability	a. % of deliveries (or files) of the total deliveries with an unknown extension, that are corrupted or cannot be opened b. % of the total file which is unreadable (MB/GB size) or number of unreadable records
Convertibility	% of objects with decoding errors or corrupted data.
File declaration compliance	% of variables in the current delivery that differ from metadata lay-out delivered or agreed upon in: i) formats and names ii) variable and attribute content iii) categories defined for categorical variables iv) ranges for numerical variables

2.2.2 Integrability and indicators

This Dimension contains indicators aimed to evaluate the ease by which the data in the source can be integrated into the statistical production system of an NSI. Since it immediately looks at the integration process, the idea standing behind this Dimension is clearly an IOQ view (Daas et al., 2011b).

It should be noted that the Integrability is a characteristic Dimension in assessing the input quality of an AD source. In facts, since AD are primarily collected for non-statistical purposes and describe non statistical concepts (e.g. administrative, fiscal) they firstly need to be converted into statistical concepts by appropriate harmonization. The reconciliation of concepts and definitions for objects and variables on AD source plays an important role for the source evaluation in terms of the Integrability Dimension before data are actually integrated and involves both Metadata and Data Hyperdimensions.

The Integrability Dimension includes indicators for objects and for variables.

Comparability and Alignment indicators evaluate the similarity of objects in the Administrative Source and their linking-ability with those used in the statistical production system by measuring the distance from the point of view of the objects definition. Administrative objects are analyzed with respect to their degree of comparability with the statistical objects and they are evaluated as identical, corresponding or incomparable objects according respectively to the fact that (a) they have exactly the same unit of analysis and definition as those used by NSI or (b) they correspond after harmonization or (c) they are not comparable.

Some measurement methods for indicators are proposed in Table 5.

Table 5 - Proposed measurement methods for Integrability indicators

INDICATORS BY DIMENSION	Measurement methods
Comparability of objects	a. % of identical objects = (Number of objects with exactly the same unit of analysis and same concept definition as those used by NSI) / (Total number of relevant objects in source) x 100 b. % of corresponding objects = (Number of objects that, after harmonization, would correspond to the unit needed by NSI) / (Total number of relevant objects in source) x 100 c. % of incomparable objects = (Number of objects that, even after harmonization, will not be comparable to one of the units needed by NSI) / (Total number of relevant objects in source) x 100 d. % of non-corresponding aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 - fraction of objects of interest at the same aggregated level in source 2) x 100
Alignment of objects	a. % of identical aligned objects = (Number of objects in the reference statistical population with exactly the same unit of analysis and same concept definition as those in the source) / (Total number of relevant objects in the reference statistical population) x 100 b. % of corresponding aligned objects = (Number of objects in the reference statistical population that, after harmonization, correspond to units or parts of units in the source) / (Total number of relevant objects in the reference statistical population) x 100 c. % of non-aligned objects = (Number of objects in the reference statistical population that, even after harmonization of the objects in the source, cannot be aligned to one of the units in the source) / (Total number of relevant objects in the reference statistical population) x 100 d. % of non-aligned aggregated objects = (Fraction of objects of interest at an aggregated level in source 1 that cannot be aligned + fraction of objects of interest at the same aggregated level in source 2 that cannot be aligned) x 100
Linking variable	a. % of objects with no linking variable = (Number of objects in source without a linking variable) / (Total number of objects in the source) x 100 b. % of objects with linking variables different from the ones used by NSI = (Number of objects in source with linking variables different from the one used by the NSI) / (Total number of objects with linking variables in the source) x 100 c. % of objects with correctly convertible linking variable = (Number of objects in the source for which the original linking variable can be converted to one used by the NSI) / (Total number of objects with a linking variable in the source) x 100
Comparability of variables	a. Use statistical data inspection methods to compare the totals of groupings of specific objects for variables in both sources. Graphical methods that can be used are a bar plot and a scatter plot. Distributions of values can also be compared b. The Mean Absolute Percentage Error (MAPE) that measures the mean of the absolute percentage error c. A method derived from the chi-square test that evaluates the distributions of the numeric values in both data sets. For categorical data Cramer's V (Cramer, 1946) could be used. The comparison could be performed either by groups (macro level) or at the micro level d. % of objects with identical variable values = (Number of objects in source 1 and 2 with exactly the same value for the variable under study) / (Total number of relevant objects in both sources) x 100

All these methods highlight the complexity of comparing sources and how important is the Integrability Dimension for evaluating the statistical usability of an AD source and its integration in the statistical process.

2.2.3 Accuracy and indicators

Indicators in this Dimension are derived from the sources of error scheme for AD firstly proposed by Bakker (2010) and developed by Zhang (2012) where errors occurring during the collection of AD are described from the moment they are collected by the ADH up to the moment when the data are linked to other statistical data sources to be used as input for NSI. Accuracy of AD is defined as the extent to which data are correct, reliable and certified. Some measurement methods for indicators are proposed for objects and variables in the following Table 6.

The indicators for objects evaluate the correctness of the units or events registered in the source and include the Authenticity which measures the objects correspondence to the real world and the Inconsistent or dubious objects indicators that check for the objects involvement in respectively inconsistent or dubious relations.

The indicators for variables evaluate the validity of the variables as correctness of the values of the units or events registered in the source. These indicators include the Measurement errors which measure the correctness in terms of deviation of data value from ideal error-free measurements and the Inconsistent or dubious values that check for objects with values involved in respectively inconsistent or dubious relations.

It is necessary to highlight the difference with the concept of Measurement errors used in the data survey quality. The Measurement errors evaluate the distance between the observed value and the true value but in the administrative source they are by definition out of the NSI control as these errors are the result of the data collection carried out by the ADH. Thus the measurement errors could only be known ex post in an indirect way by asking the ADH information about the quality checks management during the data collection phase if any exists.

Table 6 - Proposed measurement methods for Accuracy indicators

INDICATORS BY DIMENSION	Measurement methods
Authenticity	a. % of objects with a non-syntactically correct identification key b. % of objects for which the data source contains information contradictive to information in a reference list for those objects (master list and target list) c. Contact the data source holder for their % of non-authentic objects in the source
Inconsistent objects	% of objects involved in non-logical relations with other (aggregates of) objects.
Dubious objects	% of objects involved in implausible but not necessarily incorrect relations with other (aggregates of) objects.
Measurement error	a. % of unmarked values in the data source for each variable (when values not containing measurement errors are marked by AD holder) b. Contact the data source holder and ask the following data quality management questions: - Do they apply any design to the data collection process (if possible)? - Do they use a process for checking values during the reporting phase? - Do they use a benchmark for some variables? - Do they use a checking process for data entry? - Do they use any checks for correcting data during the processing or data maintenance?
Inconsistent values	% of objects with inconsistent (out of range) variable's values or objects whose combinations of values for variables are involved in non-logical relations.
Dubious values	% of objects with dubious variable's values or objects whose combinations of values for variables are involved in implausible but not necessarily incorrect relations.

2.2.4 Completeness and indicators

This Dimension is defined as the degree to which a data source includes data describing the corresponding set of real world objects and variables. Thus the indicators for objects in this Dimension mainly focus on coverage topics while the indicators for variables are related to missing and imputed values.

Actually the indicators for objects are the indicators of (1) Under-coverage, (2) Over-coverage, which measures the coverage with respect to a target statistical population, (3) the Selectivity which evaluates the coverage by specific statistical subpopulations and (4) the Redundancy that checks for duplications in the recording of objects.

The indicators for variables are indeed the Missing values - that evaluate objects with completely or partially missing values for key variables (missing units and missing items) and Imputed values that calculate objects with values imputed by the ADH.

It is useful to highlight that the way to calculate indicators of Under-coverage, Over-coverage and Selectivity is twofold. In facts they can be computed with respect to the statistical target populations (that means an IOQ view) or to the administrative target population of the source (in a DSQ view). It should be noted that most of the times the statistical target population is not available for timeliness reasons while the administrative target population generally may not exist at all.

As far as Redundancy is concerned, although it measures the quality of the delivered data by counting multiple registrations of data objects in the source, however it is important to underline that sometimes same object may have multiple registrations because it is a characteristic of the AD source (e.g. in more registrations the AD source records several information about the same employee).

Some measurement methods for indicators are proposed in the following Table 7.

Table 7 - Proposed measurement methods for Completeness indicators

INDICATORS BY DIMENSION	Measurement methods
Under-coverage	% of objects of the reference list missing in the source.
Over-coverage	a. % of objects in the source not included in the reference population b. % of objects in the source not belonging to the target population of the NSI
Selectivity	a. Use statistical data inspection methods, such as histograms, to compare a background variable (or more than one) for the objects in the data source and the reference population b. Use of more advanced graphical methods, such as table plots c. Calculate the Representativeness indicator (R-indicator; Schouten et al, 2009) for the objects in the source
Redundancy	a. % of duplicate objects in the source (with the same identification number) b. % of duplicate objects in the source with the same values for a selection of variables c. % of duplicate objects in the source with the same values for all variables
Missing values	a. % of objects with a missing value for a particular variable b. % of objects with all values missing for a selected (limited) number of variables c. Use of graphical methods to inspect for missing values for variables
Imputed values	a. % of imputed values per variable in the source b. Contact the data source holder and request the percentage of imputed values per variable

2.2.5 Time-related Dimension and indicators

The quality indicators in this Dimension are all related to time. The Timeliness, the Punctuality, and the Overall time lag indicators apply to the delivery of the administrative dataset. The Overall time lag indicator, which measures the total time lag between the reference period and the moment at which data can be used by the NSI, also includes the time required for evaluation. The Delay indicator is built up to evaluate how fresh is the information stored in the AD source with respect to the real world events as it aims to measure the extent of delays in registration.

These indicators are all relevant for both DSQ and IOQ.

The last two indicators apply to objects (Dynamics of objects) and to variables (Stability of variables). The indicator for objects aims to describe changes in the population over time by comparing the population of objects referred to time t (delivery τ) to that referred to time t-1 (delivery $\tau-1$). In this case the indicator does not express a direct evaluation of the data quality (the population dynamics depends on the phenomenon under study), but only a description of the dynamics. The indicator of Stability of variables describes the stability in terms of the changes over time in variable values on persistent objects at time t (delivery τ) compared to those at time t-1 (delivery $\tau-1$). Here the attention is focused on the variable composition covered by a source that should be stable in time between subsequent deliveries (e.g. the company Nace code).

It should be noted that concerning data over time it is very useful to consider tools for analysis of time series and longitudinal data.

Some measurement methods for indicators are proposed in the following Table 8.

Table 8 - Proposed measurement methods for Time-related Dimension indicators

INDICATORS BY DIMENSION	Measurement methods
Timeliness	a. Time difference (days) = (Date of receipt by NSI) – (Date of the end of the reference period over which the data source reports) b. Time difference (days) = (Date of receipt by user) – (Date of the end of the reference period over which the data source reports)
Punctuality	Time difference (days) = (Date of receipt by NSI) – (Date agreed upon; as laid down in the contract).
Overall time lag	Total time difference (days) = (Predicted date at which the NSI declares that the source can be used) – (Date of the end of the reference period over which the data source reports).
Delay	a. Contact the data source holder to provide their information on registration delays b. Time difference (days) = (Date of capturing the change in the source by the data source holder) – (Date the change occurred in the population)

Table 8 continued **- Proposed measurement methods for Time-related Dimension indicators**

INDICATORS BY DIMENSION	Measurement methods
Dynamics of objects	a. % Births t = (Births t / Total objects t) x 100 = (Births t / (Births t + Alive t) x 100 b. % Deaths t = (Deaths t / Total objects t) x 100 = (Deaths t /(Births t + Alive t) x 100 c. % Deaths t-1 = (Deaths t / Total objects t-1) x 100 = (Deaths t / (Alive t + Deaths t) x 100
Stability of variables	a. Use statistical data inspection methods to compare the values of specific variables for persistent objects in different deliveries of the source. Graphical methods that can be used are a bar plot and a scatter plot b. % of Changes = (Number of objects with a changed value / total number of persistent objects with a value filled in for the variable under study) x 100 c. A correlation statistical method can be used to determine to which extent values changed in the same direction for different object. For categorical data a method such as Cramer's V can be used

3. Application to a case-study: the SSD administrative source

3.1 Description of the source used for the case-study

Italian Social Security Data (SSD), used for this application of the input quality indicators, is produced by Inps (Italian Institute of Social Security) and concerns the monthly contribution declarations of employers for employees, as requested by the law of 24 November 2003, n. 326. The so-called Emens declarations are the means by which Inps collects pay data and information needed to computation of the social security contributions for each employee.

The choice of the SSD source for testing QRCA quality framework derives from several elements: it is a complex and big source including more statistical units connected to each other; due to its wealth of information it is suitable to be widely used in Istat within different statistical processes, both for the production of business and social statistics.

Particularly relevant is the central role that the SSD source has assumed in redefining the Istat production process of the Business Register (BR), Asia (Archivio statistico delle imprese attive) for the 2011 edition and in implementing innovative information about employment for the Census of Industry and Services 2011, mainly based on AD.

The SSD source covers data on the social security system for private employers as well as for other small subsets of public employers. Regarding the territorial reference, SSD includes social security contributions payable by employers resident in Italy.

Until 2010, Inps provided to Istat an annual dataset built on the basis of the monthly declarations and data referred to year *t* were delivered after 18 months.

From the supply of 2010 onwards, except for some previously test, the Istat interest has moved to the monthly version. Monthly data referred to year *t* are provided in two releases, on April *t+1* and on November *t+1* with an improvement of the timeliness (a maximum of 12 months for the final version). Data supply is very big, about 160 million records and 45 variables. For the application of the quality framework, measurement methods of quality indicators are computed on May 2010 data. This subset of the entire database contains about 13 million records and the same number of variables.

As already mentioned, more types of objects can be identified in SSD. Each record refers to the "Employee Tax Feature" defined as the set of variables useful to calculate the amount of social security contributions payable for each employee by the employer. Among these variables there are the Employer and Employee tax codes too, hence SSD source is a LEED dataset (Linked Employer - Employee Data). Other variables characterizing the Employee Tax Feature are the Type of employment contract (Fixed term/Permanent), the Contractual working time (Full/Part-time), the Professional status and the Type of contribution (tax relief for disabled or disadvantaged workers,…). The change of any of these characteristics during the month gives rise to a new record concerning the same contract. Due to this database building rule, the information is redundant.

Some units that can be derived from the Employee Tax Feature are: the Employee, the Employer and the Workplace (Municipality).

In Table 9, the administrative units are described and their identification key variable is reported. The Employee and Employer units are both identified through the Tax Code; concerning the Workplace, a code called "Belfiore" is used to identify the Italian Municipalities.

Table 9 - Administrative units in the SSD source

DEFINITION	Identification key
EMPLOYEE TAX FEATURE – primary unit The set of characteristics useful to define the amount of social security contributions payable for each employee by the employer	Complex key defined by a set of variables.
EMPLOYEE – derived unit A worker who has had at least one pay contributions to INPS as an employee during the month	Employee Tax Code.
EMPLOYER – derived unit Employer who have made at least a payment contributions for employees in May of 2010 or Employer who has employed at least one regular worker	Employer Tax Code.
WORKPLACE – derived unit Place where the work is mainly carried out	Belfiore Municipality Code.

Among the SSD administrative units there are several relationships:

a) an Employer may have more Employees;
b) an Employee can have more than one Tax Feature with the same Employer;
c) an Employee can have more than one Tax Feature with the more than one Employer;
d) a Municipality can be host for more Employees;
e) (a Municipality may not be a Workplace, in this case it is not recorded in the dataset).

The SSD source provides a wealth of information useful to describe the employment in enterprises. In addition to the main variables already mentioned for describing Employee Tax Feature, it contains the following variables: the number of paid days, the national collective agreement, the date and the reason for hiring, the date and the reason for termination and so on. The hiring date and the termination date are two events defined in the monthly data with the day of the month.

3.2 Working method and selection of the indicators

The results presented here on the SSD derive from the methodological application carried out by Istat that has been called, along with Statistics Netherlands and Statistics Sweden, to test the theoretical framework defined to evaluate AD quality within the BLUE-ETS WP8 (Daas et al., 2013).

The same results, however, were here taken up and examined in the attempt to respond to a different purpose, which is to determine the applicability of the QRCA into the Istat production processes, at the Data Hyperdimension level. We are not considering here the Source and the Metadata Hyperdimensions involving the AD acquisition task and Metadata analysis task already mentioned (Daas and Ossen, 2011).

The feasibility study for the production of the QRCA for SSD means to identify, first of all, which indicators and measurement methods included in the framework BLUE-ETS are actually applicable, useful and computable on each supply of SSD source, as soon as it is acquired by Istat. The aim, in this case, is to identify the set of information on the SSD quality that is possible to release in a timely manner in the moment in which the data are delivered to the Istat users of the source. Timeliness is a central element on which to focus, in particular for sources already included in the statistical processes, as is the case of the SSD. For these sources, in fact, the time that elapses between the acquisition (by the directorate in charge) and the delivery of data to users can be very short, and it is in this time that the BLUE-ETS indicators should be calculated.

The implementation of the quality indicators depends on two main aspects: the availability of the information requested for the measurement methods computation and the possible automatic IT procedures that should make the computation as timely as possible.

With respect to the second aspect, the activities are in progress in Istat.

This paper focuses instead on the first aspect and presents a preliminary analysis aimed at assessing the applicability of the indicators and related measurement methods verifying which information requested for the computation is actually available about SSD. In order to verify this, quality indicators have been divided between those referring to the entire dataset (Technical checks and some Time related Dimension indicators), and those applicable to selected units or variables (Integrability, Accuracy, Completeness and some Time related Dimension indicators). For the second group of indicators it was necessary to perform some preliminary actions:

- selection of objects in the SSD source to which it was possible and useful to calculate indicators;
- identification and acquisition of the reference statistical population to compare and match data (Integrability and Completeness Dimensions);
- selection of variables in the SSD source to which it was possible and useful to calculate indicators;
- identification of relevant edit rules to which it was possible and useful to calculate indicators (Accuracy Dimension).

The analysis carried out in the test phase (Daas et al., 2013) had already pointed out that problems of applicability do not occur for indicators whose calculation depends on information included directly in the dataset provided, while some issues emerge when the information requested is external to the supply and are primarily due to:

- the not comprehensive information given by the AD provider on the metadata and on the data generation process;
- the existence of the reference list (administrative target population or statistical target population) for the comparison of the units or variables.

Regarding the last point, if in the test case indicators of Comparability, Alignment, Undercoverage and Overcoverage were computed for Enterprise unit, using the BR as the reference statistical population, in this context they have been excluded for two reasons: one connected to timeliness issues and the other one to a more substantial restriction.

The supply of SSD used in this work was acquired on November 2011, while Asia register related to 2010 was made available on January 2012, in draft form, and on May 2012, in the final version. It is evident that the timeliness requirement for the computation of the indicators is not respected.

About the second reason, innovations introduced from 2011 in the Asia BR production made the SSD source part of the input sources used. This dependence relationship makes that it will never be possible to calculate the indicators mentioned using the Asia register as reference statistical population.

For Comparability and Alignment indicators it is essentially a practical problem and an approximate solution may be to use the BR of the previous year to evaluate the Integrability. The calculation was not carried out in this work. However, an analysis of Integrability was performed at the Metadata level on the Employer unit in the SSD source and the Enterprise unit in the Asia register (see Table 9 in the Section 3.3). Useful information about the Integrability of the SSD may be obtained by calculating the indicators also with respect to other AD sources, such as those that enter along with it in the Asia production process.

In the case of the Undercoverage and Overcoverage indicators, the dependence of the Asia register on SSD source poses, instead, a problem that is, not only practical, but also, and especially, of a conceptual nature. The independence between the AD source under evaluation and the reference statistical population should be an essential requirement when analyzing the coverage of an AD source in terms of input quality.

Table 10 lists the set of input quality indicators that can be calculated on the basis of the analysis carried out comparing the information requested for the computation with those actually available. For indicators related to objects, the type of object on which the indicator has been applied is reported. It is pointed out that whereas in SSD objects are repeated on multiple records (by construction), the indicators were calculated by properly counting the number of records or the number of units as specified.

Table 10 - Indicators applied to SSD

DIMENSION	Level	Indicator	Type of object			
			Employee	Employer	Municipality	Employee Tax Feature
Integrability	Object	Comparability of objects			x	
	Object	Alignment of objects			x	
	Variable	Linking variable				
Accuracy	Object	Authenticity	x	x		
	Object	Dubious objects	x			
	Variable	Inconsistent values				
	Variable	Dubious values				
Completeness	Object	Redundancy	x	x	x	x
	Variable	Missing values				
Time related Dimension	Dataset	Timeliness				
	Dataset	Punctuality				
	Objects	Dynamics of objects	x	x	x	

In the next Section results of the indicators related to the SSD are reported. For the implementation of the measurement methods we used the statistical software Sas. Initially we evaluated also the hypothesis of performing the processing in R, however, it was decided to use the Sas software because it is more suitable for processing large amounts of data as it is the case of SSD.

3.3 Case-study results

In this Section results of the quality indicators applied on the SSD administrative source are presented by quality Dimension.

Starting with the Integrability Dimension, we focused on Alignment and Comparability indicators, for objects, and on Linking variables indicator. With regard to the two indicators of objects a preliminary analysis at the Metadata Hyperdimension level is required and it is made comparing administrative concepts, already described (Table 9), with the statistical ones, as shown in the following Table 11. For the Employee and Employer units there is a similarity (*identical objects*) between administrative and statistical concepts and SSD identification variables are the same used by Istat (Tax Code). The units Employee Tax Feature and Municipality are not directly comparable with statistical units of interest. But, after a treatment process, it is possible to integrate them (*corresponding objects*). In particular, the Employee Tax Feature can be used for the identification of the statistical unit Employment relationship (or Contract of employment or Job), however, we do not have a statistical register for it. About the administrative unit of the Workplace it is possible to use an external table to harmonize administrative and statistical units.

Table 11 - Administrative units in the SSD source and Reference statistical units used in Istat

ADMINISTRATIVE OBJECTS		INTEGRABILITY LEVEL	STATISTICAL OBJECTS	
Definition	Identification key		Definition	Identification key
EMPLOYEE TAX FEATURE The set of characteristics useful to define the amount of social security contributions payable to INPS for each employee by the employer	Complex key defined by a set of variables	Corresponding ⟺	EMPLOYMENT RELATIONSHIP A formal agreement between an enterprise and a person, whereby the person works for the enterprise in return for remuneration	- (Employment relationships register not available)
EMPLOYEE – derived unit A worker for which there is at least one social security contribution paid to INPS as an employee during the month	Employee Tax Code	Identical ⟺	EMPLOYEE IN ENTERPRISE Person who works for an Enterprise on the basis of a contract of employment and receives compensation	- (Employees register not available)
EMPLOYER – derived unit Employer who has employed at least one regular worker	Employer Tax Code	Identical ⟺	ENTERPRISE Enterprise in Business Register ENTERPRISE WITH EMPLOYEES Enterprise with employment >0 in Business Register	Enterprise Tax Code
WORKPLACE - derived unit Place where the work is mainly carried out	Belfiore Municipality Code	Corresponding ⟺	Italian Municipality	Istat Municipality Code

As explained in Section 3.2, in the Data Hyperdimension, the Comparability and Alignment indicators for Employers are not reported, as the BR cannot be used in practice for the computation.

It is however possible to calculate the two indicators for the Workplace (Municipality), using the official Istat Municipalities Register[13]as reference statistical population. The list used is the one updated on the 1st of January 2011 when the official number of Italian municipalities was to 8,094 units.

As already said, the Municipality unit in SSD is not directly comparable with the statistical unit in Istat Register so it is classified as "Corresponding" (Table 11). However a table is available for the harmonization between the Municipality identification codes (named Belfiore Code) in SSD and the Istat Municipality Codes (foreign key). Administrative units are involved in (n : 1) with (n ≥1) relations with statistical units.

Comparability and Alignment indicators are presented in the following Table 12.

Table 12 - Comparability and Alignment of objects indicators

INDICATOR	Measurement method	Result
Comparability of objects	% of the SSD Municipalities corresponding to Istat Municipalities.	97,49%
Alignment of objects	% of the Istat Municipalities corresponding to SSD Municipalities.	99,62%

[13] The register is produced by Istat and updated twice a year (June 30 and December 31) on the basis of territorial and administrative changes that occurred in the country according to the Classification of territorial units for statistics (NUTS), adopted at the European level.

The results express the weight of the similarity between the two sources. In the Comparability indicator, this weight is calculated with respect to SSD, while in the Alignment indicator with respect to the Istat Register of Italian Municipalities. Users can draw their own conclusions: the degree of Comparability is good and the decoding table works fine (less than 3% of the Municipalities in SSD are not found in the Istat Register); with regard to the Alignment, it is possible to conclude that information is exhaustive as only few Istat Municipalities are not corresponding due to the fact that not all the Municipalities are workplaces (the comparison would have been perfect using the hypothetical Register of the Municipality that are places of employment).

The Linking variables indicator, reported in Table 13, gives information about the usability of units identification codes in SSD for integration with other micro data sources. It has been computed for Employee Tax Code, Employer Tax Code and Belfiore Municipality Code. The results show a high quality of the linkage variables.

Table 13 - Linking variable indicator

INDICATOR	Linking variable	Measurement method	Result
Linking variable	Employee Tax Code	% of records in SSD with missing value on the Employee Tax Code.	0,00015%
		% of records in SSD with syntactical incorrect value on the Employee Tax Code.	0%
	Employer Tax Code	% of records in SSD with missing value on the Employer Tax Code.	0%
		% of records in SSD with syntactical incorrect value on the Employer Tax Code.	0%
	Belfiore Municipality Code	% of records in SSD with missing value on the Belfiore Municipality Code.	0,06%
		% of records in SSD with Belfiore Municipality Code convertible to one used by Istat.	99,89%
		% of Municipalities in SSD with Belfiore Municipality Code convertible to one used by Istat.	97,49%

The indicators of the Accuracy Dimension allow to provide an assessment of the correctness of SSD both for objects and for variables. With regard to the objects, the Authenticity and the Dubious objects indicators are shown in the following Table 14. The first focuses on the legitimacy of objects and it is calculated for the Employee and Employer units checking the syntactic correctness of their identification codes. The result coincides with that previously reported for Linking variable indicator (see Table 13). This is an example of how the same measurement method can be used to evaluate two different aspects of the AD quality.

Table 14 - Accuracy of objects indicators

INDICATOR	Object	Measurement method	Result
Authenticity	Employee	% of records in SSD with syntactical incorrect value on the Employee Identification Code (Tax Code).	0%
	Employer	% of records in SSD with syntactical incorrect value on the Employer Identification Code.	0%
Dubious objects	*Employee (in relation to Employer)*	% of Employees in SSD which worked at more than 4 Employers.	0,0099%

Dubious objects indicator can be measured investigating the correctness of each object with respect to other types of objects in SSD. A soft rule can be defined to detect objects involved in implausible but not necessarily incorrect relations. In this application, we investigated the relation between the Employee unit and the Employer unit counting the number of attachments for each Employee with different Employers registered on May 2010. The distribution is reported in Table 15.

Table 15 - Distribution of Employees by number of Employers

EMPLOYERS	Employees
1	12.712.004
2	269.496
3	14.708
4	2.506
>= 5	1.283
Total	**12.999.997**

For the calculation of the indicator in Table 14, we considered the following soft rule for each Employee: *More than k "attachments" with different Employers during the month.* The value of the parameter k could be, for example, k = 4. The Dubious objects indicator provides the percentage of units that should be subjected to more accurate checks and inspections and possibly not considered in the statistical process if it is not possible to interpret the meaning of the relationship.

Concerning Accuracy of variables in SSD, we focused on the Inconsistent values and Dubious values indicators. For the calculation, a set of checking rules – respectively, hard and soft rules – should be defined and applied to the variables in the dataset. The rules here examined are to be considered only by way of exercise. The overall definition of these rules requires a thorough knowledge of data and therefore the involvement of the Istat researchers using SSD. Table 16 shows some results of the two indicators applied to the dataset, reporting the percentage of records for which each rule is violated.

Table 16 - Accuracy of variables indicators

INDICATOR	Measurement method	Rule	Result
Inconsistent values	% of records in SSD of which values (or combination of values) for variables are involved in non-logical relations.	*Hard rule* Full-time employment and zero part-time percentage.	0,11%
Dubious values	% of records in SSD of which values (or combination of values) for variables are involved implausible but not necessarily incorrect relations.	*Soft rules* Employee age <= 65.	0,37%

Focusing on the Completeness Dimension, on the basis of available information in SSD, Redundancy and Missing value indicators are the only indicators for which it is possible to meet the timeliness and independence criteria adopted.

The Redundancy has been measured for the different types of objects detecting duplicates for the respective identification codes. For the Employee Tax Feature unit, a multiple identification code is assumed considering the following set of variables:

Employee Tax Code, Employer Tax Code, Professional status, Contractual working time, Type of employment contract, Type of contribution. A last Redundancy indicator has been calculated also to check the occurrence of multiple records, with the same values for all variables. As shown in Table 17, in SSD there are no duplicated records for the entire set of variables, while high percentages of duplicates are found for objects. It should be noted that while the presence of duplicated records with the same values for all variables has to be evaluated as an error and detects problems of data quality, the presence of duplicates on the identification codes for objects is admissible and it depends on the mechanism of data generation (see § 3.1).

Table 17 - Redundancy indicator

INDICATOR	Object	Measurement method	Result
Redundancy	Employee	% of records in SSD duplicated for Employee Tax Code.	2,81%
	Employer	% of records in SSD duplicated for Employer Tax Code.	88,72%
	Municipality	% of records in SSD duplicated for Municipality Belfiore Code.	99,93%
	Employee Tax Feature	% of records in SSD duplicated for Employee Tax Feature multiple identification code.	0,47%
	-	% of records in SSD duplicated for all variables.	0%

Regarding the Completeness of variables, Missing values indicator has been calculated counting the number of records with missing values for the main SSD variables. This indicator can be implemented easily and in a timely manner, as it requires no additional information other than that contained in the dataset itself. As reported in Table 18, the percentage of missing values is equal, or very close, to zero for all the variables considered. For the computation, the first step is to verify for each variable what 'value' in the dataset is used to indicate a missing item and to distinguish items for which a value it is not expected. In some cases, the latter can be identified in association with the value of the corresponding possible filter variable. In SSD, this happens for the variables: Hiring reason, Job contract termination reason and Part-time percentage. In particular, for the Hiring reason and the Job contract termination reason a value is expected if the respective date is "active" while for the Part-time percentage a value is expected if the Contractual working time is equal to 'Part-time'. It can be useful to evaluate the presence of missing values considering more variables simultaneously. In this application, we calculated the percentage of records with all missing values for the set of variables, already considered in the Redundancy indicator, that it is expected to be the multiple identification code of the Employee Tax Feature.

A graphical representation of the number of missing values may be useful in preparing the quality report.

Table 18 - Missing values indicator

INDICATOR	Measurement method	Variable	Result
Missing values	% of records in SSD with missing value for a particular variable.	Professional status.	0%
		Contractual working time.	0,03%
		Type of employment contract.	0,03%
		Type of contribution.	0%
		Hiring date.	0%
		Hiring reason.	0%
		Job contract termination date.	0%
		Job contract termination reason.	0%
		Part-time percentage.	0%
	% of records in SSD with all missing for a set of variables.	Employee Tax Feature multiple identification code.	0%

With regard to the Time-related Dimension, in this application we considered the Timeliness and Punctuality indicators, referred to entire dataset, and the Dynamics of objects indicator.

Timeliness and Punctuality indicators have been calculated with the aim to measure, respectively:

- the time difference (days) between the date of receipt by Istat and the end of the reference period.
- the time difference (days) between the date of receipt by Istat and the date of receipt agreed upon, as defined in the agreement with the AD provider.

Results of both indicators (Table 19) point out a good quality of SSD and their possible use in the statistical production processes in a timely manner. The Punctuality indicator, assuming a negative value, shows that data were delivered before the receipt date specified in the official request.

Table 19 - Timeliness and Punctuality indicators

INDICATOR	Measurement method	Result (days)
Timeliness	Time difference (days) between the date of receipt of SSD by Istat and the end of the reference period.	365
Punctuality	Time difference (days) between the date of receipt of SSDby Istat and the date of receipt agreed upon, as laid down in the contract.	-31

The Dynamics of objects indicator gives information about changes over time of the populations present in SSD. As the dataset contains monthly data, it was considered useful to point out the dynamics between two consecutive months of the same provision, comparing objects in the months of April 2010 (t-1) and May 2010 (t). In Table 20, results are provided both for Employers and for Employees, in a longitudinal perspective, performing a microdata record linkage between the two monthly datasets. For the Employees, the Dynamics indicator is equal to 3% if we consider the “new workers” and to 2.3% referring to “old workers”. For Employers, values are lower (2.5% and 1.8%, respectively) showing a more limited dynamics.

Table 20 - Dynamics of objects indicator

INDICATOR	Object	Measurement method	Result
Dynamics of objects	*Employee*	% of Employees present on May 2010 but not on April 2010 (new Employees) compared to the total number of Employees on May 2010.	3,0%
		% of Employees present on April 2010 but not on May 2010 (old Employees) compared to the total number of Employees on April 2010.	2,3%
	Employer	% of Employer present on May 2010 but not on April 2010 (new Employer) compared to the total number of Employers on May 2010.	2,5%
		% of Employer present on April 2010 but not on May 2010 (old Employer) compared to the total number of Employers on April 2010.	1,8%

It should be noted that the interpretation of these results is not directly connected to a quality evaluation, as a certain population dynamics is a characteristic of all phenomena. In the case of SSD, it is connected to demographic events for Employer unit and to hiring / termination of contract for Employee. However, the availability of the indicator values of each supply in time series together with possible reference value or benchmark could be very useful. For example deviations from an average or a trend value can detect the presence of possible errors in the supply analyzed.

4. Further development

The conceptual scheme just described and experimented on SSD should be implemented in an efficient manner as the evaluation of the statistical quality of the AD plays a crucial role in the new Istat statistical production process involving the use of AD. In order for the application to be effective, the standardized tools implementing quality indicators have to meet the following requirements:

- produce documentation of quality assessment in a timely manner;
- provide information as completely as possible;
- provide general information to AD users regardless of the domain of the produced statistics;
- be concise and easy to read.

In particular, there are three tasks that the QRCA can perform addressed to different types of users within NSIs.

The first one is “Evaluating AD statistical usability” for the new potential users of AD already acquired, or for AD sources acquired for the first time. For new potential users, the QRCA enclosed to the supplied AD, and together with the Metadata level description, could provide a useful support for evaluating whether to introduce or not AD into the production process. For new AD, after the preliminary metadata analysis performed using the Checklist proposed by Daas and Ossen (Daas and Ossen, 2011), the exploratory analysis provided by the QRCA at the Data level can give the information needed to support the final decision to include or not the source in the statistics production.

The second task of the QRCA is "Monitoring AD quality" already in use in NSI, for current users. This task is of primary importance because the production processes can develop a strong reliance on AD and a tool should be developed to promptly deal with possible problems. In particular it is necessary to constantly monitor AD quality for two main reasons: a) their statistical use is secondary and regulatory changes can produce significant breaks in the periodical deliveries and may impact the statistics production process; b) before the data are introduced in the production process, a check procedure should be performed to make sure that there are no unexpected statistical errors.

The last QRCA task is "Monitoring quality in the AD acquisition process" that is to check whether data received are consistent with the requests and to support the process of data loading. Where appropriate, it is useful to define alert or warning to optimize the timing of data acquisition and release to internal users.

The quality evaluation results of the AD supplies in a time perspective can also provide interesting elements to evaluate the effectiveness of possible harmonization processes between administrative and statistical concepts agreed with the AD producers and the NSI such as: shared use of standardized classifications, changes in the process of recording data and so on.

The next challenge for Istat is how to plan and implement the quality reporting activity achieving the objectives defined and taking into account the limited resources available.
From the organizational point of view, Istat determined that the acquisition of AD should generally be made at a central level. A central organizational office, named ADA (AD acquisition and integration), is in charge of acquiring AD responding to almost all the institute AD requests. In 2013, Istat acquired about 250 supplies from more than 100 Administrative sources, so a strong coordination among departments using AD has been set up in order to plan the activities and meet the needs of the whole production process.

Recently, in order to avoid duplicate work among AD source users, this office also is building a new integrated system, called SIM (Integrated System of Microdata), which has the task to store AD supplies and to perform data pre-processing. In particular data received are coded with respect to official classification, when possible, and integrated using unique codes for the same objects in SIM. Currently unique codes are assigned to individuals and economic units. A Metadata repository is currently also under development. Of course, all operations are in compliance with the rules on data security and privacy.

The AD quality evaluation in SIM is the further task for ADA and this is another important function for Istat AD source users. From this point of view a standardized and generalized QRCA could be a support to share information defining usability of an administrative source and to monitor the quality of AD received by Istat (Di Bella and Ambroselli, 2014).

With the purpose of complying the appropriate timeliness, a system that allows to make the AD quality evaluation as automated as possible is being planning. Interesting results derived from the possible use of some statistical packages available in R (Tennekes et al. 2013). At this moment, in Istat, the implementation of the QRCA is undergoing testing on some education AD in SIM. The strategy aims to take advantage of all the available metadata, that is to make metadata "active" to the greatest extent possible for supporting the QRCA production process[14].

[14] Following the Core principles for metadata management (Common Metadata Framework Part A: Statistical Metadata in a Corporate Context), UNECE / Eurostat /OECD Group on Statistical Metadata (METIS) http://www.unece.org/stats/cmf/.

For the implementation of Source Hyperdimension quality indicators we'll try to take advantage of all the information used to manage the AD acquisition process.

In the Metadata Hyperdimension, we are experimenting some ways of interacting with the AD provider in order to acquire, together with data, also updated metadata necessary for their correct interpretation. In addition, the phase of Entity Relationship analysis of the administrative dataset and the consequent data loading in the relational database, can allow us to automatically identify the set of objects /entities to be evaluated.

The process of assigning an unique code to the same objects in SIM can provide information for the implementation of Comparability indicators of the objects in the Metadata Hyperdimension with respect to statistical units mapped in the system of data dissemination. In this case it could be possible to define equivalence classes for type of objects defined at different levels (i.e. individual-student, economic unit-enterprise).

In the Data Hyperdimension, a suitable description of the process of assigning unique codes can support the calculation of quality indicators for evaluating the record linkage procedure: some useful measurement methods can be derived for the linking variable quality indicator. It is important to underline that this is a core indicator not only for the Integrability Dimension evaluation but it also assumes a significant role for other quality indicators, such as Coverage and Dynamic of objects indicators.

A last example of making metadata "active": the coding phase of the territorial units using the official classification in SIM, can produce Comparability indicators for the classification variable in the Data Hyperdimension.

5. Concluding remarks

The AD quality evaluation is a necessity for the statistical production processes and the QRCA is a useful summary, documentation and sharing tool.

The framework for describing the AD quality adopted has proved very robust in the different applications carried out (Daas et al., 2013) and it seems to be a comprehensive instrument including the many facets of the concept of AD quality with respect to their statistical usability. The ability to implement such a tool envisaging inter-operability of processes is interesting. From the first results of the implementation procedure, it follows that some indicators can be calculated automatically using the metadata process, while for other indicators, such as indicators of consistency checks (Accuracy of variables) it is necessary the source users contribution to define the check rules or, in case of first usability analysis, a collaboration with the team who is in charge of analysing the source for the first time.

The implementation activities are proceeding steps by steps and depending on the resources available, it will be possible to image a full or partial implementation of QRCA for AD in SIM. At the same time, "AD Istat user groups" are setting up for the most important data source holders (Tax administration, Social Security Institute, Ministry of Education, Universities and Research) in order to verify the possibility of sharing information or more specific analysis possibly useful to most users. In any case it will be important to spread the framework of the QRCA tool in order to standardize as much as possible the AD quality assessment process.

References

Bakker B. (2010). *Micro-integration: State of the Art*. Paper for the Joint UNECE/Eurostat Expert Group Meeting on Register-Based Censuses, The Hague, The Netherlands.

Burger J., Davies J., Lewis D., van Delden A., Daas P.J.H. and Frost J.M. (2011). *Guidance on the accuracy of mixed-source statistics*. Deliverable 2011/6.3 of Workpackage 6 of the ESSnet on Admin Data. http://essnet.admindata.eu/WikiEntity?objectId=5452

Costanzo L., Di Bella G., Hargreaves E., Pereira H., Rodrigues S. (2011) An Overview of the Use of Administrative Data for Business Statistics in Europe, 58th World Statistics Congress of the International Statistical Institute, Dublin, August 21 – 26, 2011. http://2011.isiproceedings.org/papers/950391.pdf

Cramer H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, USA, p282.

Daas P.J.H., Ossen S.J.L., Vis-Visschers R.J.W.M. and Arend-Toth J. (2009). *Checklist for the Quality evaluation of AD Sources*. Discussion paper 09042, Statistics Netherlands.

Daas, P.J.H., Ossen, S.J.L. (2011) *Metadata Quality Evaluation of Secondary Data Sources*. International Journal for Quality Research, 5 (2), 57-66. http://www.pietdaas.nl/beta/pubs/pubs/IJQR2011.pdf

Daas P.J.H., Ossen S., Tennekes M. (CBS, Netherlands), Zhang L. C, Hendriks C., Foldal Haugen K. (SSB, Norway), Bernardi A., Cerroni F. (ISTAT, Italy), Laitila T., Wallgren A., Wallgren B., (SCB, Sweden) (2011a). *List of quality groups and indicators identified for administrative data sources*, Deliverable 4.1 of Workpackage 4 of the BLUE-ETS project. http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.1.pdf

Daas P.J.H., Ossen S., Tennekes M., Zhang L. C. (CBS, Netherlands), Hendriks C., Foldal Haugen K. (SSB, Norway), Cerroni F., Di Bella G. (ISTAT, Italy), Laitila T., Wallgren A. and Wallgren B. (SCB, Sweden) (2011b). *Reports on methods preferred for the quality indicators of administrative data sources*, Deliverable 4.2 of Workpackage 4 of the BLUE-ETS project. http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable4.2.pdf

Daas P.J.H., Tennekes M., Ossen S., Burger J., (CBS, Netherlands), Di Bella G., Galiè L., Bonardo D., Cerroni F., Talucci V., (ISTAT, Italy), Laitila T., Lennartsson D., Nilsson R., Wallgren A., Wallgren B. (SCB, Sweden), Hendriks C., Zhang L.C. and Foldal Haugen K. (SSB, Norway) (2013). *Guidelines on the use of the prototype of the computerized version of the QRCA, and Report on the overall evaluation results.* Deliverable 8.2 of Workpackage 8 of the BLUE-ETS project. http://www.blue-ets.istat.it/fileadmin/deliverables/Deliverable8.2.pdf http://essnet.admindata.eu/WikiEntity?objectId=5452

Di Bella G., Ambroselli S. (2014). *Towards a more efficient system of administrative data management and quality evaluation to support statistics production in Istat*, paper presented at the European Conference on Quality in Official Statistics (Q2014) held Vienna, 3-5 June 2014.

ESSnet AdminData (2013a). *Final list of quality indicators and associated guidance.* Deliverable of Workpackage 6 of the ESSnet on Admin Data. http://essnet.admindata.eu/WikiEntity?objectId=5452

ESSnet AdminData (2013b). *Overview of Existing Practices in the Uses of Administrative Data for Producing Business Statistics in EU and EFTA* (Database Tables + Reference Library, update 31/12/2012). Deliverable 1.2 of Workpackage 1 of the ESSnet on Admin Data. http://cros-portal.eu/content/overview-existing-practices-uses-administrative-data-producing-business-statistics-eu-and

ESSnet AdminData (2013c). *Admin Data Glossary. Definitions adopted for certain terms related to the use of administrative data for producing business statistics*. Deliverable 2013/1.1 of the ESSnet on Admin Data. http://www.cros-portal.eu/sites/default/files//SGA%202011_Deliverable_1.1.pdf

Eurostat (2003a). Definition of quality in statistics. Working group Assessment of quality in statistics, Luxembourg, 2-3 October 2003. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/ess%20quality%20definition.pdf

Eurostat (2003b). *Standard Quality Report.* Methodological Documents, Working Group *Assessment of quality in statistics*, Luxembourg, 2-3 October 2003. http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/documents/STANDARD_QUALITY_REPORT_0.pdf

Eurostat (2005). *European Statistics Code of Practice for the National and Community Statistical Authorities* - revised edition 2011. Adopted by the Statistical Programme Committee on 28th September 2011. http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/publication?p_product_code=KS-32-11-955

Frost J.M., Green S., Pereira H., Rodrigues S., Chumbau A. and Mendes J. (2010). *Development of quality indicators for business statistics involving administrative data.* Paper presented at the Q2010 European Conference on Quality in Official Statistics. Helsinki, Finland.

Hox J. J. and Boeije H. R. (2005). *Data Collection Primary vs. Secondary.* Encyclopedia of Social Measurement. http://joophox.net/publist/ESM_DCOL05.pdf

Laitila T., Wallgren A. and Wallgren B. (2011). *Quality Assessemnt of administrative Data*. Research and Development – Methodology reports from Statistics Sweden, 2, 2011.

Schouten, B., Cobben, F., Bethlehem, J. (2009). *Indicators for the representativeness of survey response.* Survey Methodology, 35 (1), 101-113.

Tennekes M., de Jonge E., and Daas P.J.H. (2011). *Visual profiling of Large Statistical Datasets*. Paper for the 2011 New Techniques and Technologies for Statistics Conference, Brussels, Belgium.

Tennekes M., de Jonge E., and Daas P.J.H. (2013). *Visualizing and Inspecting Large Datasets with Tableplots*, Journal of Data Science 11 (1), 43-58.

Unece (2007) Register-based statistics in the Nordic countries. Review of best practices with focus on population and social statistics, United Nation Publication, Geneva, 2007.

Wallgren A. and Wallgren B. (2007). *Register-based Statistics: Administrative Data for Statistical Purposes*. John Wiley & Sons, Chichester, UK.

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

Norme redazionali

La Rivista di statistica ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche corredati da una nota informativa dell'autore contenente attività, qualifica, indirizzo, recapiti e autorizzazione alla pubblicazione. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di due referenti scelti tra gli esperti dei diversi temi affrontati.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file RSU stili o nella classe LaTex, entrambi disponibili on line. La lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 35 pagine. Una volta che il lavoro abbia superato il vaglio per la pubblicazione, gli autori sono tenuti ad allegare in formato originale tavole e grafici presenti nel contributo, al fine di facilitare l'iter di impaginazione e stampa. Per gli standard da adottare nella stesura della bibliografia si rimanda alle indicazioni presenti nel file on line.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 120 parole); quelli in italiano dovranno prevedere anche un abstract in inglese.

Nel testo dovrà essere di norma utilizzato il corsivo per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

Per contattare la redazione o per inviare lavori: rivista@istat.it. Oppure scrivere a:
Segreteria del Comitato di redazione delle pubblicazioni scientifiche
all'attenzione di Gilda Sonetti
Istat
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012014002000000

ISSN 1828-1982

€ 15,00