

WP.30
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing

(Bonn, Germany, 25-27 September 2006)

Topic (v): New and emerging methods

A NEW SOFTWARE FOR DATA EDITING

Supporting Paper

Prepared by Sergio Delgado-Quintero and Juan-José Salazar-González, DEIOC,
Universidad de la Laguna, Spain

A new software for Data Editing

Sergio Delgado-Quintero and Juan-José Salazar-González

{sdelquin, jjsalaza}@ull.es

DEIOC, Universidad de La Laguna, 38271 Tenerife, Spain

June 30, 2006

Abstract

Since three years, a team at University of La Laguna (Tenerife, Spain) is developing an automatic software to help statistical agencies in data editing. The software is a friendly graphical interface under which several optimization algorithms have been implemented. These algorithms deal with the combinatorial problems of finding the minimum number of fields to modify in an incorrect record, and the problem of finding the "best" imputed value. Several optional methods are available in the software to define "best". Some methods are more suitable for continuous data, and are related to multi-regression statistical analysis. Other are more suitable for categorical data, and are related to donor-record techniques. The implementation of the methods relies on modern results in Mathematical and Constraint Programming. Setting in a proper way parameters of the software, a user may apply hybrid methods on mixed data. The graphical interface also provides tools to display and analysis the effect of the settings. Therefore, this automatic software may also be used as a simulator to measure the quality of different techniques on a given microdata. The software is named TEIDE ("Técnicas para la Edición e Imputación de Datos Estadísticos") and it has been successfully used in different survey microdata in some statistical agencies in Spain. An English version of our work can be download from <http://www.goma.ull.es/TEIDE>. Our aim is to continue improving ours software with the experience of using it in other statistical offices around the world."

Keywords: Editing; Imputation; Error localization problem

1 Introduction

Due to the importance of the Data Editing and Imputation (DEI) in all statistical agencies, several different approaches have been implemented (see K-Base [2] for a list of examples). Unfortunately, each one has been developed to solve the problem of some special datasets in a very specific statistical agency. As a consequence, the implemented procedures cannot be easily used on other surveys which makes very difficult to compare their performances, and in most

cases they are not available outside the owner agency. As far as we know, only the Italian software DIESIS and the Dutch software WAID have been partially described in scientific journals (see Bruni [3, 4] and de Waal [5]). DIESIS comes from Data Imputation Editing System – Italian Software, and it solves the error localization problem through a mixed integer programming model, while the imputation problem is solved through a hot-deck donor scheme. WAID comes from Weighted Automatic Interaction Detection and it uses a tree-based model that classifies the data in terms of the values of categorical predictor variables.

This paper presents a new automatic procedure to perform DEI. It is called TEIDE, which is the acronym of the Spanish words “Técnicas de Edición e Imputación de Datos Estadísticos” (Techniques of Editing and Imputation of Statistical Data). TEIDE is intended for the reading and writing of data in open-format files. These files are simple for different statistical agencies to process and therefore to help comparisons with other implementations.

TEIDE works on data containing both qualitative and quantitative numbers, with no special assumption on the edits. Section 2 describes the main ideas under this software. Section 3 illustrates its execution on three real-world surveys provided by the Statistical Office of the Canary Islands (ISTAC) under a confidentiality agreement.

TEIDE is a free software that we are developing at University of La Laguna. The executable can be download from <http://www.goma.ull.es/TEIDE> and it runs as an stand-alone program on a computer running Windows XP. It contains an on-line help, but additionally a manual can also be download from the same webpage. Suggestions based on experiments on your own surveys are very welcome to improve the current release. Our aim is to make available also non-confidential surveys which could be used as benchmark instances to compare different softwares, and for that reason we appreciate your contribution to it.

2 Description of the tool

TEIDE was designed to be a stand-alone executable on a standard personal computer running a Microsoft® Windows operating system. To this end, the whole program has been implemented in C++ language and it has been linked so that the executable is a self-contained binary file.

Following a classical oriented object style, our approach consists of different modules that process data and send them to the next phase, i.e. a pipelining program. We now describe the two fundamental modules.

2.1 Validation

The set of edits is divided into three groups, according to their meaning.

Group 1 contains the consistency rules referencing the valid set of values associated with each field in a record. These rules are called *range edits*, and there is one for each variable in the dataset. It allows four different types according to the nature of the variable:

- *alphanumeric*: it means that the variable is a character string;
- *continuous*: it means that the variable is a fractional number in a given interval;
- *list of discrete values*: it can be, for example, a yes/no answer;
- *range of discrete values*: it can be, for example, the age of the respondent.

The user has the option of classifying each variable as *non-modifiable*, which means that their values will not be changed, even if they are required to guarantee validity. The alphanumeric variables are non-modifiable in all cases. The continuous variables correspond to quantitative fields, while the discrete variables correspond to categorical fields.

Each variable may also assume a special value ϵ outside the range which in practice is used for the case where this variable has no meaning. For example, if a respondent has no job, then the variable “salary” may be required to contain ϵ .

Group 2 of edits also contains a rule for each field, but now it is related to other fields. Let us be more precise. Given a field i , a *filter edit* associated with i has the following type:

```
IF <condition> THEN  $i \neq \epsilon$ 
IF not <condition> THEN  $i = \epsilon$ 
```

where the string $<\text{condition}>$ is a logical clause composed of other variables. This rule determines when a variable can assume the missing value. These rules are common in practice, and that is the reason for them being located in a separate group.

Group 3 contains all the remaining edits required by the statistical agency to check the validity of a record. They are called *general edits*, and they have the form

```
IF <condition> THEN <consequence>
```

The strings $<\text{condition}>$ and $<\text{consequence}>$ are logical clauses composed of categorical or numerical variables.

Although we could work with the whole set of all edits including the three groups, in practice we have found this group partition for displaying purposes very convenient. Indeed, in a real-case study, all variables have range edits, some variables have filter edits, and in exceptional cases there are general edits.

If the dataset is successfully loaded, TEIDE creates three windows. A first window contains the variable descriptions (including the correspondent range and filter edits), a second one shows the individual values in each record and field (i.e., the full dataset), and a third one displays the general edits.

Then the user can proceed checking the edits. A first phase checks the range and filter edits for each variable. A new window displays the results, as illustrated in figure 1. It contains a grid in the top half. Each row of the grid represents a record, and each column represents a field. A cell in this

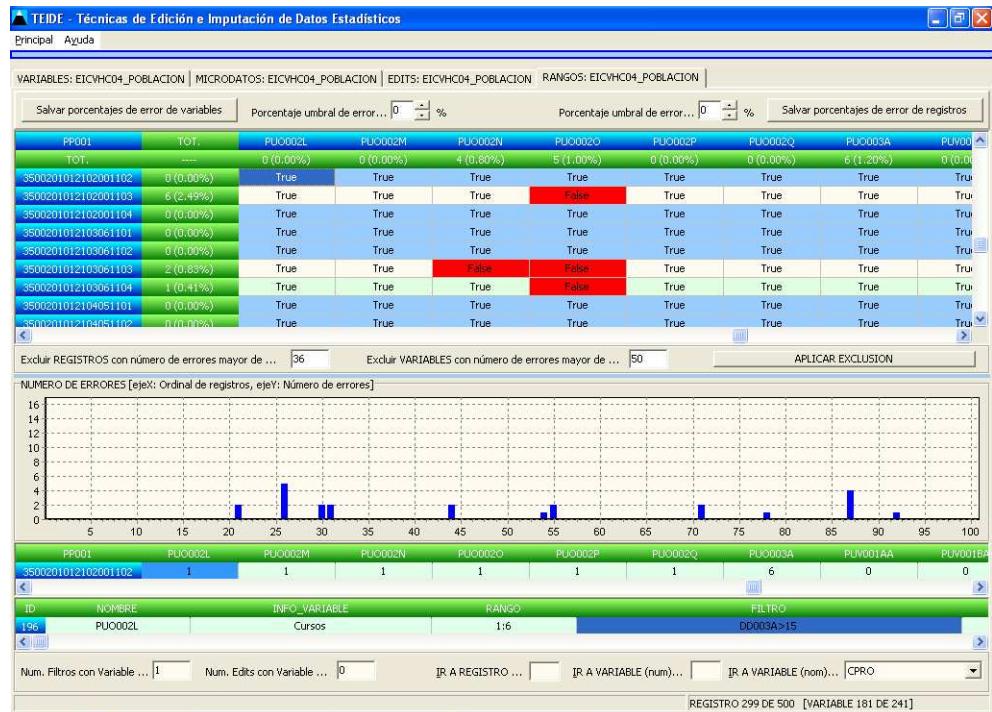


Figure 1: Evaluation of ranges and filters.

grid contains `true` when both range and filter edits hold, and `false` otherwise. Records with errors in range and filter edits are represented by blue rows, while the others by white and light green rows alternately. Cells representing wrong (range or filter) edits are in red. By moving the cursor in this grid, the bottom of the window displays the full record values and the description of the field. When a grid cell contains `false` then this information allows one to know if this is due to a range violation and/or to a filter violation. In the latter case, the other values in the wrong filters are represented in yellow. Between the grid and the bottom of the window, there is a histogram showing the number (and percentage) of variables with a wrong range or filter edit, for each record. Alternatively, it is possible to replace this histogram by another one showing the number of records with a wrong range or filter edit, for each field variable. The user can save a report with this descriptive information. Finally, the window also allows the manual modification of the original dataset values and the checking of the new dataset for filter and range edit violations. There are additional features simplifying the browsing through the grid to analyse failed rules, but we do not include these technical details in this article.

A new phase allows the user to proceed by checking the general edits. The result is displayed in a new window. It has a similar layout to the previous one, but now each column represents a general edit instead of a field variable, and each cell value is the result of the general edit evaluation instead of the result of the range and filter rule evaluation.

The decisions for designing this graphical interface have been taken based on our experiments working with practitioners in a statistical agency.

2.2 Editing and Imputation

When all the edits from the three groups have been evaluated, the set of all the records is partitioned into two subsets. A first subset \mathcal{D} contains all the records satisfying all the edits (i.e. the valid records) and is called *donor-record pool*. A second subset \mathcal{I} contains the records violating some edits, and therefore the ones needing further manipulation. An input parameter defined by the user reduces this set of invalid records by removing all those with a large number of violated edits. Each remaining invalid record creates the so-called *editing-and-imputation problem*, where we need to localize the potential field errors, and replace the original values in these fields by new ones. More precisely, the problem is to find the minimum set S^* of variables and a value y_i for each $i \in S^*$ such that the record

$$y = \begin{cases} y_i & \text{if } i \in S^* \\ a_i & \text{otherwise} \end{cases}$$

is valid and $|S^*|$ is minimal. Although one can question this definition, it is widely accepted by experts of most of the statistical offices since the article by Fellegi and Holt (see [1]).

Trivially, this definition can be extended to a more general version where each variable i in S^* has an associated weight w_i representing the probability

that this variable contains a true value a_i in a valid record. Then, the DEI problem can be seen as the Combinatorial Optimization problem of finding a set S^* which minimizes $\sum_{i \in S^*} w_i$ and such that it is possible to modify only the field variables in S^* to correct the invalid record.

To be more precise, let us represent the collection of all variables by V , and let us denote by \mathcal{R} the set of all valid records according to a given set of edits. Note that $\mathcal{D} \subseteq \mathcal{R}$ and $\mathcal{R} \setminus \mathcal{D}$ is the set of valid records not in the input dataset. Depending on the variables and the edits, \mathcal{R} can be a discrete or continuous region inside a $|V|$ -dimensional space. For example, when all variables are real numbers and all edits are linear inequations, \mathcal{R} is a polyhedron. However, when categorical variables are in the dataset, the set \mathcal{R} can be very complicated. Given a subset $S \subseteq V$ and an invalid record $a \in \mathcal{I}$, let $\mathcal{R}(S, a)$ be the valid records which coincide with a in the field variables in $V \setminus S$, i.e.

$$\mathcal{R}(S, a) := \{y \in \mathcal{R} : y_i = a_i \text{ for all } i \notin S\}.$$

Traditionally the DEI problem is split in two subproblems. The first one is named *editing problem* (or *error localization problem*) and consists of

$$\begin{aligned} & \text{Minimize} && \sum_{i \in S} w_i \\ & \text{subject to} && S \subseteq V : \mathcal{R}(S, a) \neq \emptyset. \end{aligned}$$

The second subproblem is named *imputation problem* and consists of selecting an element y in $\mathcal{R}(S^*, a)$ to replace the invalid record a in the dataset, where S^* is the solution of the first subproblem.

The imputation problem is usually solved through simple statistical tools to guarantee an unbiased dataset after the corrections. The editing problem, however, is a very difficult optimization problem, classified in the Complexity Theory as *NP-hard in the strong sense*. This is because it is a generalization of the known “Minimum Weighted Solution of Linear Equations”, arising when all the variables are real numbers and the edits are linear equations where the vector zero is not a solution. As a consequence, it is very unlikely that an algorithm running in polynomial time on the size of the input data can be proposed. However, it is still possible to derive exact approaches to solve some specific instances, as well as heuristic techniques to address complicated instances. The purpose of this article is to propose a new approach working on real-world datasets. The new technique takes into account the whole DEI problem in the same spirit as the so-called *hot-deck donor schemes* (see [6]). Therefore, the here-proposed technique also solves both the first and the second subproblems.

To approach the DEI problem we have considered a heuristic generating a near-optimal solution in a short computational time. Details on this heuristic algorithm are not given here due to the 10-page limit of this paper.

3 Cases studies

We were unable to get the same benchmark instances used by Bruni [3, 4] but they were not available due to confidentiality issues by the Italian statistical agency. For this reason, we could not compare the performance of our algorithm with other previous approaches.

We have tested the performance of the heuristic approach on three real-world surveys provided by the Statistical Office of the Canary Islands (ISTAC). All the experiments have been conducted on a PC Pentium IV 2.4 Ghz. running Windows XP and with one Gbyte of RAM. A first survey is the *2004 Canary Health Survey* (“Encuesta de Salud Canaria 2004”), and it will here be referred to as HEALTH. A second survey contains the house data from the *Canary Household Condition Survey 2005* (“Encuesta de Ingresos y Condiciones de Vida de los Hogares Canarios 2005”), and it will be referred to as HOUSE. The third survey contains the individual information from the *Canary Household Condition Survey 2005*, and it will be referred to as INDIVIDUAL. Table 1 shows some features of these microdata and also some details of the performance of our implementation.

More precisely, the rows of Table 1 are grouped into four blocks. The rows in each block give the following information:

- *Description:* It refers to the following features of the input dataset:
 - *#variables:* number of variables in the dataset, including the “non-modifiable”.
 - *#records:* number of records in the dataset.
 - *#microdata:* number of values including the “no-answer” (i.e. it is simply *#variables* times *#records*).
 - *#real microdata:* number of values excluding the “no-answer”.
 - *#continuous variables:* number of quantitative fields defined by real or fractional numbers.
 - *#discrete variables:* number of qualitative fields, or fields defined by integer numbers.
 - *#imputable variables:* number of fields whose values can be modified.
 - *#filter edits:* number of fields with a filter edit.
 - *#general edits:* number of general edits.
 - *Loading–Checking time:* time to load and check all relationships and coherence within data, variables and edits.
- *Evaluation I:* It concerns the validation of the range and filter edits. The following information is displayed:
 - *Total time:* total evaluation time (in seconds).
 - *#correct records:* records satisfying all the range and filter edits.

Block	Information	Survey		
		HEALTH	HOUSE	INDIVIDUAL
Description	#variables	375	176	234
	#records	5633	7797	22584
	#microdata	2112375	1372272	5284656
	#real microdata	1129379	1108119	2035238
	#continuous variables	2	12	16
	#discrete variables	373	164	218
	#imputable variables	335	128	210
	#filter edits	283	42	207
	#general edits	33	7	45
	Loading–Checking time	105.29	69.43	251.79
Validation I	Total time	18.48	8.31	50.88
	#correct records	4215	4934	12374
	#incorrect records	1418	2863	10210
	#correct variables	137	20	29
	#incorrect variables	238	156	205
Validation II	Total time	6.70	5.28	23.31
	#correct records	5519	7635	21613
	#incorrect records	114	162	971
	#correct edits	17	1	6
	#incorrect edits	16	6	39
Imputation	Total time	3835.47	258.14	5573.30
	#donor records	4138	5049	11915
	#correctable records	1495	2748	10669
	#non-corrected records	5	0	0
	#corrected records	1490	2748	10669
	#warning records	30	19	49
	affected var. average	7.92	7.16	12.08
	imputation var. average	5.22	6.44	11.13
	wrong-range var. average	1.47	5.93	9.75
	wrong-edits var. average	6.73	1.69	4.34
<i>connected var. average</i>		225.13	5.50	195.63
<i>donor record distance average</i>		0.98	1.10	1.35
Overall time		3965.94	341.16	5899.28

Table 1: Computational Results of using TEIDE on three real-world surveys.

- *#incorrect records*: records violating at least one edit.
 - *#correct variables*: variables whose range and filter edits are satisfied in all records.
 - *#incorrect variables*: variables whose range and/or filter edit is violated in at least one record.
- *Evaluation II*: It refers to the evaluation of the general edits. The following information is displayed:
 - *Total time*: total evaluation time (in seconds).
 - *#correct records*: records satisfying all the general edits.
 - *#incorrect records*: records violating at least one edit.
 - *#correct edits*: edits satisfied by all the records.
 - *#incorrect edits*: edits violated by at least one record.
- *Imputation*: It refers to the process of finding the new values to correct the errors, as follows:
 - *Total time*: total imputation time (in seconds).
 - *#donor records*: records satisfying all the edits.
 - *#correctable records*: records not in the donor pool, and therefore invalid.
 - *#non-corrected records*: records not automatically corrected.
 - *#corrected records*: records satisfactorily modified.
 - *#warning records*: corrected records with more modified values than variables involved in invalid edits.
 - *affected var. average*: average number of variables involved in all edits violated by an incorrect record. More precisely, this number is the number $|S(a)|$ used by the heuristic approach (not described). The average has been computed over the corrected records.
 - *imputation var. average*: average number of modified values. The average has been computed over the corrected records.
 - *wrong-range var. average*: average number of values with wrong range edit. The average has been computed over the corrected records.
 - *wrong-edits var. average*: average number of variables involved in all edits violated by an incorrect record excluding the range edits. The average has been computed over the corrected records.
 - *connected var. average*: average number of variables related to variables in violated edits. This number is the number $|U(a)|$ used by the heuristic (not described). The average has been computed over the corrected records.

- *donor record distance average*: distance average from each correctable record to its associated donor record, i.e. the average of function $d(a, b)$ used by the heuristic (not described). The average has been computed over the corrected records.

From the table it can be observed that TEIDE is able to deal with the three dataset in a very reasonable computational time, see the last row (“Overall Time”) of Table 1, and with a quite satisfactory quality result. Note that the three surveys contain both continuous and discrete variables, with a large number of variables and records. (INDIVIDUAL is the largest survey ever managed by ISTAC.)

Acknowledgments

This research was supported by “Instituto Canario de Estadística” (ISTAC, Canary Islands, Spain).

References

- [1] Fellegi, I.P., Holt, D., 1976, A systematic approach to automatic edit and imputation, *Journal of the American Statistical Association*, **71**, 17–35.
- [2] The Knowledge Base on Statistical Data Editing. Available online at: <http://amrads.jrc.cec.eu.int/k-base> (accessed 25 September 2005).
- [3] Bruni, R., 2004, Discrete models for data imputation. *Discrete Applied Mathematics*, **144**, 59–69.
- [4] Bruni, R., 2005, Error Correction for Massive Data Sets. *Optimization Methods and Software*, **20**, 295–314.
- [5] de Waal, T., 2001, WAID 4.1: a computer program for imputation of missing values. *Research in Official Statistics*, **2**, 53–70.
- [6] Ford, B.F., 1983, An overview of hot-deck procedures. *Incomplete data in sample surveys: Theory and bibliographies*, **2**, 185–207.