Working Paper.

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(Geneva, Switzerland, 15-17 April 2020)

An overview of the editing and imputation process of the 2018 Italian Permanent census

Prepared by Bianchi G., Filippini R., Lipsi R.M., Pezone A., Scalfati F., ISTAT, Italy

I. Introduction

1. For the first time, in 2018, the Italian National Institute of Statistics (Istat) conducts, differently from the traditional census, an annual survey of the main characteristics of the country's resident population and its social and economic conditions at national, regional and small areas levels. The new Permanent census of population and housing involves a sample of Italian households each year: about 1,400,000 families resident in 2,800 Italian municipalities. Final census results are produced by integrating yearly survey data with administrative and register data, related to different topics. This new census structure requires a review of the overall Editing and Imputation (E&I) process, introducing new generalized solutions and improving standard methodologies taking into account technological innovations. In this paper, we will provide an overview of the Editing and Imputation process of the 2018 Italian Permanent census describing the solutions adopted in solving consistency problems in (and between) topics related to households and personal characteristics, i.e. education and employment status. We will also introduce a generalized Data Editing and Imputation System (DEIS) for the management, scheduling and monitoring of IT procedures carried out for the processing of census data. This is an integrated system of generalized services that can be used in different context.

II. The new strategy of data editing and imputation

2. The main objective of the E&I process is to provide a complete and coherent set of data with respect to the compatibility plan, carrying out imputation operations that preserve the original distribution of the information collected. The strategy involves breaking down the overall problem into simpler sub-problems and finding appropriate solutions for each of them. The overall process is defined tacking into account the experience and tools of the 2011 Population and housing census and by adapting activities and/or by developing new ones, considering the new investigation methodologies and the new technological infrastructures used for the Permanent census.
The variables of the questionnaire are divided into thematic areas and have been treated separately taking into account the prerequisites of each area.

3. The treatment of the data follows the processing logics represented by the scheme shown in Figure 1. The validation of the final data, produced by the E&I process, requires the preparation of tables that allow verifying the results deriving from the various stages of data processing and by the use of other sources for comparative analyses. In particular, these sources are the 2011 Population and housing census, the Labour Force Survey, the Base Register of Individuals (BRI) and administrative data for education and employment.

4. The E&I process of the data, collected by the 2018 Italian Permanent census, can be divided considering the following thematic areas:

(a) dwellings and buildings;
(b) personal/family information;
(c) citizenship and residence;
(d) education and training;
(e) economic activity status;
(f) commuting;

that can be grouped into three phases (Figure 1):

- Phase 1 - E&I dwellings/buildings
- Phase 2 - E&I gender, date and place of birth, citizenship, type of family nucleus and usual residence
- Phase 3 - E&I education, economic activity status and commuting.

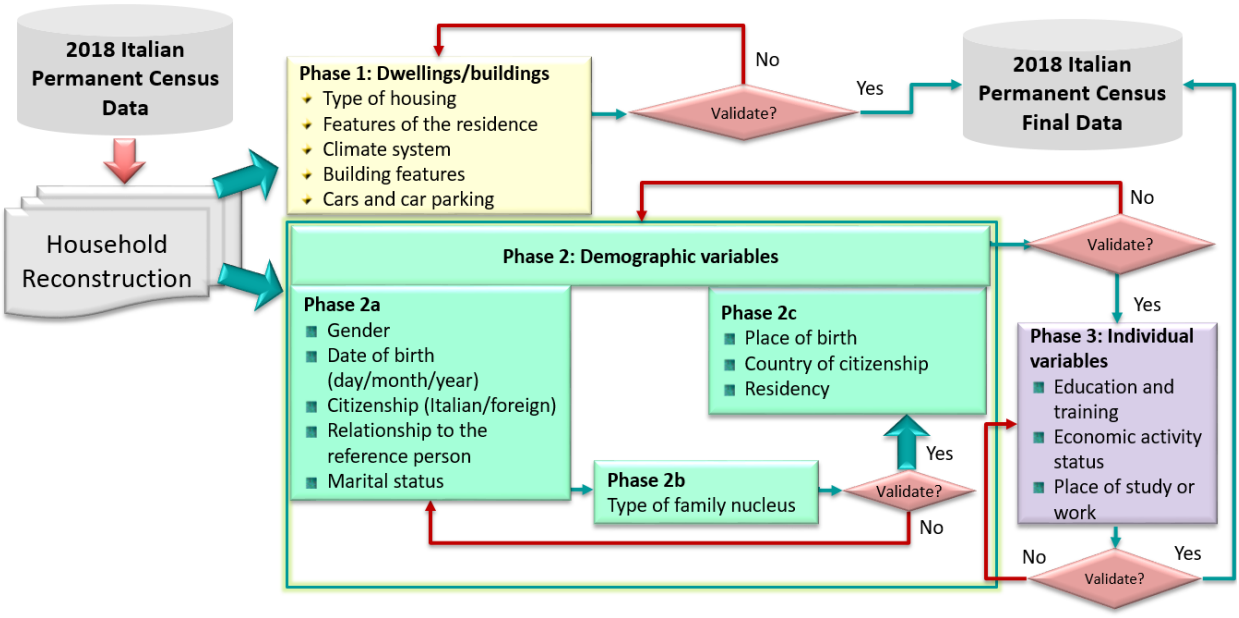
5. The processing of the three phases can take place, in some cases, independently allowing a certain degree of parallelism; in others, it is necessary to introduce a sequence of activities foreseen by the E&I process. All phases can start only at the end of the *Household Reconstruction*, which leads to the composition of the households through the linkage between detected individuals and their own households, after having verified and resolved the presence of any duplicates.

6. The E&I system of the collected data produces also quality indicators and contingency tables for the comparison between raw data and final census data. This comparison is necessary in order to verify and test distributions while preserving coherence and minimizing distortions due to possible imputations. All the checks are carried out at the microdata level, while at the macro level a set of tables are prepared for monitoring the E&I processes at provincial level.

7. Auxiliary sources, in particular administrative and register data, are used in the sample data E&I process both in a macro level, for the validation of the sample data, and in a micro level for comparisons between detected and administrative data, when the sample data is inconsistent.

8. Each of the above-mentioned phases includes one or more activities represented in the coloured boxes of Figure 1. While the grey boxes show the activities of the external team involved in the process.

Figure 1. Flowchart of the Editing and Imputation process



The following sub-paragraphs describe the main steps for the three phases.

A. Phase 1 - E&I dwellings and buildings

9. This phase involves the correction of the variables relating to private households, dwellings and buildings. The E&I consists of the following steps:

- Step 1 - type of housing ;
- Step 2 - features of the residence;
- Step 3 - climate system;
- Step 4 - building features;
- Step 5 - car and car parking.

The main imputation methodologies used in this phase by the E&I process are: deterministic techniques (in the case of deductive imputations) and probabilistic or *data driven* imputation techniques.In this phase the DIESIS (Data Imputation and Edit System - Italian Software) computer system used (see section IV), allows the application of *data driven* and minimal change methods. Where possible, the subset of the variables are partitioned into disjoint sets with respect to the set of consistency rules between variables defined in the compatibility plan. This approach also applies to the later stages.

10. For the identification of cohabitant families, it has been used a graph theory approach to reconstruct the missing link of the cohabitant families within the same house. In particular, a combinatorial optimization algorithm has been implemented in order to identify the connected subgraphs.

B. Phase 2 - E&I gender, date and place of birth, citizenship, type of family nucleus and usual residence

11. In the second phase, there are 3 sub-phases (phase 2a, phase 2b and phase 2c of Figure 1). A first one which involves the correction of the personal data gender, date of birth (day, month and year), citizenship (Italian or foreign), relationship to the reference person and marital status. In the second sub-phase, the E&I of the variables relating to characteristics of couples and type of family nucleus are performed. In the third sub-phase the correction refers to the place of birth, the country of citizenship and the previous place of usual residence.
The country of citizenship can be corrected only after calculating the derived variables relating to type of family nucleus.

The main E&I steps of phase 2 are:

- Step 1 - usual resident population by sex, age and citizenship (Italian or foreign);
- Step 2 - resident population by place of birth;
- Step 3 - resident population by country of citizenship;
- Step 4 - type of family nucleus
- Step 5 - population by previous place of usual residence.

The imputation methodologies in this phase are deterministic, probabilistic and data driven (DIESIS).

12. In his phase, the Base Register of Individuals, storing core variables such as place and date of birth, gender and citizenship for each individual from various administrative sources, represents an important source for micro level comparisons and for the validation of personal information.

C. Phase 3 - E&I education, economic activity status and commuting

13. In the third phase, the data E&I is carried out at individual level. In particular, the corrected variables are related to education and training (including the "qualification"), employment characteristics and the usual movements to reach the place of study or work.
The main steps of E&I are the following:

- Step 1 - education and training;
- Step 2 - economic activity status;
- Step 3 - place of study or work.

14. The previous mentioned E&I steps are mainly sequential, excluded the correction of the data relating to the routes (municipality of origin/destination in commuting journeys), the time taken and the mode of transport used for commuting journeys (included in step 3).
Also in this phase the main methodologies of imputation of the E&I process are: *data driven* methods, probabilistic and deterministic methods. The DIESIS computer system is available for the application of *data driven* and minimal change methods.

15. Administrative data for education and employment can be used in this phase in a micro level to restore the correctness of the data when the record violates some edit rule. The use of big data (web data) has been useful to support the correction and validation of some commuter routes taking into account the relationship between the vehicle, the time taken and the distances between the usual residence and the usual place of work or study (see section VI).

III. The generalized editing system

16. A generalized editing system allows checking the consistency of the data collected with respect to the check plan for individual records and hierarchical structure, with intra-record and inter-record rules. Furthermore, the editing system allows identifying the inconsistency and redundancy of the rules. The application allows identifying the exact and incorrect questionnaires, the individuals involved in violated edits and the fields involved in the violation of the rules. For this purpose, there is a metadata table describing the rules that is functional to the execution of the application. This table contains the following information:

(a) The type of rule: Validity, Logic, Mathematics and Logical-Mathematics;
(b) The textual description of the rule;
(c) The representation in formal logic, that is through a meta-language understandable to the editing application;
(d) If it is a hard or soft rule (i.e. blocker or non-blocker);
(e) If it is an individual or couple rule;
(f) In the case of a couple rule, if it is a symmetrical or asymmetrical rule. This indicates to the application whether it must verify the rule between the individual *i* and the individual *j* and vice versa (asymmetric rule) or if it is sufficient to verify it in only one way (symmetrical rule);
(g) A hierarchy of rules, to indicate to the application the relationship and the order of control of the rules. In the case of a violated rule, all the related rules (which are a specialization of the rule itself) of a subsequent order can be put directly to violated.

In the following paragraph, there is a description of the representation of rules in formal logic.

A. Control rules based on a logic structure

17. This section provides some useful concepts for the definition and representation of a list of check rules and for understanding how to transform the textual rules, defined in the examples described below, in compatibility or incompatibility rules when they are translated in a formal language (Bruni and Sassano, 2000; Bruni and Bianchi, 2012). Rules are expressions typically used to detect, among a possibly large set of elements, the ones verifying some conditions. It is convenient, in order to verify a set of checking rules, to express them using a structure based on propositional logic.

18. Propositional logic, sometimes called sentential logic, may be viewed as a grammar for exploring the construction of complex sentences using as building blocks atomic statements connected by logical connectives. In this type of logic, logical formulas (sentences, propositions) are built up from atomic propositions that are unanalysed. The meaning of these atomic propositions will be known for the specific domain of application. A truth assignment to such atomic propositions will determine the truth value of the whole formula according to the truth rules of the logical connectives. The traditional (symbolic) approach to propositional logic is based on a clear separation of the syntactical and semantical functions. The syntactic deals with the laws that govern the construction of logical formulas from the atomic propositions and with the structure of proofs. Semantics, on the other hand, is concerned with the

interpretation and meaning associated with the syntactical objects. A basic aspect of propositional calculus is that inferences are obtained as purely syntactic and mechanical transformations of formulas.

(a) The set of primary logic connectives {¬, ∨, ∧}, together with the brackets () to distinguish start and end of the field of a logic connective.
(b) The set of proposition symbols, such as $x_1, x_2, ..., x_n$.
(c) The only significant sequences of the above symbols are the well-formed formulas (WFFS). An inductive definition is the following:
(d) A propositional symbol x or its negation $\neg x$.
(e) Other WFFS connected by binary logic connectives (∨, ∧) and surrounded, in case, by brackets.

19. Both propositional symbols and negated propositional symbols are called literals. Propositional symbols represent atomic (i.e. not divisible) propositions, sometimes called atoms. An example of WFF is the following:

$$(\neg x_1 \vee (x_1 \wedge x_3)) \wedge ((\neg(x_2 \wedge x_1)) \vee x_3) \qquad \text{(A.1)}$$

A formula is a WFF if and only if there is no conflict in the definition of the fields of the connectives. In order to simplify the exposition, we will henceforth assume that all our formulas are well formed unless otherwise noted.

20. The calculus of propositional logic can be developed using only the three primary logic connectives above. However, it is often convenient to introduce some additional connectives, such as ⇒ which is called *implies*. They are essentially abbreviations that have equivalent formulas using only the primary connectives. In fact, if S_1 and S_2 are formulas, we have:

$(S_1 \Rightarrow S_2)$ is equivalent to $(\neg S_1 \vee S_2)$.

The elements of the set {T, F}(or equivalently{1,0}) are called truth values with T denoting True and F denoting False. When all the proposition symbols of a formula receive truth values, the truth or falsehood of that formula is obtained according to the truth rules of the logical connectives (considering their appropriate meaning of "not", "or", and "and"). As an illustration, consider the formula (A.1). Let us start with an assignment of true (T) for all three atomic propositions x_1, x_2, x_3. At the next level, of sub formulas, we have $\neg x_1$ evaluates to F, $(x_1 \wedge x_3)$ evaluates to T, $(x_2 \wedge x_1)$ evaluates to T, and x_3 is T. The third level has $(\neg x_1 \vee (x_1 \wedge x_3))$ evaluating to T and $((\neg (x_2 \wedge x_1)) \vee x_3)$ also evaluating to T. The entire formula is the "and" of two propositions both of which are true, leading to the conclusion that the formula evaluates to T. This process is simply the inductive application of the rules:

• S is T if and only if ¬S is F.

• $(S_1 \vee S_2)$ is F if and only if both S_1 and S_2 are F.

• $(S_1 \wedge S_2)$ is T if and only if both S_1 and S_2 are T.

Such a truth evaluation approach can be the basis for developing *control rules*, which are rules that allow the individuation of inconsistent or erroneous data records into a large set of similar records. We denote by P a *record schema*, that is a set of *fields* f_i, with $i = 1...m$, and by p a corresponding *record instance*, that is a set of values v_i, one for each of the above fields.

$P = \{f_1, \ldots, f_m\}$ $p = \{v_1,...,v_m\}$ (A.2)

Each field f_i, with $i = 1...m$, has its *domain* D_i, which is the set of every possible value for that field. Examples of fields f_i are age or marital status, and corresponding examples of values v_i are 18 or single.

21. A control rule should be applied to a generic record and provide a binary value. Therefore, each rule can be seen as a mathematical function r_k from the Cartesian product of all the domains to the Boolean set {0,1}, as follows (see also Fellegi and Holt, 1976; Bruni, 2004; Bruni, 2005).

$$\begin{array}{rccc} r_k: & D_1 \times \ldots \times D_m & \rightarrow & \{0,1\} \\ & p & \mapsto & 0,1 \end{array} \qquad \text{(A.3)}$$

22. The problem of error detection can be approached by formulating a set of rules $R = \{r1, ..., r_t\}$ that are verified by consistent, or correct, records, and are not verified by inconsistent, or erroneous, records. These rules are called compatibility rules, they are such that a generic record p is recognized as a correct record if and only if $r_k(p) = 1$, for all $k = 1, ..., t$. On the other hand, incompatibility rules are verified by erroneous records and not verified by correct records. The detection of erroneous records into a large set of records is a very relevant problem in the field of data E&I.
Compatibility and incompatibility rules can be expressed as disjunction (∨) and/or conjunction (∧) of conditions (also called propositions), hence with the structure of propositional logic formulas. Like to the truth evaluation technique described above, the value of each field of a record under analysis provides a truth assignment for those propositions. The truth/falsehood of the formula constituting the rule provides now the detection of inconsistent or erroneous data records.
However, differently from the case of pure propositional logic, conditions may have an internal structure. It is necessary to distinguish between two different types of structures for the conditions:

(a) A condition involving values of a single field is called a logical condition, and corresponds to an atomic proposition of propositional logic. For instance, (*age* < 14) is a logical condition.
(b) A condition involving mathematical operations between values of fields is called mathematical condition. For instance: (*age - years married* ≥ 14) is a mathematical condition.

We call *logical rules* the rules expressed only with logical conditions, *mathematical rules* the rules expressed only with mathematical conditions, and *logic-mathematical rules* the rules expressed using both types of conditions. For instance, a logical rule expressing that all people declaring to be married should be at least 14 years old is:

$$marital\ status = married \Rightarrow age \geq 14$$

This rule can be represented by the following compatibility rule:

$$\neg\ (marital\ status = married) \vee age \geq 14$$

or, equivalently, by the following incompatibility rule:

$$marital\ status = married \wedge \neg\ (age \geq 14).$$

Instead, a logical-mathematical rule expressing that all people declaring to be married should have the difference between age and years married at least 14 years, is:

$$marital\ status = married \Rightarrow (age - years\ married \geq 14)$$

This rule can be represented by the following compatibility rule:

$$\neg\ (marital\ status = married) \vee (age - years\ married \geq 14)$$

or, equivalently, by the following incompatibility rule:

$$marital\ status = married \wedge \neg\ (age - years\ married \geq 14)$$

23. A formal definition of the structure of the rules allows solving by means of automatic formal methods a number of difficult and computationally demanding problems arising in the different steps of E&I procedures. Examples are:

(a) Problems of error localization (the determination of the erroneous fields of a record).
(b) Problems of imputation (the determination of the correct values for the erroneous fields of a record. This can be done according to a minimum change principle or by means of a data driven approach).
(c) Problems of finding contradictions into the set of rules itself (the determination of a (sub)set or rules determining a logical inconsistency).

Note, in particular, that very effective solution approaches are available when encoding rules into linear inequalities. Indeed, a parallelism can be established between logic formulas and linear constraints, and between atomic propositions and 0-1 variables (see Chandru and Hooker, 1991). The above problems are converted into linear or integer linear programming problems and solved by means of efficient optimization solvers. For further details on those techniques see e.g. Bianchi and Bruni, 2012; Bruni, 2004; Bruni, 2005.

IV. The interactive editing system

24. The interactive editing system allows viewing the data of the questionnaire to check it by using compatibility rules written according to a formal logic, and to modify the data by restoring the situation of coherence between them. The sets of rules used can be prepared, in the census data processing phases, from time to time by the thematic experts and loaded into a database to make them usable to the application.

The rules are divided into two types: blocking rules (Hard) and non-blocking rules (Soft). The application based on the violated rules displays the type of error and the variables (questions) involved in the violated rules. If the violated rule is blocking, the application does not allow the final saving of the data and it is possible to continue only after the corrections of the missing or incorrect data. Obviously, the user is allowed to do partial rescues even in the presence of violated rules, to allow the correction of data in multiple work sessions. For soft rules, the application displays a warning and still allows the data to be saved in the database and to continue with the correction process.

25. All this is implemented through a graphical interface that simulates the graphic layout of the questionnaire, dynamically recreating, through metadata tables, the pages and sections of the questionnaire of interest and making the variables to be managed.

The application allows the administrator to determine which variables cannot be modified, in order to preserve the variables treated by the E&I system in the previous production phases.

Depending on the type and permissions possessed, for example administrator, corrector or viewer, the user has access to the data of the questionnaire for the visualization, the start of the automatic correction procedures, the simple consultation of data or the manual correction of data.

26. The application gives the possibility to visualize all data of the questionnaire inserted by the enumerator by entering the code of the questionnaire and to visualize all the variables inserted by the user during the compilation phase in order to be able to review their correctness and consistency.

The application also features:

(a) search function by questionnaire ID code or by applying other filters, such as the type of questionnaire, the municipality, the province and the region of origin;
(b) the function of checking the rules of a single page of the questionnaire;
(c) the function of checking the rules of a single section of the questionnaire;
(d) the function of checking the rules of the entire questionnaire;
(e) the function of downloading the rules violated in a xls file;
(f) the visualization of the paper of the searched questionnaire;
(g) a section for the management of utilities by the administrator user, for the creation, activation or turn off, modification and cancellation of the user with notification via e-mail;
(h) the function of the versioning of the modified data for each user

As concerns the architecture, the application is a three-tier web application, built in Java 2 Enterprise Edition technology, which runs on Application Server Tomcat 8 and accesses to Oracle 11g Rdbms.

V. Software for data driven imputation (DIESIS)

27. The software (DIESIS) was jointly developed by ISTAT and academic researchers (Department of Computer and Systems Science of the University of Roma “La Sapienza”) (Bruni *et al*., 2001). The DIESIS system allows to deal with qualitative and quantitative variables simultaneously, at household and individual level. After a rigorous statistical evaluation, the DIESIS system was successfully used for

imputing nonresponse and resolve inconsistent responses for the 2001 and 2011 Population and housing census (Bianchi *et al*., 2005; Bianchi *et al*., 2008).

28. Two editing approaches are implemented in the DIESIS system, the *data driven* and the (theorical) *minimum change*, through the *first donors then fields* and the *first fields then donors* algorithms.

The *first donors then fields* algorithm first identifies a subset of potential donors and then determines the minimum number of variables to impute based on these donors. The potential donors are the passed edit households as similar as possible to the failed edit household. The similarity between each failed edit household and each passed edit household is calculated by a function defined as the weighted sum of the distances (for quantitative variables) or similarities (for qualitative variables) for each household variable over all the persons. The algorithm selects, from the potential donors, the minimum (weighted) set of values to impute so that the new adjusted household will pass all the edits (minimum change given the potential donors). By using this algorithm, the imputed values for a household comes from a single donor household.

The *first fields then donors* algorithm first determines the minimum (weighted) number of variables to impute and identifies the potential donors (as previously described). Then, for each recipient person, the algorithm takes the values to impute from the donor person as similar as possible to the recipient one. This algorithm imputes the variables of one person in turn. If possible, the variables inside the person are imputed simultaneously. Note that the imputed values for a household may come from two or more donor households.

29. The two algorithms were jointly used for the treatment of the demographic variables, in order to balance the plausibility of the imputation actions with the preservation of the collected information. The *first donors then fields* algorithm was selected as default one, with the option to turn to the *first fields then donors* algorithm when, for a given failed edit household, the number of changes proposed by the first algorithm was exceedingly high in comparison with the number of changes proposed by the second algorithm (the extent was set on the basis of the household size).
The *first donors then fields* algorithm was mainly used to process the households having common structure that are usually those having smaller household size. For these households it was generally possible to find enough potential donors. Otherwise, in the treatment of households having uncommon structure, usually those with largest size, few donors were generally available, and often they were not very similar to the failed edit household. In these cases, the data driven imputation action would have required many changes to obtain an adjusted household passing the edits, therefore the minimum change approach was preferred.

VI. The use of Big Data for data validation

30. In the strategy defined for data E&I of section "place of study and work" the correction of a few inconsistent data relating to the routes (common destination for commuting), the time and means used for commuting was an interesting example of big data usage.

31. In order to validate the individual routes and means, a first comparison between the detected data and the 2011 validated data is performed. If the detected routes was already valid in 2011, it is automatically valid in 2018. For all other combinations of routes and means, distances between municipality centroids are calculated were centroid information is available. For some residual routes, we have developed a web scraping procedure that allows analysing and validating combinations of routes travel times and means.

32. The web scraping procedure (web automation) uses Selenium headless browser. This software allows simulating the behaviour of a user who visits the website and collects all the information of interest. In particular, the procedure accesses to the portals that allows to create customized maps and offer georeferenced research services by providing information relating to the place of origin and destination. The routes are validated for four modes of transport: car, public transport, bicycle, on foot.

33. Using this web scraping procedure on all detected routes could give origin to an important information asset gathering data on distances, means of transport and travel time of paths between

municipalities. This set of information could rise in time using information of yearly census surveys giving origin

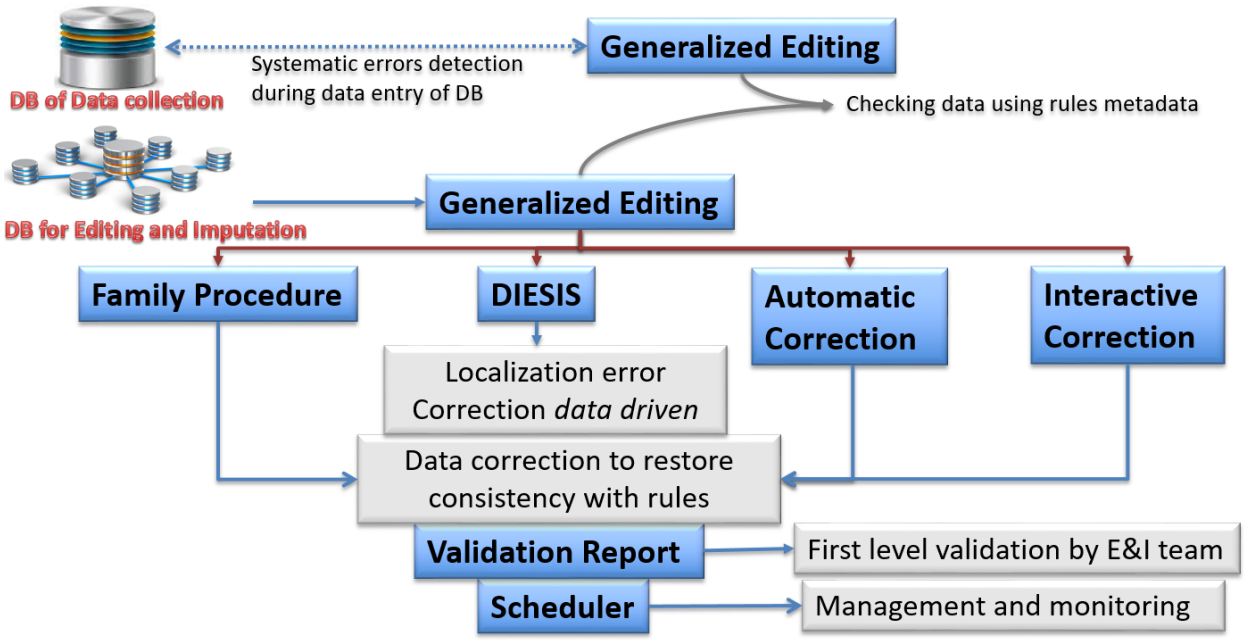
VII. Management system (DEIS)

34. DEIS is a generalized multiplatform application for the management, scheduling and monitoring of procedures carried out for the processing of census data.
The application executes and controls a set of procedures grouped in a data structure called "container". The container also takes into account the launch parameters associated with the procedures contained inside. Therefore, parallel execution of two or more containers with the same procedures but different parameters is possible.
The Access to DEIS is regulated by authentication.
DEIS allows the definition of system users. In particular, there are three types of users:

(a) the administrator: can access every function provided by the system and create new users with different privileges;
(b) the scheduler: can start, and interrupt, the execution of the containers created and defined by the administrator and monitor the progress of each process associated with the container;
(c) person who monitors: can only check the processing status of the processes.

Based on the roles attributed to a user, the application shows the permitted functions.

Figure 2. The Editing and Imputation processing cycle



The application is developed in Java according to the MVC Pattern. Once logged in, a different menu will be shown based on the type of user logged in (administrator, scheduler, monitor).

35. DEIS manages the functionalities showed in the Figure 2 of the E&I processing cycle used for the 2018 Italian Permanent census.

36. DEIS allows the interaction between procedures developed with the various programming languages. The path and the parameters for launching the procedures are metadata saved on the DB.

The system have a container start screen and a monitoring screen. In the first one the containers started have the icon of the little man running, the ones to start have the "Start" button. The traffic light shows

any execution errors. In the monitoring screen, is possible to view a procedure while it is running with the execution percentage; it is possible to stop the container or the single procedure.

VIII. Conclusion

37. The first Italian Permanent census held in 2018 required a review of the overall Editing and Imputation process to improve data accuracy and consistency, taking into account the new survey and the use of statistical registers for comparative analyses. The strategy adopted has been defined by introducing new generalized solutions and improving standard methodologies based on technological innovations. For processing census data a generalized Data Editing and Imputation System (DEIS) for the management, scheduling and monitoring of IT procedures has been realised. The whole process of Editing and Imputation can be generalized and adapted to the second Permanent census of the population, as to other social surveys.

References

Bianchi G., A. Manzari and A. Reale, An overview of editing and imputation methods for the next italian censuses (Key Invited Paper), Conference of European statisticians, Work Session on Statistical Data Editing, Geneva, 13-15 May 2008.

Bianchi G., A. Manzari, A. Pezone, A. Reale and G. Saporito, New procedures for editing and imputation of demographic variables, Conference of European statisticians, Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005.

Bruni R. and A. Sassano, Solving Propositional Satisfiability by Identification of hard Subformulae, in Proceedings of the 17th International Symposium on Mathematical Programming (ISMP2000), Atlanta, USA
2000

Bruni R. and G. Bianchi, A Formal Procedure for Finding Contradictions into a Set of Rules. Applied Mathematical Sciences 6 (126), 6253-6271, 2012.

Bruni R. and G. Bianchi, Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis, IEEE transactions on knowledge and data engineering, vol. 27, no. X, XXXXX 2015.

Bruni R., A. Reale and R. Torelli, Optimization Techniques for Edit Validation and Data Imputation. In *Proceedings of Statistics Canada Symposium: Achieving Data Quality in a Statistical Agency*, Ottawa, Canada, 2001.

Bruni R., Discrete Models for Data Imputation. Discrete Applied Mathematics Vol. 144(1), 59-69, 2004.

Bruni R., Error Correction for Massive Data Sets. Optimization Methods and Software, Vol. 20(2-3), 295-314, 2005.

Chandru V. and J.N. Hooker, Extended Horn sets in propositional logic. J. ACM 38, 205-221, 1991.

Fellegi I.P. and I.P.D. Holt, A Systematic Approach to Automatic Edit and Imputation, Journal of the American Statistical Association 71, 17-35, 1976.