Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Neuchâtel, Switzerland, 18-20 September 2018)

Selective editing in the Integrated Business Statistics System (SINTESI)

Prepared by Orietta Luzi, Antonia Manzari, Tiziana Pichiorri, Fabiana Rocci, Simona Rosati, Roberta Varriale, Italian National Institute of Statistics, Italy

I. Introduction

1. Following the experience of Statistics Canada, Istat decided to invest in the design and implementation of a new complex infrastructure in order to develop a unique environment to support the statistical production processes in the economic area. The new system, called SINTESI (Integrated Business Statistics System), will allow survey managers to perform all phases of their statistical process in a unified environment.

2. SINTESI is expected to realize important economies of scale, scope and knowledge in productive contexts with a limited number of resources. In particular a number of advantages could be achieved:
- simplifying and standardizing processes;
- reducing costs associated with operational aspects of surveys to increase efficiencies and improve timeliness;
- supporting the management of statistical production processes while eliminating redundancies and ensuring the overall coherence of the statistical processes in the business area.

3. The main components of SINTESI are: a metadata module, an informative architecture, a survey management tool, and a series of methodological statistical services, each implementing a methodology for a specific phase of the survey production process.

4. Short-term business statistics (STS) have been selected as starting point of the application of the new system. The first statistical services to be implemented in SINTESI are related to the editing and imputation process. In particular, the first objective to be achieved is the development of a statistical service for selective editing. The selective editing method which will be adopted is based on mixture models and is implemented in the SeleMix R package developed at Istat.

5. As a part of the project, experimental analyses are planned on selected STS in order to: (i) correctly design the statistical service for selective editing in SINTESI, and (ii) properly introduce the methodology in STS statistical production processes.

6. The monthly Survey on Turnover and Orders in Industry is currently under study. In this work we show the first results of the experiment carried out on this survey, focusing on critical issues and strengths of the entire selective editing process in the SINTESI framework.

II. The project

7. In 2017, Istat has started the project SINTESI with the purpose to develop a new common infrastructure for conducting business surveys and for producing economic statistics (*TO-BE model*

hereafter). The main expected benefits are the standardization and harmonization of the "key phases" of production processes of different surveys, the traceability of actions and procedures carried out, the improvement of quality and numerical consistency of statistical outputs.

8. The main problems in the implementation of the *TO-BE model* are expected to be related to the costs to be incurred in the short/medium-term for the implementation, introduction, and maintenance of the new operating model.

9. Given the complexity of the system to be developed, several professional figures have been involved in the project to form a comprehensive group of work: methodologists, IT specialists, data collection experts, subject-matter experts.

10. The first group of business surveys involved in the project are STS. This choice has been based on the need to decrease production costs (in particular, those depending on the human intervention on data in specific phases of the process), while safeguarding the timeliness requirements, and ensuring higher levels of efficiency and consistency among statistical production processes and related outputs.

11. In the current situation (*AS-IS model* hereafter), STS are characterized by highly heterogeneous production processes: even though some methodological solutions can be classified according to similar criteria, they are implemented in different ways, e.g. different solutions are adopted to manage data flows from the data collection platform, statistical registers and administrative data are used non-homogeneously, a limited use of auxiliary information from related surveys is made, and different systems for the management of the survey production processes are adopted.

12. The main characteristics of the SINTESI System will be:
   (a) Standardization of methodologies, information flows, and software components. This aspect involves the optimization of time and costs related to data processing, with a reduction of manual interventions and a consequent decrease of respondent's burden.
   (b) Metadata driven approach. The system architecture is guided by metadata both for data component (microdata, administrative archives, registers, etc.) and for process components.
   (c) Use of standardized metadata for the management of the entire statistical production process, in particular for the development of questionnaires, for sampling, for editing and imputation, and estimation process.
   (d) Standardized and efficient use of every feasible type of auxiliary data source, as administrative data and statistical registers. The controlled and standardized use of registers and other auxiliary sources of information guarantees robustness, efficiency and consistency of the statistical production system.

13. The system will consist of the following main elements:
   (a) a Data Repository, which is the logical environment the SINTESI information sources are referred to (Business Register, auxiliary sources of information, raw data, working data, etc.);
   (b) a Metadata Repository containing all the metadata on survey units, variables, and statistical processes;
   (c) the Interpreter, which is responsible for generating procedures for querying the Data Repository and some processing of data driven by metadata;
   (d) a (unique) Survey management system, which provides the management functions facilitating the activities of the subject matter experts related to the production of statistical outputs;
   (e) a Repository of generic Statistical Services, i.e. software components which allow the execution of statistical activities that can be used in a variety of application contexts.

14. Generally speaking, in order to develop the survey management system, a set of proper statistical services for every survey phase are to be implemented. The first service to be implemented in 2018 is related to the editing and imputation (E&I hereafter) process.

III. Editing and imputation in SINTESI

A. Introduction

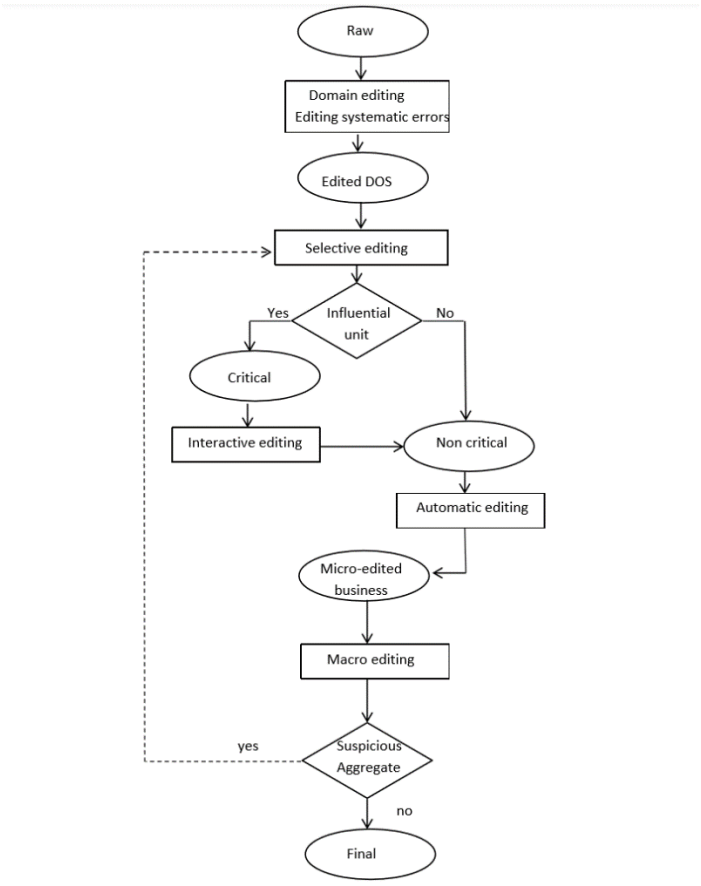
15. The reference framework which has been adopted for designing an E&I process in SINTESI is the Generic Statistical Data Editing Model (GSDEM; UNECE, 2015). In GSDEM the E&I process is interpreted as a set of standardized, consistently described information *objects* that are the inputs and outputs in the design of the overall data processing flow. More specifically, the E&I process is primarily composed of statistical *functions*.

16. An E&I work flow is a configuration of generic E&I functions which is specified in terms of the mapping from the input to the output of each E&I function, and the associated metadata including the relevant concepts, data structure, routing conditions, stopping rules, etc.

17. In general, the GSDEM describes the elements that properly combined allow to design any E&I process (a *SDE flow model*), as well as the main elements that determine the choice of a specific E&I procedure. Also, according to the Generic Statistical Information Model (GSIM) terminology (UNECE, 2013b), the E&I process is represented as a 'business process', composed of 'process steps' (e.g. *Editing systematic errors*, *Selective editing*, *Interactive editing*, *Automatic editing*, *Macro editing*, etc.) together with their associated functions, and 'process steps control' (e.g. *Influential units*, *Suspicious Aggregates*, etc.), (UNECE, 2015).

18. GSDEM also proposes some examples of 'generic E&I model flows' for different types of statistical production processes on enterprises and households. In the area of economic statistics, the *generic SDE flow model for STS* is taken as reference in the context of SINTESI (Figure 1). This model is the one to be implemented in SINTESI.

Figure 1: SDE flow model for Short-Term Business Statistics (STS)



B. The strategy to be implemented

19. The main steps of the overall strategy to move from the *current STS E&I models* to the *generic SDE flow model for STS* to be implemented in SINTESI are:

(a) select a set of "pilot" STS;
(b) provide the workflow of the current production process of each pilot survey;
(c) analyse the workflows to extract their operating structure (data, steps, flow) in order to identify common steps that can be performed through the use of generic statistical services and for designing a generalized *SDE flow model* for STS;
(d) build the statistical services;
(e) fine-tune the mode of application of each service to each specific STS.

20. We describe each step as follows.
Step (a). Three STS surveys have been selected for the first implementation and testing of SINTESI and for the first migration: the *Monthly Survey on turnover and orders of the industry*, the *Quarterly Survey on turnover in the services* and the *Monthly Survey on retail sales*. These surveys were selected as pilot because:

- they have relatively simple and sufficiently homogeneous processes and methodological characteristics;
- their current information systems would gain in efficiency introducing product innovations of methodology and software. and generalization.

21. Step (b). For each pilot survey, the detailed description of the current E&I process were provided, together with the associated data and metadata. Particular attention was given to the activities included in sub-processes "5.3 Review and Validate" and "5.4 Edit and Impute" of the GSBPM Version 5.0 (UNECE, 2013a).

22. Step (c). This step allows to identify the optimal location of each statistical service within the *SDE flow model* for each pilot survey, while the (e) step permits to ensure the effectiveness of each methodological function implemented in it together with its optimal mode of application.

23. In the *generic SDE flow model for STS*, the process step to be implemented with priority is considered *selective editing*: actually, in this type of surveys influential errors need to be timely identified while optimizing the trade-off between costs (associated to the interactive review of data) and accuracy of output statistics. *Selective editing* is then the first Statistical Service which will be made available in SINTESI, together with its associated functions "*Identification of units affected by influential errors*" and "*Selection of units for interactive treatment, selection of units for non-interactive treatment, selection of units not to be amended*".

24. It has to be underlined that the development of the generic Statistical Services will exploit as much as possible the statistical functions implemented in the IT tools currently available in the area "Methods and IT tools for statistical production" (https://www.istat.it/en/methods-and-tools/methods-and-it-tools) of the Istat's website. Concerning *selective editing*, the methodology based on contamination models which is implemented in the SeleMix R package has being chosen (Guarnera and Buglielli, 2013; Buglielli and Guarnera, 2016). This methodology is currently used in many processes at Istat as a suggested approach for the identification of potentially influential errors in continuous variables (https://www.istat.it/en/methods-and-tools/methods-and-it-tools/process/processing-tools/selemix).

25. As a consequence, for the first release of SINTESI, steps (d) and (e) of the above delineated implementation strategy are currently focused on selective editing via SeleMix.

26. In order to develop a statistical service able to implement the selected methodology, a mapping of the current E&I processes needs to be made for each pilot survey. Mapping consists in a description phase-by-phase of all the actions involved in the E&I processes. These actions can be summarized as follows:

1) describe procedures or methods used to detect outliers (e.g. edit rules or graphical editing);

2) provide any details on how to identify erroneous outliers and how they are treated;
3) specify how missing or inconsistent data items are corrected;
4) explain which of the macro-editing techniques are being applied to validate estimates and any auxiliary information included (e.g. historical data or other sources of information).

27. On the basis on the results of mapping, the aim is to redesign the current E&I strategy according to the *generic SDE flow model for STS* and, at the same time, to improve and to standardize the E&I process. On the other hand, the specific characteristics of each survey should be preserved and it is well known that they may influence the design of an E&I process.

28. This represents an additional challenge for the Project; the role of metadata in relation to GSDEM can be crucial for achieving the goal of producing a flexible and robust system able to adapt to different survey contexts.

29. The analysis of the first mapping revealed what several authors have already emphasized, that is business surveys incur substantial E&I costs, especially due to intensive follow-up and interactive editing (Luzi *et al*., 2008). The question is: "Do E&I actions, performed in business surveys, produce a substantial increase in data quality that could justify the resources and time dedicated to them?".

30. In case of over-editing, selective editing techniques could be the solution to reduce costs and increase efficiency of an E&I process in business surveys. However, further research is needed in this context to achieve this aim.

C. Selective editing and SeleMix

C.1 Characteristics of the SeleMix package

31. The aim of selective editing is to identify observations affected by influential errors in order to reduce the cost of the E&I phase maintaining at the same time a certain level of quality of estimates (Lawrence and McKenzie, 2000; Lawrence and McDavitt, 1994).

32. The method for selective editing used in this context, implemented in the SeleMix package, is based on a latent class model, that takes advantage of a probabilistic specification of the true data and of the error mechanism. More specifically, a Gaussian model is assumed for true data and an "intermittent" error mechanism such that a proportion of data is contaminated by an additive Gaussian error. Observations are prioritized according to the values of a score function that expresses the impact of their potential error on the estimates of interest (Latouche and Berthelot, 1992). All the units above a given threshold are selected since they potentially represent the observations affected by important errors.

33. The approach used in SeleMix is based on explicitly modelling both true (error-free) data and error mechanism. True data (possibly in log-scale) are thought of as n realizations from a random p-vector Y that, conditional on a set of q covariates X, is normally distributed with mean vector BX and covariance matrix Σ. The intermittent nature of the error is modelled through a Bernoullian random variable I, with parameter w, assuming value 1 or 0 depending on whether an error occurs in data or not respectively. The parameter w can be interpreted as the marginal probability of an observation of being affected by an error in at least one variable Y. Conditional on I=1 (presence of error), a Gaussian additive error is assumed, with zero mean and covariance matrix proportional to Σ, the proportionality constant being some positive number λ. Thus the model parameters are $\theta = (B, \Sigma, w, \lambda)$.

34. The previous assumptions allow us to explicitly derive, via Bayes formula, the distribution of the true data conditional on the observed data that is a mixture of a mass density corresponding to absence of error and a Gaussian distribution corresponding to presence of error. This mixture is the central object for the proposed selective editing method and is completely identified by the set of parameters θ. The model parameters can be estimated by maximizing the likelihood function based on the observed data using an EM-type algorithm. Once the model parameters have been estimated, they can be plugged into the functional form of the conditional distribution of true data given observed data. The selective editing strategy consists in using this estimated distribution to build up a score function. Once all the

observations have been ordered according to this score function, we are able to estimate the residual error remaining in data after the correction of the first k units (k=1,..,n). The number of most critical units to be edited can be chosen so that the estimate of the residual error is below a prefixed threshold. Details on model specification and parameter estimation can be found in Buglielli *et al*. (2011), Di Zio and Guarnera (2013).

35. The SeleMix package includes R functions to implement selective editing method based on such a model. In particular, SeleMix is composed of three functions *ml.est*, *pred.y*, *sel.edit* performing each a specific step within the selective editing procedure: estimation of the parameters of the model, prediction of variable values, selection of the subset of observations affected by influential errors based on local and global scores). A further function, *sel.pairs*, provides graphical tools.
If the analysis is performed on sample surveys, in order to select the most influential errors SeleMix uses the sampling weights. In the context of economic surveys when positive variables are analyzed, the method is more realistically applied to logarithms of data instead of data in their original scale.

C.2 The implementation strategy of selective editing in SINTESI

36. In this paragraph we describe the key elements of methodological and operational characteristics of the SeleMix algorithm, in order to implement a valid strategy for selective editing as a service in SINTESI, according to the survey process needs.

37. Three steps are required to perform selective editing when SeleMix is used:
1. estimate model parameters;
2. predict the "true" values for the target variables;
3. identify influential errors.

38. The workflow from step 1 to step 2 can be changed, depending on the survey process and its features. A preliminary data analysis is expected to give information about how to set the process for the specific survey under study.

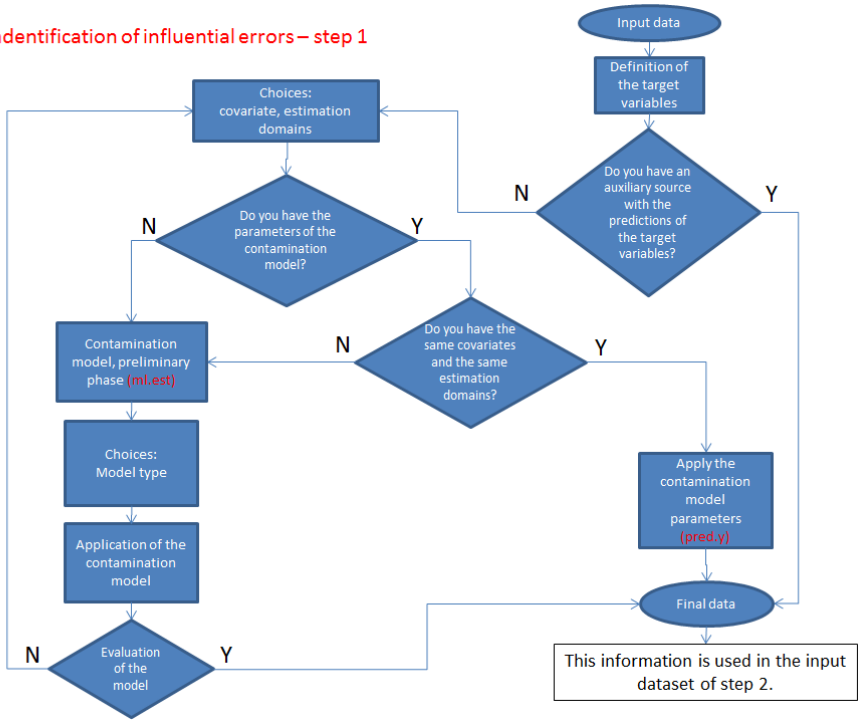
39. In step 1, the strategy to be implemented has to take into account the following steps:
- ✓ select one or more target variables, e.g. Y_1. Y_2;
- ✓ identify any available auxiliary variables, e.g. X_1 and X_2 that are highly correlated with the target variables (they could also refer to the same target variables observed at different reference period or from another data source);
- ✓ choose the model: univariate ($Y_1 \mid X_1, X_2$ and $Y_2 \mid X_1, X_2$) or multivariate ($Y_1, Y_2 \mid X_1, X_2$);
- ✓ define the estimation domains.

40. Usually Step 2 is performed immediately after step1. Nevertheless, a different approach could be adopted for short-term surveys. Indeed, performing step 1 could not be essential when parameters estimates are available from a previous time of the same survey on a monthly or quarterly basis. This approach assumes low sampling variability of model estimates.

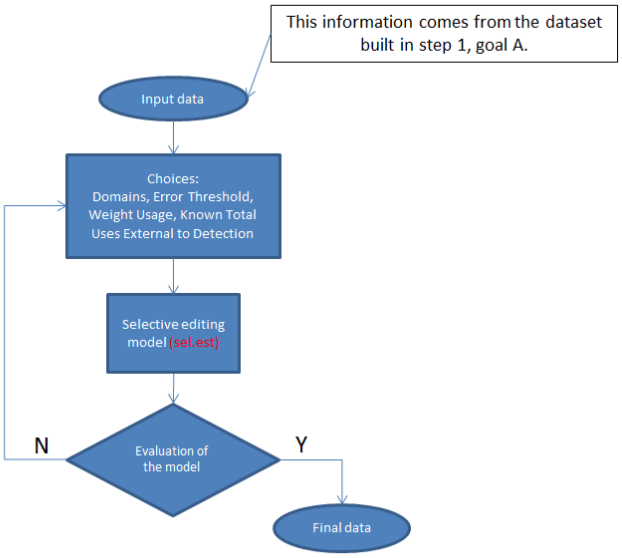
41. In step 3, the workflow of the process is different depending on:
- ✓ domain for selective editing application;
- ✓ error threshold;
- ✓ whether or not sampling weights are used;
- ✓ whether or not known population totals are used.

42. The following diagrams aim to represent step 1 and step 2 as above described.

A. Indentification of influential errors – step 1



A. Indentification of influential errors – step 2



IV. The Survey on Turnover and Orders in Industry: first results

43. STS aim to describe the most recent developments and changes of the economy, under the EU Regulation (EC) no. 1165/98 of the Council, and subsequent amendments, which define the level of detail, the standards and the frequency according to which the indicators must be produced. The development in the different economic domains is described with a series of STS indicators, such as production, turnover, prices, number of persons employed, gross wages and several more. The STS indicators are published as indices which show the changes of each indicator in comparison with a fixed reference year (base year).

44. Among others, the Survey on Turnover and Orders in the Industry (limited to mining and manufacturing economic activities, NACE Rev. 2 sections B and C) has been chosen as the first pilot

survey in the SINTESI project, given its relevance and representativeness of STS production processes. Two indices are released by means of the monthly survey data on turnover and orders of enterprises in the manufactory sector: the turnover index measures the trend over time of the sales of industrial companies, while the specific index about the orders captures the dynamics of the value of the orders that companies receive from customers.

45. The survey is based on a panel of enterprises, selected at the base year on a quota sampling criteria, in order to reach the 70% of the total turnover as measured by the Italian Businesses Register (ASIA). This results in about 7.000 surveyed enterprises, whose data are elaborated according to their Kind of Activity Unit (KAU). Statistical results are released monthly, with only 30 days of delay with respect to the end of the reference month. The survey process runs continuously during the period of every month.

46. For the SINTESI purposes, the survey process has been analyzed in order to design its current picture (*AS-IS survey model*). Focusing on the E&I phase, the purpose is to study to which extent the current SDE process can be designed according to the GSDEM model. At this stage, the focus is about how selective editing can be integrated in the process itself in order to make more efficient the interactive data treatment. To this aim, the description of the current process highlights the phases during which the subject-matter experts revise the enterprises data through interactive inspection based on deterministic rules.

47. The current E&I process has been represented in macro-phases, along which few deterministic edit rules are run, based on a comparison with the longitudinal profile of the enterprise itself. The main rule calculates the variation over the same month of the previous year, for which records with ±30% variation are considered to be anomalous. For these records further controls are done to be able to distinguish between systematic errors, for which the correction is given, and other kind of errors which reason and the following correction has to be evaluated case by case. In these cases, the interactive controls are based on an intense work of collecting information, both directly from the respondents both from other auxiliary information delivered by other survey approximately about the same period.
It is important, indeed, to understand whether the high variation is caused by an actual error to be amended or otherwise if it hides a structural change of the enterprises (e.g. either splitting or fusion of the enterprises) that could change their representativeness into the panel.

48. In order to test the performance of SeleMix approach, a first experiment has been implemented, based on data of the years 2017 and 2016. A simulation is designed to run over every month of the year 2017, trying to model the same longitudinal relationship as the current process is based on. In this view, the target variable is the monthly raw data on turnover Y_t, to be related to the turnover of the same month of the previous year $Y_{t-12,}$ as the longitudinal information represents commonly an auxiliary information for the STS editing rules.

49. A proper benchmark has to be defined in order to evaluate the results of the model with respect to the current data survey. To this aim, for each month the comparison between raw data and the corresponding final ones makes it possible to identify those data that are changed according to the current E&I process. This set of data would include every kind of potential erroneous or anomalous data. At first, for the simulation purpose, the different data are treated according to the type of potential errors: on one side the systematic errors are eliminated (as the process in Figure 1 suggests selective editing have to be run after systematic errors detection); on the other side the remaining data are flagged as being erroneous for other reasons, for which further analysis should be done to understand what caused them. Then selective editing model is estimated for the set of data originally corrected plus those records that were flagged. The final aim would be to assess which data the model identify as influential errors. The results should be compared to the flagged observations, over which the performance of selective editing can be ruled out, in terms of capability of correctness, identifying the errors by minimizing the overall amount of manually revised units.

50. The first results (see Table 1) show how the model, which exploits the longitudinal behavior of enterprises, identifies subsets of flagged data (change for not systematic errors), depending on the specified threshold (tsel).

Table 1. Results of SeleMix for the Survey on Turnover and Orders in Industry: surveyed units and influential data per month – Year 2017

Month	Total n. obs	flagged (a)	Outlier and Influential (b)			Flagged and Outlier and Influential (a∩b)		
			tsel 0.003	tsel 0.004	tsel 0.005	tsel 0.003	tsel 0.004	tsel 0.005
1	6713	107	363	321	137	20	16	13
2	6729	84	398	252	214	21	16	15
3	6758	72	179	128	94	13	7	7
4	6740	67	394	328	260	10	7	6
5	6743	76	308	140	133	13	11	11
6	6748	82	300	214	181	15	12	10
7	6743	89	365	313	215	16	16	12
8	6623	100	604	494	357	19	19	18
9	6715	102	299	192	183	9	5	5
10	6685	111	316	173	144	11	6	6
11	6678	141	317	244	193	12	9	9
12	6655	186	484	268	144	25	16	12

51. Additional analyses are needed to evaluate performance of the proposed selective editing method in the specific context to better define a strategy to release a generalized service.
As an example, the following methodological issues need to be addressed:

- ✓ the impact of the influential errors on the final indices needs to be quantified;
- ✓ a comparison between the set of flagged data and the influential ones can suggest which errors mechanism the model identifies;
- ✓ any possible seasonal effects need to be assessed and adjusted.

V. Conclusions and next steps

52. The implementation of the new common infrastructure SINTESI for conducting business surveys and for producing economic statistics (*TO-BE model*) implies a deep analysis of the current production processes. Concerning STS, the AS-IS description of statistical processes has still to be finalized, because some issues results difficult to be standardized.

53. Nevertheless the first mapping of the data processing flow of the Survey on Turnover and Orders in the Industry revealed some interesting aspects in order to understand how to proceed.

54. With regard to Statistical Services, Selective Editing has been evaluated as the first service that can be efficiently implemented for STS processes. For the SINTESI project, the methodology implemented in the SeleMix software has been selected. At present, the “transformation” of SeleMix algorithms in a Statistical Service in the SINTESI infrastructure is an ongoing activity.

55. In parallel, experimental applications of selective editing models to the selected STS are to be carried out in order to properly integrate the methodology in the survey data processing flow. The first experiments on monthly data of the Survey on Turnover and Orders in the Industry should allow to understand which kind of errors are detected on the basis of the current procedures in respect of SeleMix approach.On the one hand, selective editing can allow to save time and costs for interactive editing, since only units that result to be both potentially erroneous and influential are revised. On the other hand, the adopted selective editing approach could improve the quality of survey data by identifying those errors which are not edited based on the current E&I process**.**

References

Buglielli, M.T., Di Zio, M., Guarnera, U., and Pogelli, F.R. (2011). Selective Editing of Business Survey Data Based on Contamination Models: an Experimental Application. NTTS 2011 New Techniques and Technologies for Statistics, Brussels, 22–24 February 2011.

Buglielli, M.T. and Guarnera, U. (2016). SeleMix: Selective Editing via Mixture Models, Version 1.0.1, available at: https://CRAN.R-project.org/package=SeleMix.

Di Zio M., Guarnera U. (2013). Contamination Model for Selective Editing. *Journal of Official Statistics*, Vol. 26, n. 4, pp . 539-556.

Guarnera U., Buglielli M.T.(2013). SeleMix: an R Package for Selective Editing, available at https://www.istat.it/it/files/2014/03/SeleMix-vignette.pdf.

Latouche M., Berthelot J.M. (1992). Use of a score function to prioritize and limit recontacts in editing business surveys. *Journal of Official Statistics*, 8, n.3, 389-400.

Lawrence, D., McDavitt, C. (1994). Significance Editing in the Australian Survey of Average Weekly Earnings. *Journal of Official Statistics*, 10, 437-447.

Lawrence D., McKenzie R. (2000). The General Application of Significance Editing. *Journal of Official Statistics*, 16, n. 3, 243-253.

Luzi O., De Waal T., Hulliger B., Di Zio M., Pannekoek J., Kilchmann D., Guarnera U., Hoogland J., Manzari A., Tempelman C. (2008). Recommended practices for editing and imputation in cross-sectional business surveys.

UNECE (2013a), Generic Statistical Business Process Model. Version 5.0, December 2013, available at: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

UNECE (2013b). Generic Statistical Information Model (GSIM),Version 1.1, December 2013, available at: http://www1.unece.org/stat/platform/display/gsim.

UNECE (2015). Generic Statistical Data Editing Models - GSDEMs, Version 1.0, October 2015, available at: https://statswiki.unece.org/display/sde/GSDEMs.