Results Of A Project On Record Linkage, Statistical Matching And Micro Integration: The ESSnet On Data Integration

Scanu, Mauro
Istat, DPTS/DCET
via Cesare Balbo, 16
Rome (00152), Italy
E-mail: scanu@istat.it

The ESSnet on Data Integration (ESSnet – DI)[1] focuses on statistical methodologies for data integration and on the statistical aspects to be considered for ensuring the usability of the integrated data sets. These are the statistical methodologies under investigation:

- Record linkage: complete records at unit level are obtained by fusing records of two or more data sets with appropriate unit identifiers. Within this setting, two broad groups of methods can be considered: deterministic record linkage (including exact linkage with the use of an identifier); probabilistic record linkage.
- Statistical matching: complete (synthetic) records at unit level are obtained with appropriate imputation procedures, whereby the data sets to be integrated play the role of donor and recipient files respectively.

Once an integrated data set has been produced, it may be appropriate to use actions that ensure better quality of the matched results. These actions are usually called:

- Micro integration processing: quality and timeliness of the matched files; defining checks; editing procedures to obtain better estimates; imputation procedures to obtain better estimates; weighting (to population totals) issues of matched files.

For record linkage, statistical matching, and micro integration processing, the project main goals are:

1. to develop and organise common knowledge;
2. to develop methods in specific sub domains;
3. to provide users with tools for their use;
4. to foster knowledge transfer;
5. to build and maintain sustainable capacity.

A former ESSnet (ESSnet on Statistical Methodology: area Integration of Surveys and Administrative Data) was active between December 2006 and June 2008. Results can be found in the webpage: http://cenex-isad.istat.it. The results of the new project can be found in the ESSnet portal: http://www.essnet-

[1] The project is composed by people working in the following National Statistical Institutes: ISTAT – Italy (Mauro Scanu, Tiziana Tuoto, Monica Scannapieco, Cristina Casciano, Nicoletta Cibella, Paolo Consolini, Marco Di Zio, Marcello D'Orazio, Marco Fortini, Daniela Ichim, Filippo Oropallo, Laura Peci, Francesca Romana Pogelli, Giovanni Seri, Luca Valentino), CBS - Netherlands (Jeroen Pannekoek, Arnout van Delden, Bart Bakker, Paul Knottnerus, Léander Kuijvenhoven, Frank Linder, Nino Mushkudiani, Dominique van Roon, Eric Schulte Nordholt), SFSO - Switzerland (Jean-Pierre Renfer, Daniel Kilchmann), GUS Statistical Office in Poznań - Poland: Marcin Szymkowiak, Adam Ambroziak, Dehnel Grazyna,Tomasz Józefowski, Tomasz Klimanek, Jacek Kowalewski, Ewa Kowalka, Andrzej Mlodak, Artur Owczarkowski, Jan Paradysz, Wojciech Roszka, Pietrzak Beata Rynarzewska, Magdalena Zakrzewska), INE - Spain (Francisco Hernandez Jimenez, Gervasio-Luís Fernández Trasobares, Miguel Guigó Pérez), SSB - Norway (Johan Fosen, Li-Chun Zhang).

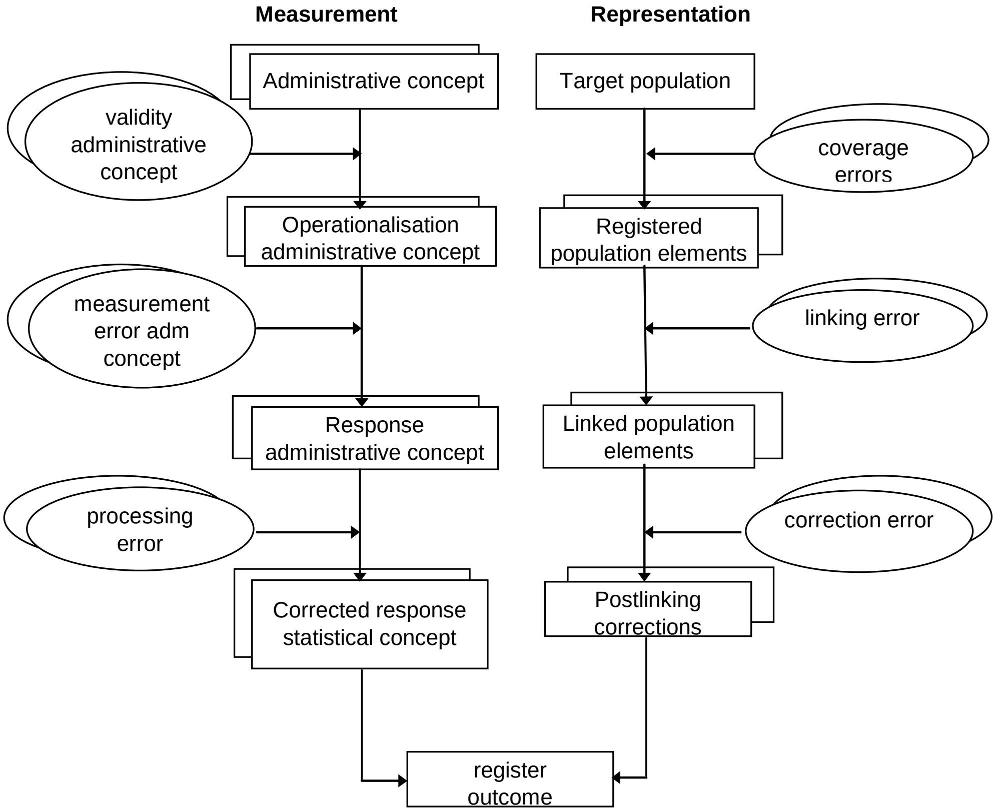
portal.eu/project-information/data-integration.

State-of-the-art

An updated state-of-the-art on record linkage and statistical matching has been made available focusing mainly on the new papers appeared in the period 2008-2011. For this reason record linkage state-of-the-art focuses mainly on comparison functions, blocking criteria, and alternative classification techniques with respect to the traditional Fellegi-Sunter model (Fellegi and Sunter, 1969), including Bayesian procedures and support vector machines. For statistical matching, interest has been focused mainly on how to manage the possible complex survey designs used for drawing samples, the uncertainty affecting statistical matching results (mainly the fact that estimates of parameters on the distribution of never jointly observed variables are sought), and the use of nonparametric procedures, instead of the assumption of parametric models for the variables at hand (usually as normal or multinomial variables). For both record linkage and statistical matching, particular attention has been given to areas that resemble the problems tackled in the two areas. Significant connections have been discovered with the methods used in confidentiality, and (only for statistical matching) for those used in ecological inference (King, 1997).

There was not so much available literature on micro integration. Instead of a state-of-the-art document, project members tackled the problem of defining a theoretical framework (Bakker, 2010). Micro integration can be defined as the method aiming at improving data quality by searching and correcting (measurement and representation) errors affecting data from the integrated use of administrative sources and surveys. These issues have been distinguished: completion, harmonization, and correction for the remaining measurement errors. In Bakker (2010) the different kinds of errors are defined, examples of these errors from the daily practice (from the Social Statistical Database and the Virtual Census) are provided and operating procedures to correct them are proposed. Space has been given to consistent repeated weighting, a method that can be used for consistent estimation if one combines register data with sample survey data. Finally, the position of micro-integration in the total statistical process has been described.

Figure 1- The 'life cycle' and errors in a combined register situation


Measurement
Representation
Administrative concept
Target population
validity administrative concept
coverage errors
Operationalisation administrative concept
Registered population elements
measurement error adm concept
linking error
Response administrative concept
Linked population elements
processing error
correction error
Corrected response statistical concept
Postlinking corrections
register outcome

The description of micro integration is essentially based on a frame for errors in statistics obtained combining sources, as in Figure 1. This description has led to a comparison with the current application of data integration in the different statistical institutes, leading to different case studies, reported in a different workpackage (WP4).

Methodological developments

One of the main objectives of this project is to allow application of statistical methods for data integration in each NSI. For this reason, a number of targeted methodological developments was identified and tackled.

For record linkage, the attention has been given to the possibility of application of Bayesian procedures. The research focuses mainly on the possibility to get quality evaluations directly from the applied method. Furthermore, attention has been given to the statistical analysis performed on linked data sets, and on the use of models that take into account the probability of errors. The main reference for this problem is Tancredi and Liseo (2010).

For statistical matching, the problem that makes difficult the application of these methodologies in NSIs is the complex sampling design that characterizes most of the sample surveys. In the statistical matching literature there are at least three statistical matching methods that have been defined: file concatenation (Rubin, 1986), use of calibration procedures (Renssen, 1997), and maximum empirical likelihood inference (Wu,, 2004). We performed a comparison between these approaches, in the context of absence of external information on the variables that are never jointly observed. First results indicate that there is not a preference in terms of the quality of the obtained results. Their application depends heavily on the context and on the characteristics of the available information.

For micro integration, different topics have been considered:

- Methods for variance estimation (based on bootstrapping) for estimates based on a combination of one administrative source with a survey.
- Methodology to detect and correct linkage errors, where linkage errors comprise both incorrect positive or incorrect missing links.
- Applicability of methods for linkage of micro-data in combination of linkage at aggregated level, using a study on the creation of labour market statistics as a motivating example.
- Definition of an algorithm that makes survey data consistent with linked administrative source information on the same variables.
- Imputation and other methods for handling incoherencies caused by differences in units, when units in the different sources are at different levels.

Preliminary results on these issues will be available in July 2011.

Software tools

In order to make these methods applicable in the NSIs, it is important to provide appropriate software tools. In the former project (Cenex on ISAD) a review of the software tools for record linkage and statistical matching was prepared. Two of these tools (RELAIS for record linkage and StatMatch for statistical matching) have been created by staff working in ISTAT (Italian National Institute of Statistics), and they include functionalities useful for NSIs. This project will develop these tools extending the existing library of applications for data integration, supported by appropriate documentation.

Record linkage - In the context of a record linkage project, the pre-processing phase is an important activity that typically requires a lot of human effort and is extremely time-consuming. Hence, we plan to design and develop a software application to automate this phase as much as possible. Such an application will be part of the record linkage system RELAIS that currently implements both deterministic and probabilistic record linkage (http://www.istat.it/strumenti/metodi/software/analisi_dati/relais/).

Statistical matching - An R package for statistical matching named StatMatch has been already developed in Istat. This package includes a set of methods based on donor imputation and on maximum likelihood estimation. As a project activity, this package has been extended in order to include many functionalities:

- new functions that perform statistical matching when dealing with data sources originating from complex sample surveys, i.e. the typical samples available in NSIs (see the section on methodological developments and references therein);
- a function that allows the evaluation of uncertainty in the matching process when dealing with categorical variables. More precisely, it will be possible to compute the bounds for each relative frequency in a contingency table of two variables that are never jointly observed in the original data sources by using the Frechét bounds (Di Zio, 2010).

All the new functionalities are documented with help pages written in English in accordance with the R guidelines for the development of new packages. The help pages contain some short examples. A short "vignette" is also compiled in order to provide a detailed guideline for the application of the new functionalities to the real applications of statistical matching. The package StatMatch, as all R packages, is an open source package that is already available to all users via the R-project web-pages, with a GPL licence.

Case studies

Micro integration is a very common problem in data integration that has not been formalized and described in detail. In this project, an attempt to formalize micro integration has been tackled in the state-of-the-art. Practical application with real life examples are considered extremely important. For this reason, great attention has been posed on case studies.

A first case study has been developed by Fosen and Zhang: how to measure the quality of estimates for a register variable that is constructed by micro integration. This case study considers the Norwegian register-based employment estimates. Making assessments of the quality of a variable made by data integration is very interesting since data integration means that the inconsistence between different administrative data is dealt with through the data integration process, and it is important to afterwards evaluate the quality of this process. The data set used in this context is a random sample from a register containing one record for each person living in Norway (it would be inconvenient to use the complete register as data set since it is very large). All variables on the data set will be register variables or register-based variables. One of the variables will be the register-based employment status which has been constructed by data integration from several administrative registers. In addition to administrative register information used in the data integration, there will also be additional information that can help us in assessing quality.

A second case study has been developed by Linder and van Roon: a method for estimating the variable educational attainment from a combination of administrative sources and surveys. In this case, the following data sets are combined to form the educational attainment database: the Labour Force Survey (LFS) and several administrative data sets, such as the Pupil Number Registers for secondary education, the Central Register for Enrolment in Higher Education and the Exam Results Register. A quality-check confirms that most of these sources are appropriate for this purpose. An important advantage of combining these sources of information is that in general, estimates on education level are more reliable than those exclusively based on the LFS, in particular when smaller populations are involved. As education registers do not cover the entire population, the LFS still plays an important part in filling the gaps. Most older citizens, for example, have completed their education prior to the administrations. For this part of the population, the use of the LFS data source is indispensable. The innovating aspect in this micro-integration approach is that data from registers and sample surveys are combined to produce one single variable. Statistics Netherlands has a great deal of relevant experience with the combination of administrative sources and sample surveys. However, so far data for one single variable originated from either registers or sample surveys, never from both sources at the same time. With the new approach some thorny methodological issues need to be solved, such as the question of how to deal with out-of-date information on education and the weighting mechanism.

A third case study has been developed by Pogelli and Oropallo: the framework of error in an integrated business survey in Istat. In a general plan towards a modernization of the structural business statistics, also the Italian NSI has decided to intensify the use of administrative data with the aim to reduce the statistical burden on enterprises and to improve the statistical quality of surveys, in terms of comparability with other sources and reduction of non response bias. In this context a new integration process, concerning the sample survey on Small and Medium Enterprises (SME), has been developed with the use of all available administrative sources that have relevant economic information for a business survey. This case study focuses the attention on the methodological framework of errors for register based statistics, proposed by Bakker (2010) and illustrated in Figure 1, trying to classify the tasks of the integration process between survey data and administrative data regarding to the SME survey, for the year of reference 2007.

Dissemination

The main objective of the project is the dissemination of the statistical methods of data integration in the European Statistical System. For this reason, different dissemination tools have been considered.

Course – it will be a three-day course in September 2011. It will cover the areas of record linkage (first day), statistical matching (second day) and micro integration processing (third day). Each day will be divided into three modules: description of methods; description of applications; description of software tools (for record linkage and statistical matching). It will be held in Rome by researchers from Istat and CBS.

Final workshop - Following the success of the ESSnet - ISAD workshop held in Vienna (29-30 May 2008), this project includes a workshop open to other EU member participants whose aim is:

a) dissemination of the ESSnet - DI results to the workshop attendees, with room for discussion on the reports and results obtained during the project;

b) presentations by researchers in the ESS on results, methods, problems in the area Data Integration.

The workshop will be held in Madrid, at INE premises (24-25 November 2011). The workshop will also be open to research projects in those areas (e.g. disclosure control, imputation and editing, sampling, archive based statistics) that may need data integration. More details are available on the workshop webpage: http://www.ine.es/e/essnetdi_ws2011.html.

Training on the job – The project organizes three on-the-job training courses. A call was launched in April 2010, and 7 countries asked for holding a training course in their premises. These are the training courses performed so far:

- 20-22 October 2010: On-the-job training on statistical matching, Statistical Office in Poznan (Poland)
- 25-28 January 2011: On-the-job training on record linkage, Office for National Statistics (Southampton, UK)
- 5-8 July 2011: On-the-job training on record linkage, Central Statistical Bureau of Latvia (Riga, Latvia)

Project web page – The project output is available on a project webpage in the ESSnet portal: http://www.essnet-portal.eu/. This webpage includes:

a) a repository of documents related to data integration (WP1 result)

b) tools useful for the application of data integration methodologies (booklet of prerequisites and course slides; presentations, abstracts and papers presented in the workshop; links to software web pages),

c) web-pages on the topic in the Internet,

d) the ESSnet - ISAD project webpage.

Conclusions

In the last years, data integration assumed a major role in the production of statistics in National Statistical Institutes. Nonetheless, there is a lack of a community of statisticians working in this area. The lack of such a community seems to be strictly linked to the way statistical offices are organized: there are centralized offices for many methodological areas (survey sampling, confidentiality, editing and imputation, …). On the contrary, data integration is rarely centralized. This is not necessarily a bad thing: people with a different background can face the same problem. The risk is that knowledge gained from some people in performing data integration is not transferred to other parts of the same institute. This project has the ambitious goal to establish contacts among researchers and practitioners at European level. It would be very appealing to think about sustainable actions for a worldwide community of practitioners on the topic of data integration.

REFERENCES (RÉFERENCES)

Bakker B. F. M. (2010) Micro-Integration: State of the art. In ESSnet – DI (2010). *State-of-the-art on statistical methodologies for data integration*, Chapter 5.

Di Zio M. (2010). Uncertainty in statistical matching. In ESSnet – DI (2010). *State-of-the-art on statistical methodologies for data integration*, Section 2.3.

ESSnet – DI (2010). *State-of-the-art on statistical methodologies for data integration.* Draft report, available at http://www.essnet-portal.eu/project-information/data-integration.

Fellegi I.P. and Sunter A.B. (1969) A theory for record linkage. *Journal of the American Statistical Association*, **64**, 1183-1210.

King, G. (1997). *A solution to Ecological Inference Problem.* Princeton: Princeton University Press.

Renssen, R.H. (1998) Use of statistical matching techniques in calibration estimation. *Survey Methodology*, **24**, 171-183

Rubin, D.B. (1986) Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.*, **4**, 87-94

Tancredi A. and Liseo B. (2010). A hierarchical Bayesian approach to record linkage and size population problems. *Annals of Applied Statistics*, to appear.

Wu, C. (2004) Combining information from multiple surveys through the empirical likelihood method. *Can. J. Stat*., **32**, 1-12