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