Statistical Data Editing

Claudio Conversano
University of Cagliari, Italy

Roberta Siciliano
University of Naples, Federico II, Italy

INTRODUCTION

Statistical Data Editing (SDE) is the process of checking and correcting data for errors. Winkler (1999) defines it the set of methods used to edit (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 60's with the contributions of Nordbotten (1965), Pritzker et al. (1965) and Freund and Hartley (1967). A first mathematical formalization of the editing process is in Naus et al. (1972), who introduce a probabilistic criterion for the identification of records (or the part of them) that failed the editing process. A solid methodology for generalized editing and imputation systems is developed in Fellegi and Holt (1976). The great break in rationalizing the process came as a direct consequence of the PC evolution in the 80's: Editing started to be performed on-line on PCs even during the interview and by the respondent in computer assisted self-interviewing (CASI) models of data collection (Bethlehem et al., 1989).

Nowadays, SDE is a research topic in academia and statistical agencies. The European Economic Commission periodically organizes a workshop on the subject concerning both scientific and managerial aspects of SDE (www.unece.org/stats).

BACKGROUND

Before the computers advent, editing was performed by large groups of persons undertaking very simple checks and detecting only a small fraction of errors. The computers evolution allowed survey designers and managers to review all records by consistently applying even sophisticated checks to detect most of the errors in the data that could not be found manually. The focus of both methodologies and applications was on the possibilities of enhancing the checks and of applying automated imputation rules to rationalize the process.

SDE Process

Statistical organizations periodically perform a SDE process. It begins with data collection. An interviewer can quickly examine the respondent answers and highlight gross errors. Whenever data collection is performed using a computer, more complex edits can be stored in it in advance and can be applied to data just before their transmission to a central database. In such cases, the core of editing activity is performed after completing data collection. Nowadays, any modern editing process is based on the a-priori specification of a set of edits, i.e., logical conditions or restrictions on data values. A given set of edits is not necessarily correct: important edits may be omitted and conceptually wrong, too restrictive or logically inconsistent edits may be included. The extent of these problems is reduced by a subject-matter expert edits specification. Problems are not eliminated, however, because many surveys involve large questionnaires and require the complex specification of hundreds of edits. 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 ones, however, may not be detected at this stage. 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 which would assist in the most efficient specification of edits would therefore be welcome.

The final result of a SDE process is the production of clean data and the indication of the underlying causes of errors in the data. Usually, an editing software is able to produce reports indicating frequent errors in the data. The analysis of such reports allows to investigate the data error generation causes and to improve the results

of future surveys in terms of data quality. Elimination of sources of errors in a survey allow a data collector agency to save money.

SDE Activities

SDE concerns two aspects of data quality; (1) Data Validation: the correction of logical errors in the data; (2) Data Imputation: the imputation of correct values once errors in data have been localized. Whenever missing values appear in data, missing data treatment is part of the data imputation process to be performed in the SDE framework.

Types of editing

The different 'kinds' of editing activities are:

- **Micro Editing:** The separate examination of each single record for the assessment of the logical consistency of data, using a mathematical formalization in the automation of SDE.
- **Macro Editing:** Examination of the relationships between a given data record and the others, in order to account for the possible presence of errors. A classical example is outlier detection, i.e. the examination of the proximity between a data value and some measures of location of the distribution it belongs to. Outlier detection literature is vast and it is possible to refer to any of the classical text in the subject (for instance Barnett and Lewis, 1994). For compositional data, a common outlier detection approach is provided by the aggregate method, aimed to identify 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 use both data visualization tools (De Waal et al., 2000) and statistical models describing changes of data values over the time or across domains (Revilla and Rey, 2000).
- **Selective Editing:** An hybrid between micro and macro editing: the most influential among the records that need imputation are identified and their correction is made by human operators, whereas remaining records are automatically imputed by the computer. Influential records are often identified looking at the characteristics of the corresponding sample unit (e.g. large companies in an industry survey) or applying the "score variable method" (Hidiroglou and Berthelot, 1986) that accounts for the influence of each subset of observations on the estimates produced for the whole dataset.
- **Significance Editing:** A variant of selective editing introduced by Lawrence and McKenzie (2000). The influence of each record on the others is examined at the moment the record is processed and not after all records have been processed.

MAIN THRUST

Editing literature does not contain relevant suggestions. The Fellegi-Holt method is based on set theory concepts, which helps to perform efficiently several steps of the process. This method represents a milestone, since all recent contributions are aimed to improve it, particularly its computational effectiveness.

The Fellegi-Holt Method Data

Fellegi and Holt (1976) provide 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 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 are 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 an edit is specified in the form $\{Age \le 15,\ marital\ status = Married\}$. The record *r* falling in the set of edit restraints fails the edit. It is intuitive one field (variable) in a record *r* must be changed for each failing edit. A major difficulty arises if fields (variables) associated with failing edits are changed: then, other edits that did not fail originally will fail.

The mathematical routines code in the Fellegi-Holt model can be easily maintained. It is possible to check