Machine Learning for Data Editing Cleaning in NSI (Editing & Imputation): Some ideas and hints

Fabiana Rocci
-Istat-

1. Introduction

What is Data Editing-E&I

Data that have been collected by a statistical institute inevitably contain errors, this is true for both data obtained by means of surveys and data originating from external sources, as Administrative Data. An error occurs when the observed value is different from the actual "true" value. Typically they are called non sampling errors. The event of occurrence of such kind of errors can be characterized by different aspects, which the most important are considered to be their source, their nature, their impact on data and how they appear in data.

In order to produce statistical output of sufficient quality, it is important to detect and treat them. The process that regards the set of all actions to detect and treat errors is referred to as *statistical data editing* or *Editing & Imputation*.

Statistical institutes carry out an extensive process of checking the data and performing amendments, to improve the data quality.

Several schemes for detecting errors are proposed, most literature about E&I starts from classifying the types of errors, according to several criteria, in order to understand for each phase of a process which kind of errors are most probable to happen and which method can be defined to detect them.

One main important consideration is that only sometimes it is possible to find out a data to be surely erroneous. But other times it is only possible to regard a data as suspicious, because it sounds to be strange with regards to some characteristic of the data that is expected to be respected. In these cases, further evaluation is needed to understand whether it is a non sampling error or instead it is a correct but anomalous data.

In this sense, Editing is considered as being the problem to define the right function in order to *label* records between correct and erroneous/suspicious with regard to several controls.

Schemes usually starts from describing the error according to its the nature and to its effect on the final target estimation, to provide a kind of guidance of which method and in which point of the process it better suites to the errors to be identified and treated.

What is ML

Machine Learning (ML) is the science of getting computers to automatically learn from data and generalize the acquired knowledge to new settings. It represents a way to use computers to perform tasks that require the ability to learn from experience, to help in gaining a new level of understanding from the data.

ML uses a variety of algorithms that iteratively learn from data to describe data and predict outcomes.

Broadly speaking, there are two ways of approaching ML:

a. Supervised ML

Supervised learning is when a dataset is available, where two sets of variables can be regarded as being input (X) and output variables (Y), hence a linking function between the two set of variables is known to exists. An ML algorithm can learn the mapping function from the input to the output:

$$Y = f(X)$$

The goal is to approximate the mapping function so well that when a new input data (X) is available, then it is possible to predict the output variables (Y) for those data.

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. Supervised learning problems can be further grouped into regression and classification problems.

- **Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.
- **Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

The algorithms are trained using preprocessed examples, *called training set* and at first performance of the algorithms is evaluated with test data.

b. Unsupervised learning

Unsupervised ML is where only a set of input data (X) is available and there are not any corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. These are called unsupervised learning because unlike supervised learning above there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data.

Unsupervised learning problems can be further grouped into clustering and association problems.

- **Clustering**: A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association**: An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

Some popular examples of unsupervised learning algorithms are for example: k-means for clustering problems, apriori algorithm for association rule learning problems.

Unsupervised learning algorithms segment data into groups of examples (clusters) or groups of features.

Our task

To analyse/understand to which extent ML algorithms can be used to efficiently improve Editing and Imputation.

In order to identify which aspect and/or step of a E&I process can benefit from ML technique, we need to start from a general scheme that represent the main issues about E&I.

A kind of synthesis among several schemes and approaches will be done, to try to delineate a list of topics, for kind of errors and methods to treat them, to understand for each of them whether ML technique is already used, or if it is thought to be implemented and in which way. The main aspects will evaluated are related to the capacity of ML:

- to learn from data;
- to deal with huge amount of data, for which to look for unknown underlying pattern can be very challenging.

2. E&I reference scheme

The GSDEM is envisaged as standard references for statistical data editing from a methodological point of view, by providing standard terminology and models, the GSDEM aims to facilitate understanding, communication, practice, assessment and development in the field of statistical data editing[1].

The Generic Statistical Data Editing Model (GSDEM) was developed under the High-Level Group for the Modernisation of Official Statistics (HLG-MOS). It is intended as a reference for all official statisticians whose activities include data editing.

The GSDEM is designed to be consistent with other standards and models related to statistical modernisation, in particular, the Generic Statistical Business Process Model (GSBPM)[2] and the Generic Statistical Information Model (GSIM)[3]. It should be seen as part of the coherent toolkit of models and standards promoted by the HLG-MOS under the “ModernStats” initiative.

Together with other guidelines (as Edimbus, MEMOBSUT, etc.), it is possible to picture a scheme according which to approach to the E&I issue.

[1] The statistical data editing (SDE) process can be interpreted according to the Generic Statistical Information Model (GSIM).

[2] The GSBPM version used throughout this document is GSBPM v5.1. For more, see UNECE GSBPM Wiki: http://www1.unece.org/stat/platform/display/GSBPM

[3] GSIM provides a set of standardized, consistently described information objects that are the inputs and outputs in the design and production of statistics. The GSIM version used throughout this document is GSIM v1.2. For more, see UNECE GSIM Wiki: https://statswiki.unece.org/display/gsim

The starting point is based on the various consideration of type of errors, according to the source and nature of them. GSDEM delivers definition of the functions, that specify *what* action is to be performed in terms of its purpose, and for each category of functions the type of methods, that specify *how* the required action is performed. Hence, process step are defined as a set of functions to be designed accordingly to the type of process and to the type of errors to be treated step by step during the process flow. An introduction to all those concepts is done, to achieve the final scheme where they are puzzled all together and ML methods can be better pictured as useful or not.

1.1 Errors

The nature of non sampling error can be roughly divided in systematic and random. In general, systematic errors have a deterministic effect once they occur, while random errors are characterized by a probability distribution with its own variance and expectation. This distinction is also important since the techniques used to deal with systematic and random errors are significantly different (EDIMBUS, 2007).

On the other hand, the impact of an error on the final estimates is also an important aspect. It is not an absolute concept, but depends on the target parameter and/or on the aggregation level: in fact, an error may be influential with respect to a certain estimate, but it may have a negligible impact on some other. This characteristic should be considered when balancing the trade-off between data accuracy and costs of E&I. Finally, the way the errors appear in data (outlier, missing, and so on) will influence the choice of methods to be used to detect them in the E&I process. All the above mentioned aspects overlap, so in order to deal with the variety of problems concerning error detection and treatment, all of them are to be considered.

So that, the most common scheme about type of errors is:

- **Systematic errors:** they can have several effect on the data, the main ones are:
  - **Domain**: (in terms of units and variables) Check of structural informative objects defining the target population and the variables: e.g., verification and selection of eligible units, classification variables (e.g. ISIC/NACE, legal status)
  - **Obvious and other systematic errors**: Obvious errors are the ones easily detectable and treatable. The remaining Systematic errors, that are less recognizable than the previous ones, can be reported consistently across units in the same release of the survey or over time in repeated surveys by responding units. It is a phenomenon caused either by the consistent misunderstanding of a question during the collection of data, or by consistent misinterpretation of certain answers in the course of coding. Systematic errors do not lead necessarily to consistency errors but always seriously compromise statistical results (because they lead to bias in an estimate. A well-known type of systematic error is the so called unity measure error.
- **Influential errors**: Influential errors are errors in values of variables that have a significant influence on publication target statistics for those variables. Strictly related to the concept of influential error is that of influential observation. An influential observation is an observation that has a large impact on a particular result of a survey, i.e. a statistic.
- **Completely at random not influential errors** Random errors are errors that are not caused by a systematic reason, but by accident and they do not result to be influential. They primarily arise due to in-attention by respondents, interviewers and other processing staff during the various phases of the survey cycle. An example of a random error is an observed value where a respondent by mistake typed in a digit too many. Random errors are often defined as non systematic errors. In the statistical context the expectation of a random error is typically zero. In our context, however, the expectation of a random error may also differ from zero. This is, for instance, the case in the above-mentioned example.

GSDEM provides schemes according to the type of errors and the characteristic of the process under study, to propose function and method to be used in order to at first detect errors, to select units or variable to be amended and then to choose which method better impute the value to make data consistent.

Different types of methods are currently used, both for the detection and for the treatment of the different types of errors. Methods are characterized by different aspects as well. They differ by the kind of information that is used, how it is used, the type of error that can be detected and treated and the amount of resources needed.

Hence the process step are introduced to identify which method for which kind of error at which phase of the process is suggested to be designed.

1.2 Function

The GSIM refers to a *business function* as “something an enterprise does, or needs to do, in order to achieve its objectives”. A *statistical data editing function* is a business function that performs a specific purpose in the chain of activities defining the data editing process, and can be categorized into three broad function types.

Data editing The functions are classified into function categories, which refer to the task they are assigned to, the type of output they produce and whether they apply to units or variables. Of course, other classifications based on different criteria are possible as well, however we feel this classification can be used widely. The descriptions of the function categories are as follows:

- **Review**. Functions that examine the data to identify potential problems
- **Selection**. Functions that select units or fields within units for specified further treatment
- **Treatment**. Functions that change the data in a way that is considered appropriate to improve the data quality. The modification of specific fields within a unit (i.e. filling in missing values or changing erroneous ones) is referred to as imputation

1.3 Rules and Methods

A process method specifies how (parts of) the data editing functions in a process flow are to be performed in real life situations. A method may rely on a set of rules defining its functionality in a specific context, that can be distinguished as follows:

- **Edit-rules** An edit rule, or edit, is a restriction to the values of one or more data items that identifies missing, invalid or inconsistent values. Edits are often distinguished between *fatal* or *hard* edits and *query* or *soft* edits, depending on whether they identify errors with certainty or not. They aim at describing the valid (*hard*) or plausible (*soft*) values of variables or combinations of variables.
  Edit (as imputation) techniques can be divided into two main classes, depending on the kind of data: techniques for numerical data and techniques for categorical data. Generally, numerical data occur mainly in surveys on businesses whereas categorical data occur mainly in social surveys—for instance, surveys on persons or households. In business statistics there are often large sets of hard and soft edit rules such as linear equalities (balance edits), inequalities and ratio edits (soft edits). Hard edit rules are also used by methods for selection of values presumed to be in error, e.g. implementations of the Fellegi-Holt method.

- **Score functions** asses the plausibility and influence of the values in a unit as whole. They are functions defined to release a score to measure of the plausibility of a data being error or not, together with an influence on the final target estimate. Hence it has to be the result of some method that relates each observation to the whole set of data of the aggregation the estimates is about.
  For those type of errors, there are many methods specifically designed. Anyway, sometimes also outlier detection methods are used, event thought they do not produce a score function based also on a probability of being erroneous, they can help in guiding to put more attention on a set of units that sound to be too much different form the rest of the data distribution. Further methods should then help in understanding whether they are a different population or they are like that because of an error.

- **Correction rules** combine detection, selection and imputation of missing data or erroneous values. In particular, in case of specific “obvious” errors, they are used for the correction of systematic errors or, more generally, of errors with a detectable cause and known error mechanism. They can be formulated as IF-THEN type rules of the following form: IF (condition) THEN OldValue = NewValue. This type of rules is usually applied during micro-editing. IF-THEN type rules can also be used for automatic error detection. They can be expressed in IF-THEN form as: IF (condition) THEN FlagValue = ErrorCode.

Sometimes these rules may further be fine-tuned using parameters. Methods are defined according to the function they serve for and the type of rule they are based on.

Table 1. The Main Process Steps of GSDEM Process

Process steps	Function types	Methods (how)
Domain/obvious errors editing	Review, Selection	IF-THEN
	Review, Selection, Treatment	IF-THEN
Editing other systematic errors	Review, Selection, Treatment	IF-THEN
	Review	Cluster analysis, latent class analysis, edit rules, graphical editing (e.g. log for 1000 error)
	Selection	IF-THEN, cluster analysis, latent class analysis
	Treatment	Deductive imputation, model-based imputation
Selective editing	Review	Score calculation
	Selection	Selection by fixed threshold
Automatic editing	Review	Analysis of edit failures
	Selection	IF-THEN, Fellegi-Holt paradigm, NIM (Nearest-Neighbour Imputation Method)
	Treatment	IF-THEN, deductive imputation, non-random imputation, random imputation, prorating, NIM
	Treatment	IF-THEN, deductive, non-random imputation, random imputation, NIM
Macro editing	Review, Selection	Outlier analysis, aggregate comparison within data set, aggregate comparison with external sources, aggregate comparison with results from history

3. ML for E&I

As reported in the Beck et al. (2018), ML methods are often used in NSI for classification, identification and imputation. Although the participating institutions pursue very different tasks and therefore have very different project objectives, it can be stated that problems tackled with ML procedures are often similar ones, that often also involves similar methods.

It resulted that with regards the E&I related actions during a survey process, ML algorithms are already used mostly for the imputation part, that corresponds to the *treatment* function as defined in the GSDEM.

Much less experiences have been counted relating the Editing part, i.e. for the *review* and *selection* functions.

In this sense, this project has the task to analyse how the ML algorithms can be used to help in detecting possible non sampling error into the surveyed data, in particular how ML can help in improving the current methods in terms of timeliness and efficiency of results. Hence, with regards the E&I GSDEM process, the ML algorithms can be related to the Methods to be used to determine how to put in action the function.

As it has been described, data editing processes and procedures are governed by an interaction between available computer technologies and decision-making 'rules': it has been shown how much important is to define a proper rule criteria in order to detect a non correct data or at least a suspicious one.

Very few can be done when variables are checked trough hard edits, on the other side when rules leave only a certain degree of suspiciousness there is field for improvement in finding hidden rules.

1. Hard edits: they release data divided between correct and not correct. In the second case, in some cases it is possible to know which data of the record is wrong, in some other case not (for example, the sum is difference from the total, but it is not possible to know which component of the several is wrong). In this case, we need to define the method for the selection function.
2. Soft editing: they usually describe plausible relationship among variables, sometime these are surveyed in the same process, sometimes they relate to other sources as external auxiliary information. Hence, those methods release data classified in two set one of plausible and not

plausible data correct. In the second case, as it happens for the hard edit, in some cases it is not possible to understand which data out of the entire record is the one to be wrong.

But there is a major reflection about the definition of the rules: these are of different nature than the hard ones, because they come from hypothesis about the relationship between the data and some other characteristic. But those rules are not to be taken for given, no across the several units neither along the time on the same units.

This is the reason why the final set to be further analysed is labelled as 'not plausible', to get to the finale label of being erroneous.

When soft edits are run, the final labelling of 'right' or 'wrong' is obtained through further methods or through expert human intervention. In both cases, the final result can come to a high cost, or because of the timing occurring to run some more methods or because of an human intervention is asked, hence with a degree of subjective, even very expert, approach. This situation can be very costly when the number of variables is huge and the constraints are a lot and soft.

3. Score function and outlier detection: they are methods that release a flag to be potentially erroneous and critical of the final estimates and non critical, because they release an indicator based on a plausibility concept, the problem related to them can be considered similar to the case where soft edits are put on the data.

   For those type of errors, there are many methods specifically designed. Sometimes also outlier detection methods are used, event thought they do not produce a score function based on a probability of being erroneous, they can help in guiding to put more attention on a set of units that sound to be too much different form the rest of the distribution. Further methods should then help in understanding whether they are a different population or they are like that because of an error.

2.1 ML possible value added: some hints

It is asked if ML methods can be used for editing (intended as detection of non sampling error) and which value added they could deliver.

Going through the analysis of the main criteria on which the detection of non sampling errors is based, it sounds that there is a need of further developments in two main cases:

a. Supervised ML: it can be thought to be used when a dataset of data where data are labelled already between 'plausible' and 'not plausible' is available, this can be regarded as the *training set*.
   Hint:
   i. it is possible to approximate the function labelling the data as correct or not, this is possible where rules are suppose not to change over time.
   ii. it is possible to run a ML procedure in order to identify the whole rules describing the decision of labelling as wrong, this can help in completing the soft rules with the reasoning of the human intervention whether this hide a major reasoning that could not be identified at first.

b. Unsupervised ML: it better suits when the problem deals with data that are unlabelled and/or it is believed that data are characterized by completely hidden patterns of errors. In this sense, unsupervised ML has the potential to identify relationships among variables.
   Hint:
   i. If the problem is thought to be faced through edit rules, an unsupervised ML approach can help to classify the data based on the patterns or clusters. In essence, this process adds labels to the data so that it becomes supervised. Therefore, unsupervised learning can be used as the first step before passing the data to a supervised learning process.
      This can be useful in both cases when edit rules are not known, that happens for new data about phenomena not studied until that point, or when edit rules and error pattern are though could change over time, because the underlying pattern can change.
   ii. Where the problem is to identify group of data, this could lead to outliers detection or to pattern detection of errors strange data with regard the rest of the distribution, this can happen when a record does not fail any edit, but still they are though they could contain errors. In this regards, the wish is to learn the inherent structure of our data without using explicitly-provided labels. Since no labels are provided, there is no specific way to compare model performance in most unsupervised learning methods, this is why a common use-cases for unsupervised learning are exploratory analysis because it can automatically identify structure in data.

Generally speaking, ML methods could be used to achieve a better independency from the hypothesis to be done at first in order to classify observation among group, in our case between 'right' and 'wrong' (or at least between plausible or not). This can done through a training set, where labels are available over a representative dataset of the data, in order to approximate the function of labelling and re-producing it on new data.

As far this is possible, as a further result could help in revising the production and to invest less on human intervention or at least to lead it to a minor number of controls, for example to the ones appearing with a major risk of being wrong.

On the other side, it can help in discovering hidden patterns, that would help in an understanding the data beyond the mentioned hypothesis coming from previous analysis of the data. This could help in learning how to formulate new hypothesis. This side of the problem seems to be particularly useful in front of the new kind (big data) of data ONs are approaching to use, that hide so much difference from the usual type of data are treated until now.

References

Martin Beck, Florian Dumpert, Joerg Feuerhake (2018). Machine Learning in Official Statistics

Waal, T.de, Pannekoek, J. and Scholtus, S. (2011). Handbook of Statistical Data Editing and Imputation. Wiley, Hoboken.

EDIMBUS (2007). Recommended Practices for Editing and Imputation in Cross-sectional Business Surveys, EDIMBUS project report, https://ec.europa.eu/eurostat/documents/64157/4374310/30-Recommended+Practices-for-editing-and-imputation-in-cross-sectional-business-surveys-2008.pdf.

GSDEM (2019). Generic Statistical Data Editing Models - GSDEMs, Version 2.0, April 2019, UNECE. Available at: https://statswiki.unece.org/display/sde/GSDEM

GSBPM (2019). Generic Statistical Business Process Model. Version 5.1, January 2019, UNECE. Available at: https://statswiki.unece.org/display/GSBPM/Generic+Statistical+Business+Process+Model.

GSIM (2019). Generic Statistical Information Model, Version 1.2, May 2019, UNECE. Available at: http://www1.unece.org/stat/platform/display/gsim.

MEMOBUST (2014). Handbook on Methodology of Modern Business Statistics, CROS-portal, Eurostat, https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en.

Van del Loo M. (2015) *A Formal Typology of Data Validation Functions*, UNECE, Conference of European Statisticians, Budapest. Available at: http://www.markvanderloo.eu/files/statistics/WP_5_Netherlands_A_formal_typology_of_data_validation_functions.pdf