WP. 9
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Neuchâtel, Switzerland, 5-7 October 2009)

Topic (i): Automated editing and imputation and software applications

TEIDE 2: A modern tool for Editing and Imputation of Statistical Data

Supporting Paper

Prepared by Juan José Salazar González and Mª Salomé Hernández García, University of La Laguna, Tenerife, Spain

I. INTRODUCTION

1. Techniques for editing and imputation of data are used to detect and correct "errors" or "lost data" contained in surveys. This is a complex process that needs sophisticated and effective procedures to guarantee good solution quality. Therefore it is essential to take advantage of modern technology, including computational tools (computers, programming languages, databases, GUI, etc.).

2. The software TEIDE ("Técnicas de Edición e Imputación de Datos Estadísticos", which means "Techniques for Editing and Imputation of Statistical Data") was born in order to make faster, reliable and more comfortable several tasks related to Statistical Data Editing. The software was designed and implemented by a research team at University of La Laguna (Tenerife, Spain) and as a result it:

- is multi-platform (32-bits and 64 bits running under Linux, Windows, Mac,...).
- Makes use only of free software (C++, Qt library).
- Reads and writes in standard formats including XML, Microsoft Access and ORACLE.

3.Since an early version TEIDE1 started in 2003, the software has been considerably improved thanks to the use of it by several statistical agencies, including ISTAC (Statistical Office of Canary Islands) and IEA (Statistical Office of Andalucía). TEIDE has been used to debug and clean different real-world surveys, and has learned a lot from these experimentations. The current version is TEIDE2. It works on surveys with both quantitative and qualitative variables, allowing a very wide variety of consistency rules. For the imputation TEIDE2 combines the standard donor record procedure with a regression analysis, both based on a subset of correct records selected from the whole survey. Many parameters allow a user to control and debug the whole process through a set of friendly windows.

II. DEVELOPMENT

4. TEIDE2 was used to debug and clean the "Survey on Income and Living Conditions of Households in Canary Islands", "Survey on the Implementation of Information and Communication Technologies in Households in Canary Islands", "Survey on the Implementation and Use of Technology in Information and Communication in Business in Canary Islands", "Tourist Expenditure Survey", and others.

5. TEIDE 2 is a new version, totally re-implemented after our experiences using the first version on several real-world surveys. The old version TEIDE1 was exclusive for computers running Windows and Microsoft Access, and compiled and linked using the commercial software Borland C/C++.. The new version TEIDE2 has been re-implemented in standard C++ and linked only using the free Qt library.

5.1. **C++** is a general-purpose programming language. It is regarded as a middle-level language, as it comprises a combination of both high-level and low-level language features. It was developed by Bjarne Stroustrup starting in 1979 at Bell Labs as an enhancement to the C programming language and originally named "*C with Classes*". It was renamed to *C++* in 1983.

C++ is widely used in the software industry. Some of its application domains include systems software, application software, device drivers, embedded software, high-performance server and client applications, and entertainment software such as video games. Several groups provide both free and proprietary C++ compiler software, including the GNU Project, Microsoft, Intel, Borland and others.

5.2. **Qt** is a cross-platform application development framework, widely used for the development of GUI programs (in which case it is known as a *widget toolkit*), and also used for developing non-GUI programs such as console tools and servers. Qt is most notably used in KDE, Google Earth, Skype, Qt Extended, Adobe Photoshop Album, VirtualBox and OPIE. It is produced by Nokia's Qt Software division, which came into being after Nokia's acquisition of the Norwegian company Trolltech, the original producer of Qt, on June 17, 2008.

Qt comes under the following licenses:

Qt Commercial Version	The Qt Commercial version is the appropriate version to use for the development of proprietary and/or commercial software. This version is for developers who do not want to share the source code with others or otherwise comply with the terms of the GNU Lesser General Public License version 2.1 or GNU GPL version 3.0.
Qt GNU LGPL v. 2.1	This version of Qt is appropriate for the development of Qt applications (proprietary or open source) provided you can comply with the terms and conditions contained in the GNU LGPL version 2.1.
Qt GNU GPL v. 3.0	This version of Qt is appropriate for the development of Qt applications where you wish to use such applications in combination with software subject to the terms of the GNU General Public License version 3.0 or where you are otherwise willing to comply with the terms of the GNU General Public License version 3.0.

A clear advantage of working with TEIDE2 is that it is all open and free software. Then it is the ideal framework to implement and test new methods, and compare how these new methods works respect to other previously implemented.

III. WAY OF USE

6. The application TEIDE2 reads a metafile which should contain the name of the database Oracle, Microsoft Office Access (.mdb) or in format XML. These databases contain the necessary tables:

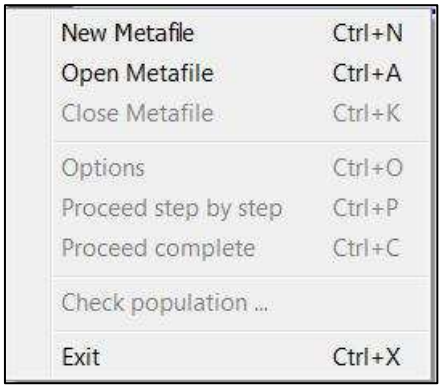
Variables: The variables that we have to study, with the necessary information (ID: identifying of the variable; NAME: name of the variable; INFO_VARIABLE: information about the meaning of the variable; TIPO: it indicates if it is discreet in range, discreet in list or constant; RANGO: values that it can take (of a list or of a range of values); FILTRO: necessary filter to indicate when the variable must be NO_PROCEDE; INFO_FILTRO: information about the meaning of the filter; SENTIDO_FILTRO: be able to take the values a, b or c “a) if (non-filter) then (valor=NP)” “b) if (filter) then (value! =NP)” “c) a and b”; IMPUTABLE: it indicates if variable debit or not to be imputed; NO_PROCEDE: it indicates if it can or not to take the value NO_PROCEDE; NO_SABE: it indicates if it can or not to take the value NO_SABE; NO_CONTESTA: it indicates if it can or not to take the value NO_CONTESTA; NS_NC: it indicates if it can or not take the value NO_SABE_O_NO_COTESTA; PESO: it indicates the weight that this variable takes with regard to other; MAPPING: the table indicates the mapping and contains the meaning of the values that this variable may take; IMP_NUM: type of imputation, by default it is the record donor).

Microdata: it contains all the values that take the variables in every record.

Edits: they are the rules that have to meet the variables. The table contains ID: identifying of the edit; CONDICION: the edit; DESCRIPCION: information about the meaning of the edit.

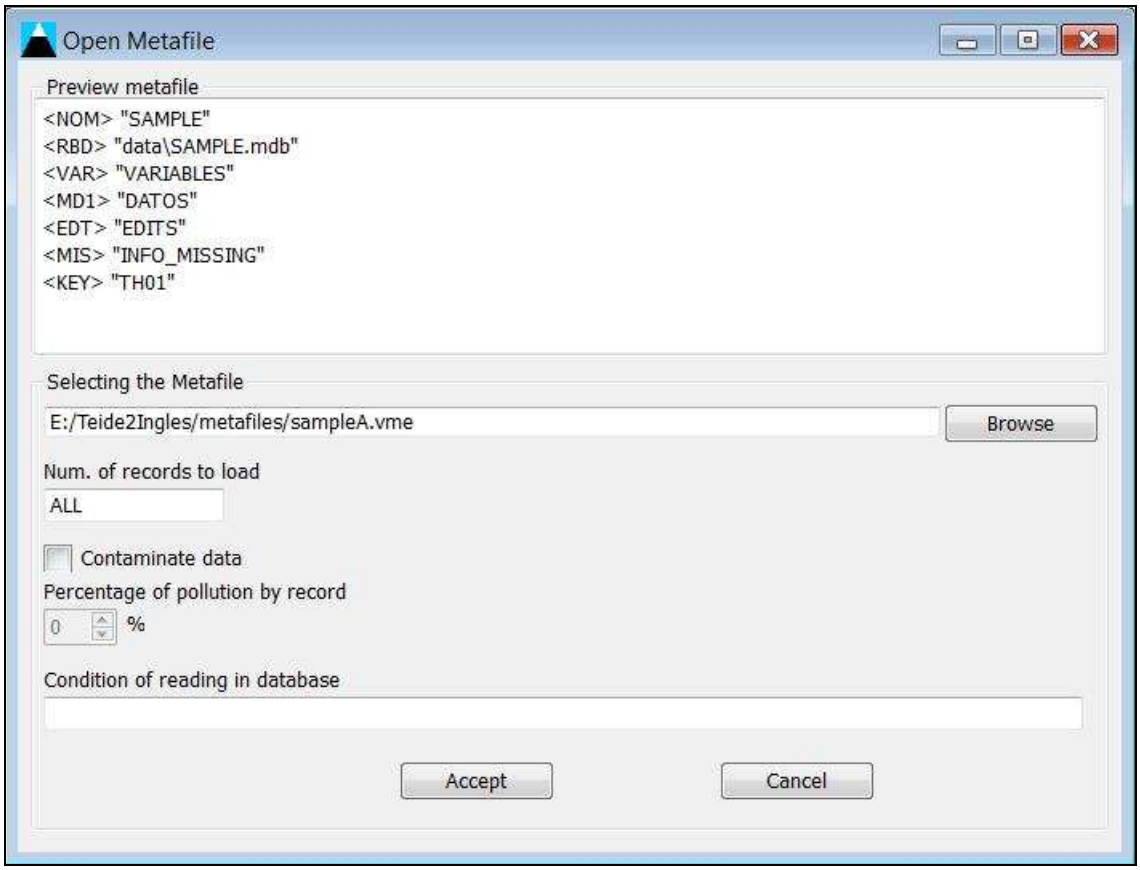
Missing: it is the table that contains the values missing (lost) that takes the variables.
Mapping: they are the tables with the meaning of every value that takes this variable.

7. As soon as the application is executed the user has to choose an option of the menu that appears later:



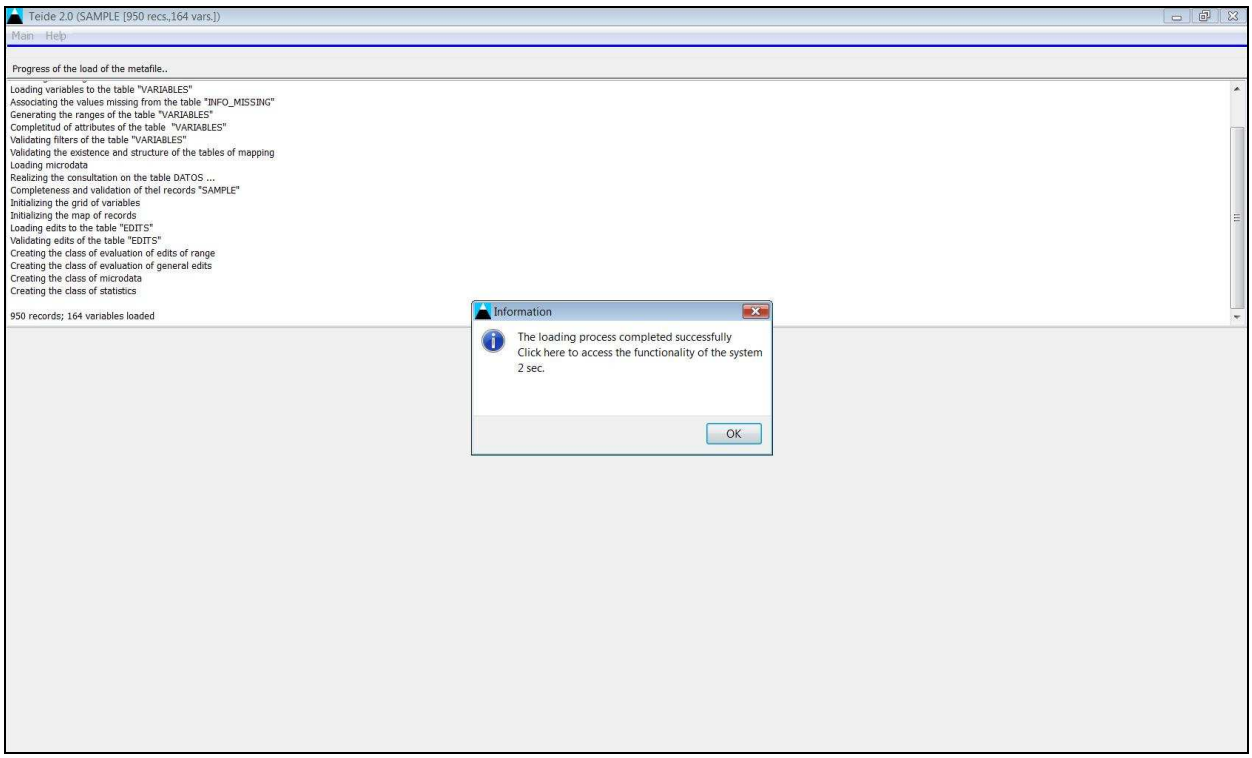
Menu.

8. The user has to open a metafile created in order to load data.

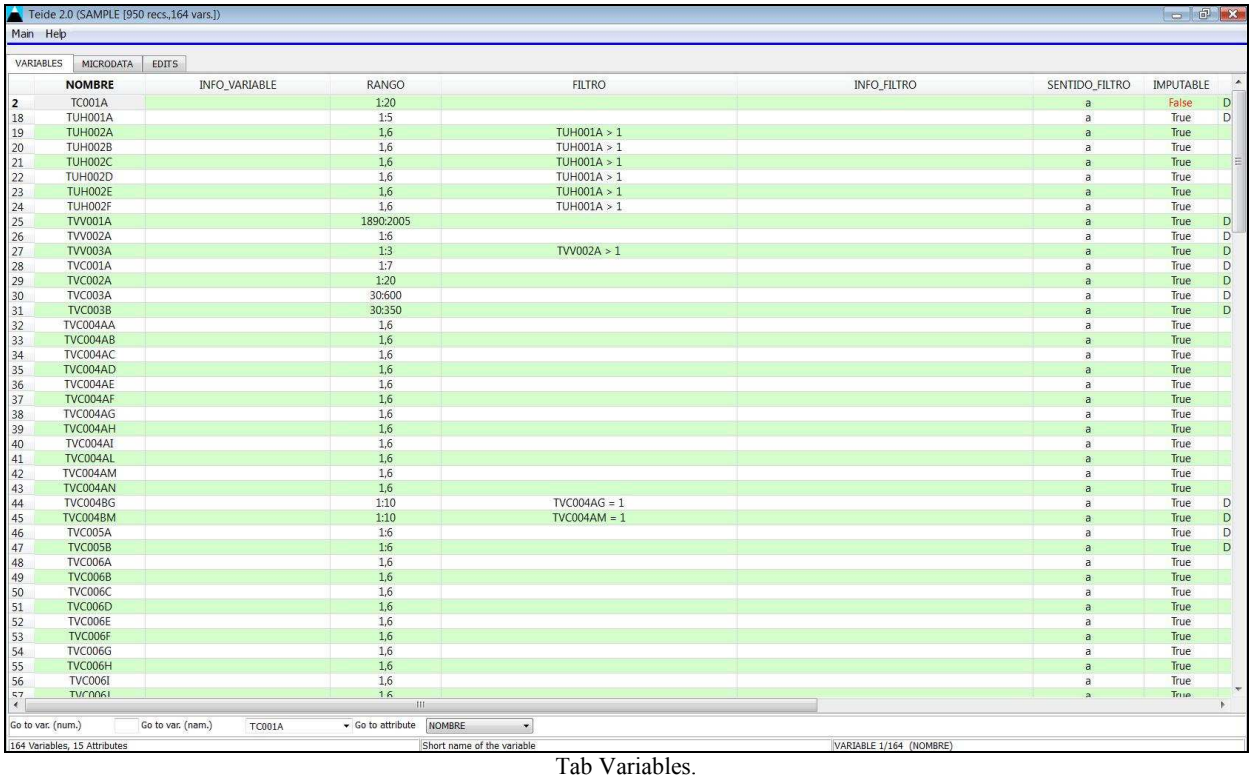


Open metafile.

9. When the metafile is selected, it begins the process of load in which the program will read all the tables and will show the information in the tab: Variables, Microdata and Edits.



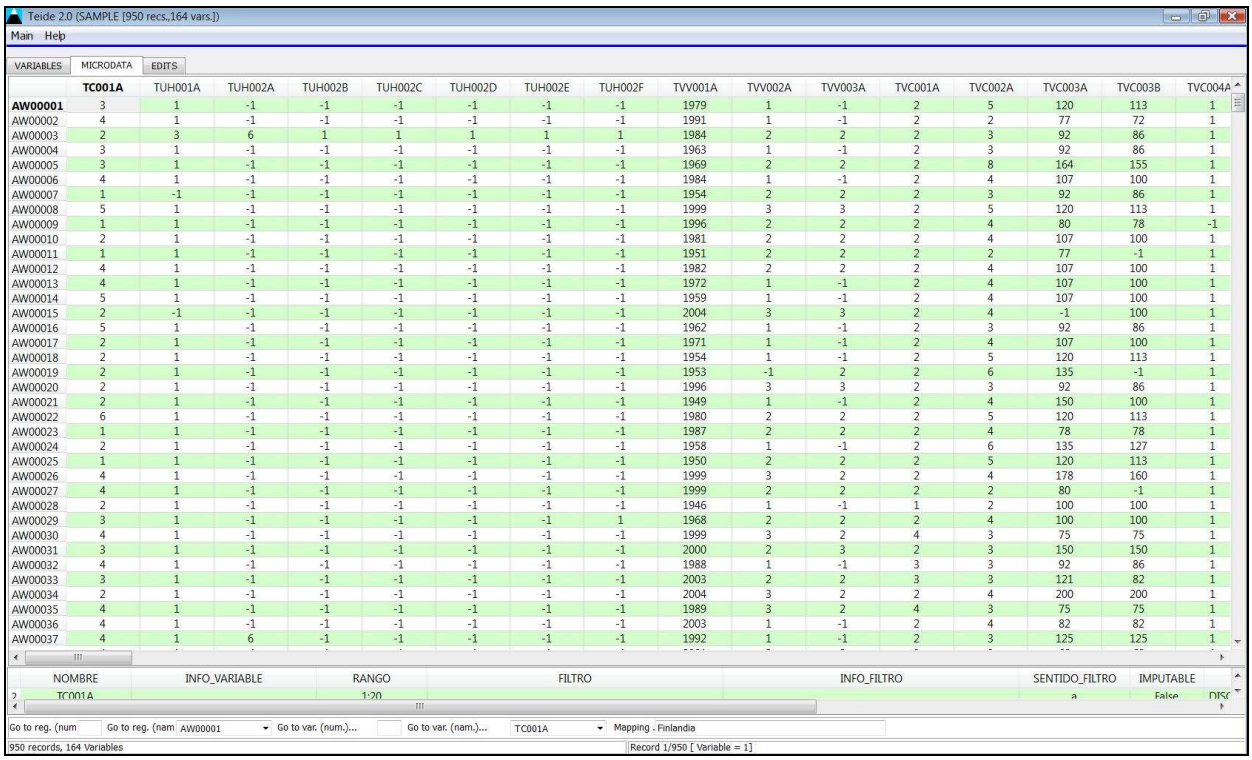
Progress of the load.



Tab Variables.

10. In the tab of variables, there appears the information of every variable: name, range, filter (in case it has it), if it is attributable, etc.

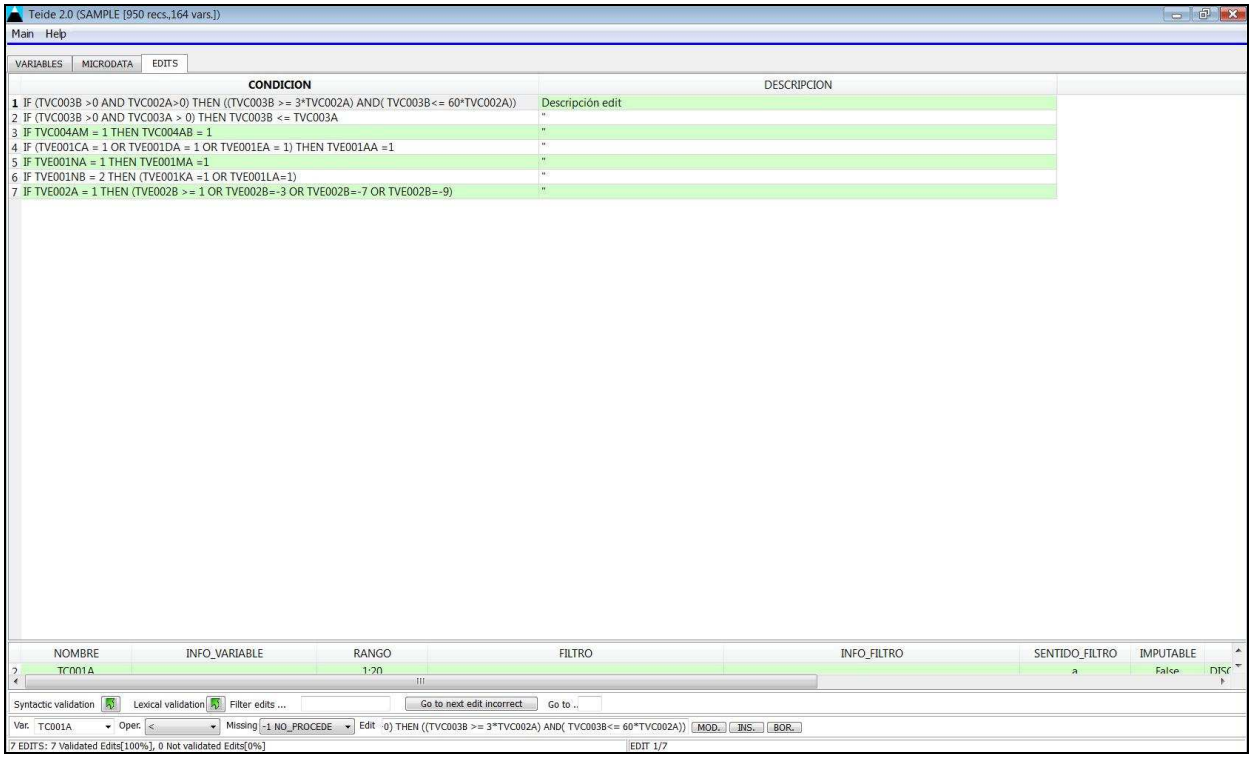
11. The tab of microdata shows all the records in the survey and the values that take the variables in each of them. Also it shows information about the selected variable.



Tab Microdata.

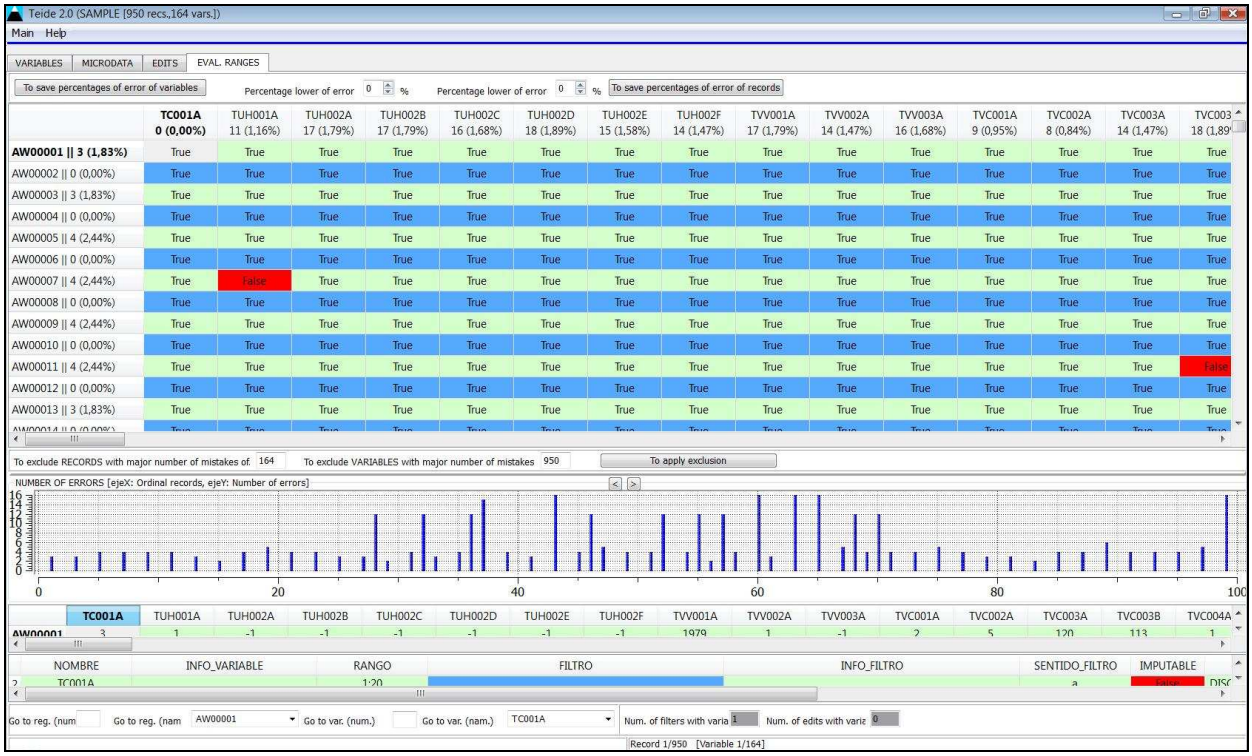
12. The tab of edits shows all the consistency rules. It indicates if they are correct or not. It allows to modify, erase or insert a new edit in an easy form.

13. Once one has seen the information showed after the process of load, the application allows that it should follow with the process of edition and imputation. It is possible to do it in one over two ways: stepwise or complete form. If the chosen option is stepwise, the user will be able to see each of the tabs that are created. If the option is complete form, the user will not be able to stop neither the process of edition nor that of imputation, and the application will show it all the tabs once it has finished the whole data-editing process.



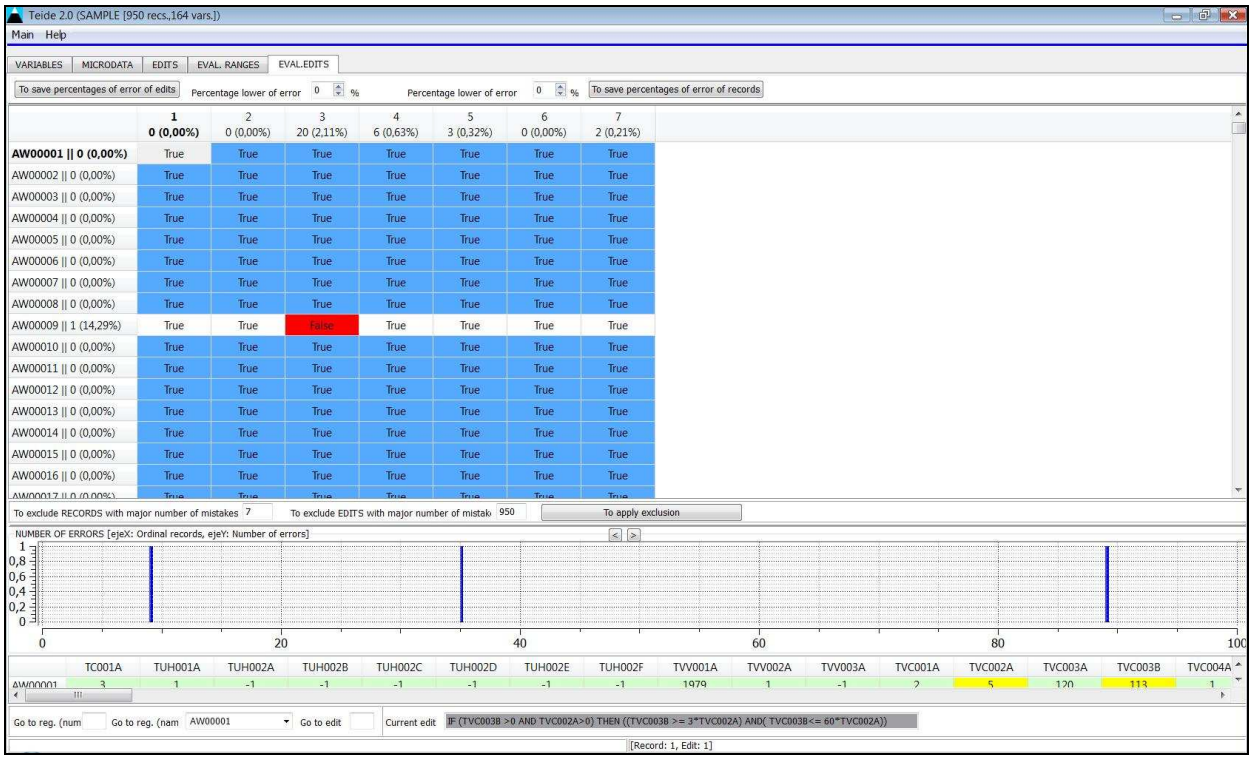
Tab Edits.

14. The following tab shows the evaluation of variable ranges, in which all the records will appear with variables in colors, indicating if the range or the filter are correct or not.



Tab Evaluation of Ranges.

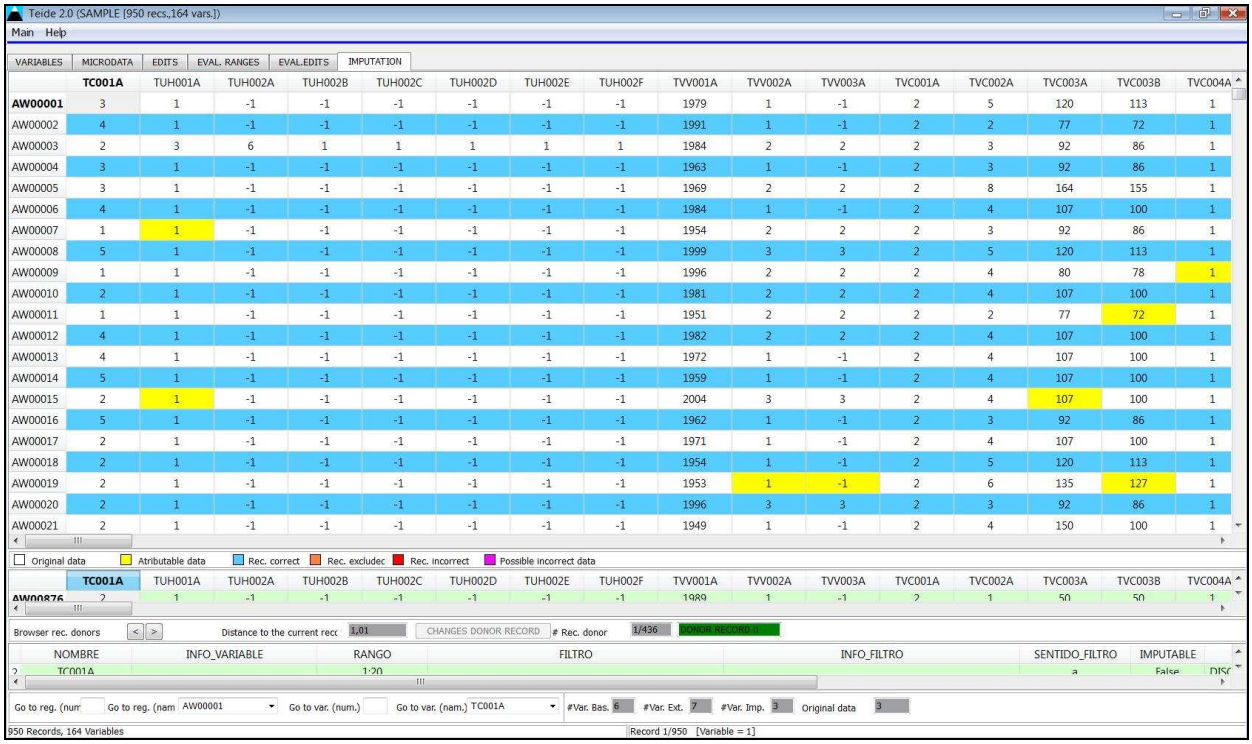
15. When the ranges and filters have been studied, the following step is the study of the edits. The tab of evaluation of edits shows what records with it variables fulfill or do not fulfill the proposed edits.



Tab Evaluation of Edits.

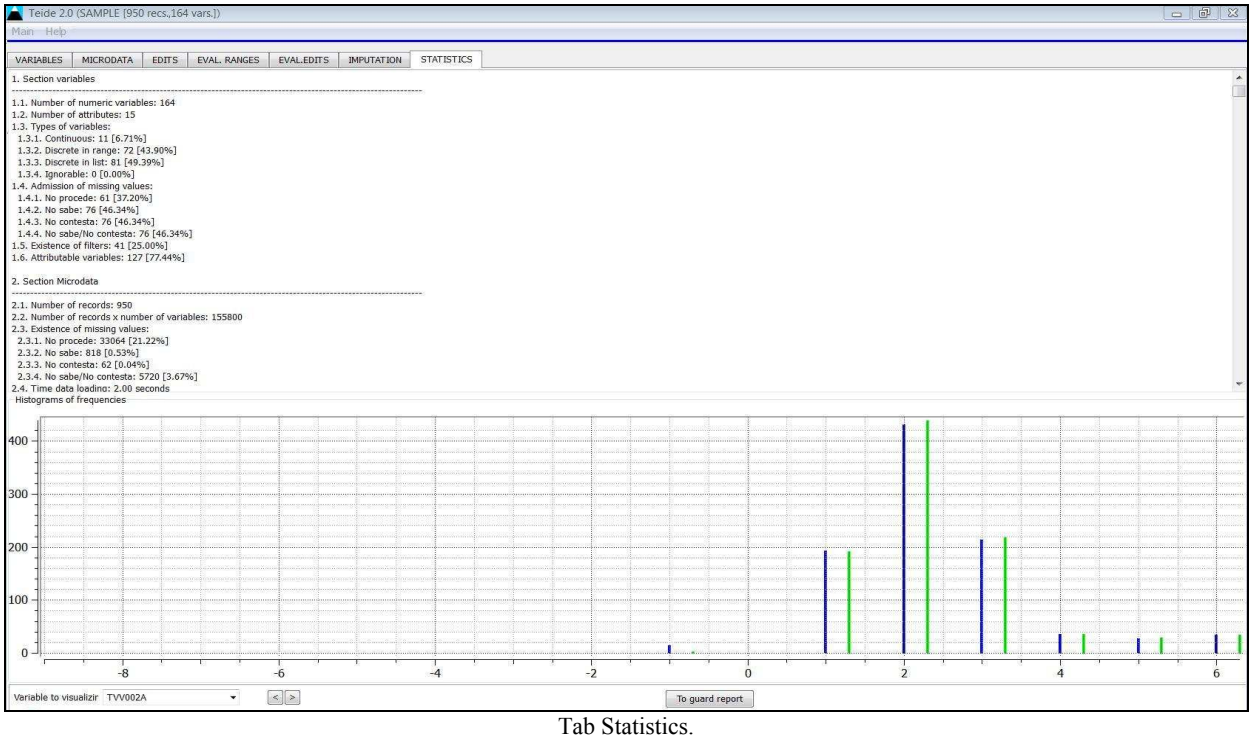
16. Once finished the processes of evaluation, the following step is that of imputation. The tab of imputation shows for all the records and it variables the following information:

- Records donors, in blue color.
- Incorrect records (they could not have corrected), in red color.
- Values of an incorrect record that TEIDE could not have corrected but that can that he is the culprit of whom this record is incorrect, in color magenta.
- Excluded records, in orange color.
- Incorrect records corrected in white color.
- Values of an incorrect record that TEIDE has modified, in yellow color.



Tab Imputation.

17. When the edition and imputation phases have finished TEIDE shows the tab of statistics (a report) which shows all the information relative to the obtained results. As well as a graph in which it is possible to see for every variable the values that had before and that have now.



Tab Statistics.

18. The application displays the results and computational statistics, i.e. the errors found and corrections made. At this stage we analyze the performance of executive tasks and data, to decide whether the results have an acceptable quality level. We analyze and evaluate the coverage obtained, the level of non-response, accuracy of responses, errors occurring during the recording, the level and distribution of errors made and the imputations.

IV. TESTS OF COMPILATION AND EXECUTION

19. To give a proper idea on how TEIDE2 works on a real-world survey, here we illustrate technical details using TEIDE2 on one survey. Due to confidentiality issue we cannot give the survey for further experiments. The database consists of 289 variables, 2789 records and 66 edits. Most of the variables are qualitative and the rest are quantitative. We conducted tests on two machines with the following characteristics:

Intel Core 2 Duo 3 GHz. Memory (RAM) 4 GB Operating Systems Ubuntu 9.04 64 bits Windows Vista Business 64 bits
Intel Core 2 Duo 3 GHz Memory (RAM) 3GB Operating Systems Ubuntu 9.04 32 bits Windows Vista Home 32 bits

As for C/C++ compilers we have used:

In Windows: Qt includes MinGW compiler (g++). We have also used Microsoft Visual Studio 2008 to obtain a different executable on the same computer. The MinGW which provides Qt is 32 bits and they do not have a 64-bit version. However, with Visual Studio we have a version of 32 and 64 bits.

In Linux we have been used g++ and icpc (Intel compiler c++). Both are 32 and 64 bits.

Executables have been compiled on two different machines, with a 32-bit OS for 32 bit executable and 64-bit OS for 64 bit executable. All tests were performed in 64-bit OS.

A. WINDOWS

20. Time of load of data

MinGW 32 bits	**Visual Studio 32 bits**	**Visual Studio 64 bits**
13 seconds	8 seconds	8 seconds

Process time in the section of ranges

MinGW 32 bits	**Visual Studio 32 bits**	**Visual Studio 64 bits**
7 seconds	5 seconds	4 seconds

Process time in the section of test

MinGW 32 bits	**Visual Studio 32 bits**	**Visual Studio 64 bits**
7 seconds	6 seconds	5 seconds

Process time in the section of imputation

MinGW 32 bits	Visual Studio 32 bits	Visual Studio 64 bits
1741 seconds	1356 seconds	1152 seconds

B. LINUX

21. Time of load of data

G++ 32 bits	ICPC 32 bits
10 seconds	10 seconds

G++ 64 bits	ICPC 64 bits
9 seconds	8 seconds

Process time in the section of ranges

G++ 32 bits	ICPC 32 bits
10 seconds	10 seconds

G++ 64 bits	ICPC 64 bits
22 seconds	22 seconds

Process time in the section of test

G++ 32 bits	ICPC 32 bits
8 seconds	8 seconds

G++ 64 bits	ICPC 64 bits
20 seconds	20 seconds

Process time in the section of imputation

G++ 32 bits	ICPC 32 bits
1251 seconds	1165 seconds

G++ 64 bits	ICPC 64 bits
1053 seconds	1026 seconds

C. COMMENTS

22. According to this in Windows it seems that the best version is compiled in Visual Studio 2008 to 64 bits, resulting in the imputation phase a difference of nearly 600 seconds with respect to MinGW compiled to 32-bit.

Linux is the best compiled with icpc to 64 bits.

V. RESULTS OF THE IMPUTATION

23. The database is read in XML as it is the only format that currently supports all versions. The results were as follows:

Donor records	2023
Records to correct	766
Correct records	671
Incorrect records	95

Records warning	33

Average imputed variables per record (total)	1.94
Average imputed variables per record (no reg. warning)	1.56
Average of errors in range per record	0.00
Average of variables involved in edits incorrect for record	3.18
Average of variables involved in total mistakes for record	3.18
Average of variables involved in comp. connected with mistake for record	37.51
Average distance to donor records	0.06

Over the total (2789 records) 72.53% are correct records, thus 27.47% records have been modified.

VI. CONCLUSIONS AND FUTURE

24. TEIDE2 can impute great variety of surveys, with quantitative and qualitative variables. Since the beginning of the data collection, TEIDE2 is a nice too to start checking that the collecting is being doing as expected, or to detect any kind of unexpected misunderstanding on the query or other type of fundamental errors. The whole editing and imputation process is carried out step by step so that the statistician can see the results of each study or modification. It is also possible to make a complaint about a given fact. There are different parameters to drive the imputation process as desired. TEIDE2 will never remove the human in the data editing process, but will help him/her to make the best of his/her time. TEIDE2 is also a great simulator to measure the impact of different consistent rules on a data survey.

Any application can be improved and in this case the future tasks we intend to address are:

- Reduce processing times. This will require thinking about improvements in the algorithms of evaluation of expressions and data structures.
- Improve gradually the imputation procedures that are being used.
- Exploit the parallelization that it is now available in my computers to make our data-editing algorithm faster.
- Add new formats to better integrate TEIDE2 with other software.
- Test and compare TEIDE2 with other software on the same data surveys.

25. We invite statistical agencies to download and check our software TEIDE2 from our webserver www.goma.ull.es, and to help us to improve it with new experiments on different surveys. Since TEIDE2 is a self-contained software, fully implemented inside a research team in a public university, it is the best framework to also implement and test different imputation techniques. Therefore, please, fill free to suggest the authors all type of suggestions that would make TEIDE2 more useful for the data editing processes in your statistical agency. Contact jjsalaza@ull.es for further details.

26. As a consequence of the economical crisis in all Europe, we are currently finding troubles to find economical support to continue improving TEIDE2. We would be gratefully to join any kind of proposal to collaborate in any National or European funded project or contract.

27. In summary, our objective is to show and share with you an application (TEIDE2) which is now faster, more flexible and more reliable to automate as far as possible the process of purification of a survey. In this way, the statistical institute can devote their resources to explore more deeply sensitive and more specific tasks.