

**UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (ii): Managing and supporting changes related to editing and imputation

Managing changes in the E&I strategy of the Italian SBS

Prepared by Orietta Luzi, Italian National Statistical Institute, Italy

I. Introduction

1. Italian structural business statistics (SBS in the following) are currently subject to a transition period that will generate in the near future a major shift from direct reporting based on large samples and limited use of administrative data (AD in the following), to an intensive use of AD coupled by the development of a new system of business surveys based on limited samples and oriented to collect new business relevant data.
2. Actually, in 2013 the Istat Department of National Accounts and Economic Statistics has started the transition process by implementing a new information system (called *Frame SBS*) for the production of key economic accounts statistics based on the massive, integrated use of administrative data, complemented by survey data for the estimation of the remaining economic accounts statistics (Luzi *et al.*, 2014).
3. As expected, the new production system has determined gains in terms of accuracy (as estimates of the main variables are free of sampling errors), consistency of business statistics over time, lower costs and respondent's burden. However, the transition process implied high initial costs. In effect, moving from the traditional survey-based estimation strategy to the new SBS production model has determined a deep change in terms of both the methodological and organizational strategies adopted, as it implied to pass from a traditional process under the complete control of the survey manager, to a new production model where different competencies, organizational structures and responsibilities have to cooperate to guarantee the final outputs.
4. This paper focuses in particular on the strategy which has been adopted to manage and support the transition from the previous editing and imputation (E&I) strategy to the new methodological and operational organization of the data verification phase. The transition, which took about three years to be actually completed, was highly supported by the Istat top-management, and required the strong cooperation between survey managers, methodologists and IT experts, AD specialists.
5. Managing changes was a time and resources consuming process not only due to the need of implementing new methodologies and tools (such as the new data and metadata systems), but also to the need of training survey managers, ensuring that they fully understand and accept their role in the new integrated production process, and more generally changing their view on data E&I principles and scopes.
6. The transition was facilitated by the efforts made in terms of standardization and industrialization of the overall production process, e.g. by maximizing the use of generalized routines and IT tools. In particular, a strong collaboration with IT experts is still in progress to exploit the CORA/CORE infrastructure to fully industrialize the new SBS production process.

7. The paper is structured as follows. In Section II the statistical register Frame SBS is illustrated, together with the main features of the new E&I strategy which has been developed to support its annual production. Section III focuses on the strategy adopted to manage the transition from the previous survey-oriented E&I model to the new model, the main problems encountered and the organizational solutions adopted.

II. The SBS statistical production strategy

A. The “old” survey-based SBS E&I procedure

8. In Italy, SBS has been traditionally estimated based on data collected through two direct annual surveys: the sample survey on Small and Medium Enterprises (SME) (about 100,000 sampled enterprises with less than 99 persons employed representing a population of about 4.3 million of units), and the total survey on Large Enterprises (LE) (about 11,000 enterprises with 100 or more persons employed). Both surveys estimate *totals* of profit and loss accounts variables, employment, investments etc. in the industrial, construction, trade and non-financial services sectors. A large number of secondary variables are also estimated, also for National Accounts estimation purposes.

9. In case of SME, the editing process essentially consists of a preliminary analysis of the distributions of the main economic variables (e.g. Turnover, Costs, etc.) for outlier detection, followed by a deterministic procedure for the identification of residual consistency errors in micro-data. Imputation is basically performed by using AD (if available in either the *Financial Statements* or *Sector Studies* archives) for the main economic variables, and within-cell nearest-neighbour donor imputation.

10. Concerning LE, the sub-set of very large enterprises are carefully followed-up, while for the remaining businesses an E&I strategy similar to the one adopted in SME is in place: the main difference consists in the fact that the main economic variables for the LE non respondents are always available in the *Financial Statements*¹. It is straightforward to mention that final SBS are obtained by properly combining the SME and the LE outputs.

11. As mentioned above, in this strategy AD are essentially used as indirect auxiliary information, in particular for data editing, for the imputation of (item and unit) non-response, for the validation of final estimates. Any E&I activity is performed on AD data.

12. In this framework, the utilized administrative sources are edited exclusively at the Istat central Department for Business Registers (BR): at this level, relations and data flows with external Agencies are managed, metadata are maintained and regularly updated, and some basic treatments on each administrative archive are performed, such as quantitative data analysis (e.g. verification of number of records), certification of the archive completeness (in terms of internal items), elimination of duplicates, unit identification and assignment of a unique ID for the direct link of the source with the BR.

B. The “new” E&I strategy in a register-based production model

13. The traditional SBS estimation strategy has been completely reversed in 2013 with the development of the *Frame SBS*, as in this system administrative and fiscal data² are used as primary source of information, with the complementary use of the SME and LE data.

14. In the *Frame SBS*, firm-level data for the *main economic variables*³ are acquired from the AD sources (covering about 95% of the whole target population), and after a validation and an harmonization

¹ there is nearly 5% of units under-coverage of this source w.r.t. the SBS target population, mainly due to delays in data transmission.

² *Financial Statements, Study Sectors, Social Security Data* and the Italian “*Unico*” fiscal model

phase, they are directly used as measurements of these variables. As a consequence, the corresponding estimated totals can be obtained at any level of detail by summing-up variables micro-data. It has to be mentioned here that for these variables an imputation step is performed to compensate for the sources under-coverage (see below).

15. The other economic account variables (*components* of the *main economic variables*) are estimated at pre-defined levels of detail using a design based/model assisted approach (known as *projection estimator*, Kim and Rao, 2011) which exploits the randomization process of the SME sample selection under consistency constraints (as the estimated totals of the *components* variables which contribute to a given *main economic variable* are to be coherent w.r.t. the estimated total of the aggregate itself, at domain level).

16. Using AD in the SBS production process has implied the design of a complex E&I strategy, consisting of specific approaches for the verification and treatment of the specific sources of errors/inconsistencies (at both unit and variable level) on each AD source and on the integrated data (see Zhang, 2012, for a wide-ranging discussion of this topic).

17. It has to be mentioned that the AD sources feeding the *Frame SBS* actually cover different - yet partially overlapping - sub-populations of enterprises, and that some sources provide information on (partially overlapping) variables: for each source, this “common” information has been used in the *Frame SBS* production process for assessing the quality of input data and for micro-data E&I (e.g. identification of data inconsistencies and possible measurement errors, missing data imputation).

18. The overall E&I strategy developed for the *Frame SBS* consists of the following main phases:

(a) Editing each AD source:

- domain editing⁴;
- deterministic E&I of micro-data inconsistencies;
- variables harmonization⁵.

(b) Editing the integrated AD sources:

- resolving influential errors;
- deterministic editing of micro-data inconsistencies;
- imputation of not covered units/variables;
- macro-editing.

(c) Assessing and documenting E&I.

19. It has to be remarked that in the *Frame SBS* E&I strategy, “classical” methods and tools have been used in all the process phases, e.g. selective editing and imputation methods like within-cell nearest-neighbour donor imputation and predictive mean matching (Luzi *et al.*, 2014).

20. However, these approaches had to be tailored w.r.t. the specific nature and size of the used AD in order to ensure the E&I process efficiency (results accuracy under time and costs constraints):

- (longitudinal and cross-sectional) selective editing approaches have been adopted taking into account the large size of administrative databases and the fact that no follow-up with enterprises is possible in this case;
- all the available information (both auxiliary and historical data) on target units has been exploited in the implemented E&I methods;
- the latter especially refers to the imputation strategy, where a combination of different imputation techniques have been applied to separate sub-populations/subsets of variables taking

³ like *Value Added, Revenues, Labor Costs*, etc.

⁴ e.g. selection of eligible units, deletion of incomplete and other invalid units, verification of classification variables, etc.

⁵ based on deep analysis of the administrative sources contents, SBS target variables are derived from AD based on the SBS European Regulation definitions.

- into account their characteristics, as well as the type of variables distributions and relationships and available information (see Luzi *et al.*, 2014);
 - where possible, standard methods and tools have been used in order to guarantee as much as possible the reproducibility and the documentation of the E&I phases, e.g. the software SeleMix (Di Zio and Guarnera, 2011) has been used for selective editing.
21. In addition, an effort in terms of overall standardization of the E&I process based on a Service Oriented Architecture (SOA) has been performed. In particular, experimental applications and first implementation of the CORA/CORE services (Bruno *et al.*, 2013) have been performed in order to integrate some of the E&I functions in this standard platform.

22. It is evident that implementing the *Frame SBS* E&I strategy has implied a major investment of resources and the decisive management of relationships among different Istat Departments/experts:

- (a) methodologists, for the design and the implementation of the new E&I process, including the adoption of generalized E&I tools, the revision of the current SME and LE procedures and the management of links among all the procedures;
- (b) IT specialists, for the design of the database (data and metadata), and for the management of processes, data and metadata flows among the various phases of the new SBS E&I process;
- (c) AD experts working the BR Dept, for the monitoring and the management of changes of the sources contents, as well as for the management and the documentation of the quality of AD in input to the *Frame SBS* system;
- (d) external experts from fiscal Agencies, for the analysis of AD contents and for supporting the harmonization phase;
- (e) SME and LE survey managers for the revision (as far as possible) of their own E&I procedures and their integration in the new SBS data E&I process.

C. Towards a revised E&I strategy in the new Istat organization context

23. In late summer 2014 Istat has initiated a modernization program for the transition from the current survey-oriented statistical production model to a completely integrated statistical system based on the more intensive and efficient use of AD, with the development of a new, centralized System of Integrated Registers (SIR). As a consequence, a deep revision of the AD acquisition policy is planned, together with the reorganization of the Istat Departments in charge of the AD collection from external owners, and the first assessment of their quality. In this perspective, the *Frame SBS* production process is expected to be further revised according to the modernization process, together with its *E&I* procedure.

III. Managing changes in the SBS E&I strategy

A. The overall transition strategy

24. A medium-term strategy has been adopted to manage the transition from the “old” to the “new” E&I model, based on a gradual evolution which has been accomplished by spreading changes (together with the related risk factors and costs) on several intermediate achievable stages.

25. In the following, we also consider the further transition of the *Frame SBS* E&I process which will be determined by the Istat modernization process (“to be” *Frame SBS* E&I process).

26. In figures 1, 2 and 3 the “old” (survey-based), the current (*Frame*-based), and the “to be” *Frame SBS* production models and associated E&I strategies are represented. In the following, the intermediate steps of the overall transition strategy are described.

Figure1. Model A. Survey-based E&I model

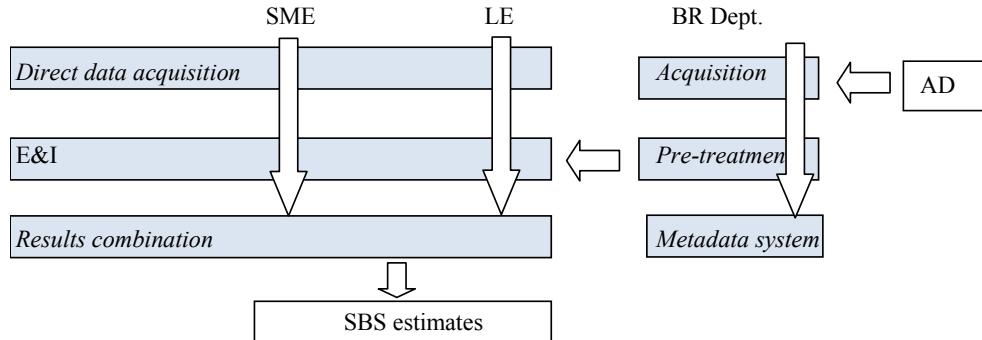


Figure2. Model B. Current Frame SBS-based E&I model

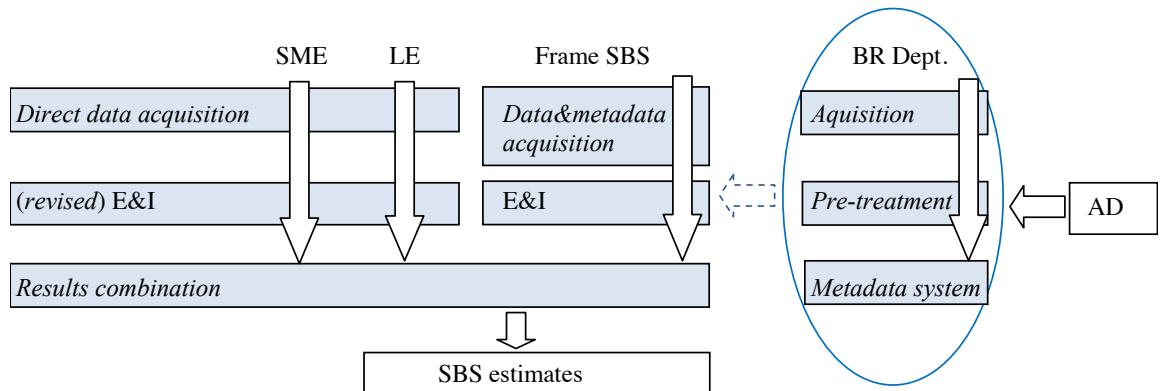
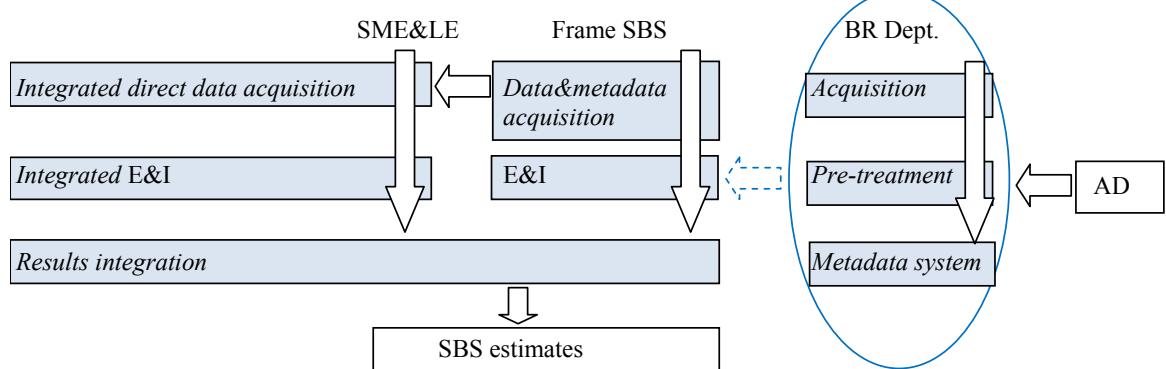


Figure 3. Model C. “To be” Frame SBS-based E&I model



From model A to Model B

27. As already mentioned, moving from the “old” survey-based E&I strategy (where the SME and the LE surveys were conducted based on independent processes and used distinct E&I strategies under the complete control of the two survey managers) to the “new” one has implied the management of changes not only from a methodological point of view, but also from an operational, organizational and infrastructural point of view.

28. This transition has been achieved by two subsequent inter-Departmental working groups (WG) that overall have worked for about three years, under a strong mandate, a continuous monitoring from the top-level management, a rigorous scheduling in order to guarantee the final achievement of results with the required timeliness. In this phase, the SME and the LE surveys continued to be conducted in the

traditional way, and to use their own E&I strategies, in order to ensure the annual release of the statistical outputs as required by the SBS European Regulation.

29. In parallel, the statistical register *Frame SBS* has been developed by the WGs, and its production process has been progressively integrated with the existing survey processes. An essentially new E&I strategy has been established for AD, in order to directly use them for estimating a subset of economic variables (previously estimated using SME+LE data). Some phases of the SME and LE E&I procedures have been revised in order to harmonize them each other and w.r.t. the new overall production context as far as possible.

30. One of the key points of the new E&I strategy is that it is under the *coordinated control* of several managers, who share the common responsibility of the SBS outputs as they are in charge of *complementary* stages of the overall E&I process:

- the BR Department, where preliminary data treatments are performed on AD sources and a metadata system is continuously updated and managed;
- the survey manager who is responsible for the *Frame SBS* production process;
- the SME and LE survey managers, who control their own E&I processes;
- the Methodological Department experts, who are in charge of introducing innovation (as the E&I procedure is continuously evolving), and training the involved survey managers on methods and tools.

From model B to Model C

31. The current developments are focusing on the overall re-design of the SME and LE surveys, in order to better integrate them with the *Frame SBS* information system, ensuring the increase of the efficiency of statistical collection, E&I and estimation: to this aim, direct surveys will use small size samples designed to complement the *Frame SBS* information, i.e. information which is not already available (with enough accuracy) in AD sources, and/or to investigate specific SBS sub-populations.

32. Taking into account the ongoing Istat modernization process, the transition from model B to model C will be managed under a strong mandate from the Istat top-level management, with a deep involvement of the SIR, the IT and the Methodological Dep.ts..

B. Strengths and weaknesses of the SBS transition strategy

33. The systematic use of AD through the *Frame SBS* has implied the design of a new E&I strategy and its integration in the pre-existing SME+LE procedures, characterized by low level of harmonization among SME and LE E&I processes, by the limited use of AD and by independent, scarce verification activities on such data.

34. In this section the most relevant issues encountered and the solutions adopted for managing the transition from the “old” to the “new” E&I strategy are discussed. They relate to the following general areas:

- a) **Managing the methodological transition**
- b) **Managing the technological and infrastructural transition**
- c) **Managing the organizational and operational transition**

a) Managing the methodological transition

35. From a methodological point of view, the main issue to be managed was the need of developing the *Frame SBS* E&I strategy in parallel with the ongoing SME and LE E&I procedures, and integrating them *ex-post*: the need of ensuring the timely release of SBS has allowed only a partial harmonization of

the three E&I procedures at this stage of the transition. This also implied problems in the documentation of the quality of the overall SBS E&I process and final results.

36. Furthermore, in the first phases of the project, the survey managers did not recognize the need of revising their own E&I strategies, and were not aware about the need of adopting a different E&I strategy due to the fact that AD were directly used for statistical production. On the other hand, SME and LE editing staff probably had concern about being judged by external experts on the quality of their E&I processes, so they tended to resist about methodological changes. A further issue was represented by the need of fulfilling the knowledge gap on E&I methods and tools which normally occurs in the case of survey managers/editing staff who work on production processes which do not change for a long time.

37. In order to manage these problems, a lot of resources and time have been spent at the beginning of the project on preliminary discussions and presentations in order to clarify the reasons underlying the need for changing the overall SBS production strategy and for adopting innovative E&I approaches for dealing with integrated AD and survey data. Ad hoc meetings and seminars were organized in order to share knowledge on the proposed E&I methodologies and tools, illustrating their potentials in terms of quality and efficiency gains as well as their scope in the new, more complex information and production context involving the *Frame SBS*.

38. A strong investment has been performed for training survey managers/editing staff not only on methods and tools, but also about the specific E&I problems arising when dealing with AD: concerning the latter, in the new E&I strategy survey managers are completely aware of the issues related to the quality assessment and to the treatment of AD, and they have acquired skills a basic knowledge of the specific methods and treatments adopted on AD as they are fully involved in the E&I processing of AD.

39. In addition, survey managers and editing staff of *Frame SBS* production process, were fully supported in the first runs of the new E&I procedures designed by methodologists through training on the job activities. The same approach has been adopted to manage changes in the already existing SME and LE E&I procedures for their first harmonization: as an example, new (longitudinal) imputation methods have been adopted to better exploit historical and auxiliary information, selective editing methods were tested for some specific variables/sub-populations, etc.

40. An additional issue to be managed relates to the lack of harmonization between data treatments performed at the SBS and at the BR Departments: in the transition period, this issue has been approached by establishing a temporary task force in charge of managing the exchange of data and metadata about the *Frame SBS* input data, monitoring the changes of the AD size and contents and alerting the IT staff for the appropriate adjustments. It is straightforward to mention that in the “to be” model, the preliminary (at the BR) and the SBS E&I procedures will result completely harmonized, mainly due to the complete revision of the BR Dept. activities associated to the ongoing Istat re-organization process.

b) Managing the technological and infrastructural transition

41. Relating to the infrastructural and IT aspects, the main issue has been the management of the transition from the “old” E&I data and process flow (where data obtained based on two independent E&I procedures were combined *ex-post*) to the new strategy where a third E&I process (related to the *Frame SBS*) has been developed, and which also involves the use of generalized tools.

42. In addition to that, the management of large databases as the ones associated to the processes AD sources has required the identification of appropriate, efficient IT solutions (in terms of both hardware and software devices).

43. Also in this case, the transition has been managed by establishing specific task forces, composed by IT experts from the SBS area and IT experts from the Istat central Dept., who was in charge not only to design and implement the new production infrastructure and integrate it with the pre-existing procedures, but also to adopt standard solutions (as the SOA architecture) to achieve higher levels of standardization and process efficiency.

44. Concerning metadata, it has to be remarked that at the time the *Frame SBS* project has started, the BR Unified System of Metadata (SUM) associated to administrative sources was still at an embryonic level of development: for this reason, a tailored metadata system had to be developed for the *Frame SBS*, taking into account the ongoing developments at the BR Dept., including metadata for all the production process, including the E&I phase. At present, SBS and BR experts are collaborating for the fully harmonization of the SUM and the *Frame SBS* metadata systems.

c) Managing the organizational and operational transition

45. As already mentioned, one of the most critical issues in this context was obtaining the support and cooperation from SME and LE survey managers and editing staff at the beginning and during the transition: in particular, survey managers tended to resist to the changes for several reasons: as an example, they prefer to be autonomous in decisions limiting dependencies from other domains, they considered the time spent for retraining themselves as useless and the process redesign too much expensive in a context of scarce resources.

46. An additional issue in the transition towards the new organizational and operational model was the need of making different Units in different Departments collaborate in strong connection in a new, more complex organization of the work where the responsibility of the quality of final data is equally shared. The main risk-factor however consists in the “not structural” nature of this new organization, where responsibilities are clearly defined but collaboration is still managed within temporary WGs. This aspect will be assessed within the Istat modernization process.

47. The decisive element for the success of the project has been the strong commitment from top-level management: actually, the overall change from the “old” SBS production process to the new estimation strategy has been managed by establishing a mandatory project, which has been defined and highly supported by the Istat top-level management. This project has been also included among the Istat strategic tasks in both years 2013 and 2014: actually, it completely fulfilled the recommendations provided at both National and European level on the need of modernizing statistical production processes and increasing their efficiency by overcoming the stove-pipe production model and by a more intensive use of already existing sources of information.

48. In this context, SME and LE survey managers have been not only fully involved in the overall transition process, but they have been assigned direct and formal responsibility for specific portions of the overall SBS E&I process: they have been continuously asked for their collaboration in developing and possibly re-orienting the overall E&I strategy, by exploiting their knowledge on both the investigated phenomena and on the used AD sources (that they had used from several years as sources in their own surveys).

49. In addition to that, in order to get long-time cooperation from methodologists, IT experts, and experts from the BR area, formal roles and coordination responsibilities have been assigned also to them inside the established working groups.

50. In 2015, the further development of the SBS production system will represent one of the pilot projects for testing the operational and organizational changes deriving from the Istat modernization: it is straightforward to mention that the associated E&I strategy will evolve accordingly, following the complete transition of the SBS production system towards its final configuration.

IV. Concluding remarks

51. Managing changes from the traditional survey-based SBS E&I strategy to a new strategy where administrative and survey data are used in combination has been a time and resources consuming process not only because of the need of implementing and new methodologies and tools, integrating them and training survey managers on their use, but also from a technological and organizational point of view.

52. A medium-term strategy has been adopted at Istat to drive the transition from the previous stovepipe model to a new configuration where the SME and the LE surveys and their associated E&I procedures are efficiently redesigned in order to complement the *Frame SBS* information content.

53. At the present stage of development, the E&I strategy of the *Frame SBS* is combined with partially harmonized SME and LE E&I procedures: ongoing activities aim at the further development of the current model, also driven by the modernization process in progress at Istat, which will determine a deep revision of the overall strategy of AD acquisition, treatment and integration for statistical production.

54. The main goals in the next future are ensuring the efficient transition towards a more efficient, harmonized and standardized SBS production process and its associated E&I strategy, in which administrative and survey data are used in a completely integrated and industrialized framework, and which is supported by the efficient, structured management of relationships among the different Istat organizational structures involved in the process.

References

- Bruno, M., Scannapieco, M., Vaccari, C., Vaste, G., Virgillito, A., Zardetto, D. (2013). CORE: a Standard Platform for Statistical Production Processes. Proceedings of NTTS 2013, Brussels, 2013.
- Di Zio, M., Guarnera, U., Varriale R. (2014). *Imputation with multi-source data: the case of Italian SBS*. Paper presented at United Nations Economic Commission for Europe, conference of European statisticians.
- Di Zio M., Guarnera U. (2013). Contamination Model for Selective Editing. *Journal of Official Statistics*, Vol. 26, n. 4, pp . 539-556.
- Di Zio M., Guarnera U. (2011) SeleMix: an R Package for Selective Editing via Contamination Models. 2011. *International Methodology Symposium*, Statistics Canada. November 1-4, 2011, Ottawa, Canada.
- Luzi, O., Guarnera, U., Righi, P. (2014) The new multiple-source system for Italian Structural Business Statistics based on administrative and survey data. European Conference on Quality in Official Statistics (Q2014). Vienna, 3-5 June.
- Kim, J. K. K., Rao, J. N. K. (2011). *Combining data from two independent surveys: a model-assisted approach*. Biometrika. No.8, pp. 1–16.
- Zhang, L.C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*. Vol. 66, No. 1, pp. 41–63.