Statistical Data Editing

Claudio Conversano
University of Cassino, Italy

Roberta Siciliano
University of Naples Federico II, Italy

INTRODUCTION

Statistical Data Editing (SDE) is the process of checking data for errors and correcting them. Winkler (1999) defined it as the set of methods used to edit (i.e., clean up) and impute (fill in) missing or contradictory data. The result of SDE is data that can be used for analytic purposes.

Editing literature goes back to the 1960s with the contributions of Nordbotten (1965), Pritzker, et al. (1965), and Freund and Hartley (1967). A first mathematical formalization of the editing process was given by Naus, et al. (1972), who introduced a probabilistic criterion for the identification of records (or part of them) that failed the editing process. A solid methodology for generalized editing and imputation systems was developed by Fellegi and Holt (1976). The great break in rationalizing the process came as a direct consequence of the PC evolution in the 1980s, when editing started to be performed online on personal computers, even during the interview and by the respondent in CASI models of data collection (Bethlehem et al., 1989).

Nowadays, SDE is a research topic in both academia and statistical agencies. The European Economic Commission organizes a yearly workshop on the subject that reveals an increasing interest in both scientific and managerial aspects of SDE.

BACKGROUND

Before the advent of computers, editing was performed by large groups of persons undertaking very simple checks. In that stage, only a small fraction of errors was detected. The advent of computers was recognized by survey designers and managers as a means of reviewing all records by consistently applying even sophisticated checks requiring computational power to detect most of the errors in the data that could not be found by means of manual review. The focus of both the methodological work and, in particular, the applications was on the possibilities of enhancing the checks and applying automated imputation rules in order to rationalize the process.

SDE Process

Statistical organization periodically performs an SDE process. It begins with data collection. An interviewer can examine quickly the respondent answers and highlight gross errors. Whenever the data collection is performed using a computer, more complex edits can be stored in it and can be applied to the data just before they are transmitted to a central database. In all these cases, the core of the editing activity is performed after completing the data collection. Nowadays, any modern editing process is based on the a priori specification of a set of edits. These are logical conditions or restrictions on the values of data. A given set of edits is not necessarily correct. It may omit important edits or contain edits that are conceptually wrong, too restrictive, too lenient, or logically inconsistent. The extent of these problems is reduced by having subject-matter experts specifying the edits. Problems are not eliminated, however, because many surveys involve large questionnaires and require hundreds of edits, which makes their specification a very demanding task. As a check, a proposed set of edits is applied on test data with known errors before application on real data. Missing edits or logically inconsistent edits, however, may not be detected. Problems in the edits, if discovered during the actual editing or even after it, cause editing to start anew after their correction, leading to delays and incurring larger costs than expected. Any method or procedure that would assist in the most efficient specification of edits would, therefore, be welcome.

The final result of an SDE process is the production of clean data as well as the indication of the underlying causes of errors in the data. Usually, editing software is able to produce reports indicating frequent errors in the data. The analysis of such reports allows the researcher to investigate the causes of data error generation and to improve the results of future surveys in terms of data quality. The elimination of sources of errors in a survey allows a data collector agency to save money.

SDE Activities

SDE concerns two different aspects of data quality; namely, data validation (the activity concerning the correction of logical errors in the data) and data imputation (the activity concerning the imputation of correct values once errors in the data have been localized). Whenever missing values appear in the data, missing data treatment is part of the data imputation process to be performed in the framework of SDE.

Types of Editing

It is possible to distinguish among different kinds of editing activities:

- **Micro Editing:** concerns the separate examination of each single record aimed at examining the logical consistency of the data it contains using a mathematical formalization of the automation of SDE.
- **Macro Editing:** concerns the examination of the relationships among a given data record and the others, in order to account for the possible presence of errors. A classical example of macro editing is outlier detection. It consists of the examination of the proximity between a data value and some measures of location of the distribution to which it belongs. Outlier detection methods literature is vast, and it is possible to refer to any of the classical text in the subject (Barnett & Lewis, 1994). For compositional data, a common outlier detection approach is provided by the aggregate method, aimed at identifying suspicious values (i.e., possible errors) in the total figures and to drill down to their components to figure out the sources of errors. Other approaches are based on the use of data visualization tools (De Waal et al., 2000) as well as on the use of statistical models describing the change of data values over time or across domains (Revilla & Rey, 2000).
- **Selective Editing:** can be meant as a hybrid between micro and macro editing. Here, the most influential among the records that needs imputation is identified, and the correction is made by human operators, whereas the remaining records are automatically imputed by the computer. Influential records often are identified by looking at the characteristics of the corresponding sample unit (e.g., large companies in an industry survey) or by applying the Hidiroglou-Berthelot score variable method (Hidiroglou & Berthelot, 1986), taking account of the influence of each subset of observations on the estimates produced for the whole data set.
- **Significance Editing:** a variant of selective editing introduced by Lawrence and McKenzie (2000). Here, the influence of each record on the others is examined at the moment the record is being processed and not after all records have been processed.

MAIN THRUST

The editing literature does not contain many relevant suggestions. The Fellegi and Holt method is based on set theory concepts that help to perform several steps of the process more efficiently. This method represents a milestone, since all the recent contributions are aimed at improving (even in a small part) the Fellegi-Holt method, with particular attention to its computational issues.

The Fellegi-Holt (FH) Method

Fellegi and Holt (1976) provided a solid mathematical model for SDE, in which all edits reside in easily maintained tables. In conventional editing, thousands of lines of if-then-else code need to be maintained and debugged.

In the Fellegi-Holt (FH) model, a set of edits is a set of points determined by edit restraints. An edit is failed if a record intersects the set of points. Generally, discrete restraints have been defined for discrete data and linear inequality restraints for continuous data. An example for continuous data is $\sum_i a_{ij} x_j \le C_j, \forall j = 1, 2, \ldots\ldots, n$, whereas for discrete data, edits can be specified in the form $\{Age \le 15,\ marital\ status = Married\}$. If a record r falls in the set of restraints defined by the edit, then the record fails the edit. It is intuitive that one field (variable) in a record r must be changed for each failing edit. There is a major difficulty: if fields (variables) associated with failing edits are changed, then other edits that did not fail originally will fail.

The code of the main mathematical routines in the FH model can be maintained easily. It is possible to check the logical validity of the system prior to the receipt of data. In one pass through the data of an edit-failing record, it is possible to fill in and change values of variables so that the record satisfies all edits.

Checking the logical validity often is referred to as determining the consistency or logical consistency of a set of edits. The three goals of the FH methods are as follows:

1. The data in each record should be made to satisfy all edits by changing the fewest possible variables.
2. Imputation rules should derive automatically from edit rules.