TEIDE: A new software for data editing

Juan José Salazar González
Universidad de La Laguna
Tenerife - Spain

Work partially funded by
"Instituto Canario de Estadística" & "Instituto Andaluz de Estadística"

UNECE, Bonn, September 26, 2006
Content

1. Introduction
   - Preface
   - Previous experience
   - What is TEIDE?
   - Implemented methodologies
   - What is TEIDE 1?
   - What is TEIDE 2?
   - A few details

2. On-line demonstration
   - Statistical agencies using TEIDE
   - Computational experiments

3. Conclusions
1 Introduction
   - Preface
   - Previous experience
   - What is TEIDE?
   - Implemented methodologies
   - What is TEIDE 1?
   - What is TEIDE 2?
   - A few details

2 On-line demonstration
   - Statistical agencies using TEIDE
   - Computational experiments

3 Conclusions

Juan-José Salazar-González  TEIDE
Aiming to replace humans by an automatic guided software is impossible when editing statistical data.

However, using hardware and software easily available today, we can reduce part of the effort, specially when the same operations (checks, computations, ...) must be done on hundreds or thousands of records.

Then, the saved effort can be devoted to improve the more complex and challenging parts of the survey.
As noticed by Tom de Waal ("Processing of Erroneous and Unsafe Data", 2003), there is a narrow relationship between Statistical Disclosure Control and Statistical Data Editing:

- Finding suppressions in SDC is like finding errors in SDE.

- Finding the expected attacker value of a suppression is like finding the imputation of an error.

University of La Laguna (Tenerife) was involved in SDC:

- European project, SDC 1996-1998

- European project, CASC 2001-2003

- U.K. project funded by ONS, 2004
What is TEIDE?

TEIDE is the name of the highest mountain in Spain, which is a vulcan in Tenerife with 3718 meters.
What is TEIDE?

In our context, it is a software for editing and imputation:

1. Técnicas de Edición e Imputación de Datos Estadísticos comes from "Técnicas de Edición e Imputación de Datos Estadísticos", which means "Techniques for Editing and Imputation of Statistical Data".

2. It works on continuous, categorical, and mixed data.
   - Discrete
     - list \((x_i \in \{v_{i1}, v_{i2}, \ldots, v_{il_i}\})\)
     - range \((x_i \in \{v_{i\min}, \ldots, v_{i\max}\})\)
   - Continuous \((x_i \in [v_{i\min}, v_{i\max}])\)

3. The edits for categorical data are logical constraints, like:
   - Arithmetic \((x_1 = 2 \times x_4 + x_2)\)
   - Logical \((if \ (x_1 = a) \ then \ (x_3 \leq b))\)
   - Mixed \((if \ (x_4 \times x_2 = 6) \ then \ (x_1 = 1))\)
Assumptions in TEIDE

1. The edits for numerical data must be linear equalities or inequalities.
2. On categorical data, unfeasible edits are automatically detected. On numerical data, no yet!
3. Both on categorical data and numerical data, the syntax of all edits are checked. Invalid edits are detected, so the user can redefine them. Redundant edits are not checked (no yet!)
4. Data are automatically checked according to the edits, but data cannot be modified by the users inside TEIDE.
Methodologies under TEIDE

1. The syntax of all edits are checked.
2. Validation: Each record is validated on each edit, individually.
3. Error localization: the minimum number of modifications is computed through optimal or near-optimal approaches – Fellegi y Holt (JASA 1976)
4. Imputation: Donor and regression are optional to all (categorical or numerical) variables
How "TEIDE 1" is being implemented?

- It is written in C++ Programming Language
- Compiled and linked in Borland C++ Builder
- Running under Microsoft Windows
- It is a stand-alone executable, so it can be copied into a new computer and executed.
- Read and Write in Microsoft Access database.
- It is free software.
- Download a copy from [http://www.goma.ull.es](http://www.goma.ull.es)
How "TEIDE 2" is being implemented?

- It is written in C++ Programming Language
- It is open source
- It can be compiled on any C++ compiler, and therefore, it runs on any computer under any operating system (including Windows, Linux, Mac, Sun, HP, ...)
- no virtual machine is required (no JAVA, no .NET, ...)
- Read and Write in XML format.
- For the moment we have a beta version in progress.

Finishing this version depends on the economical support.
How are displayed the edits

Edits are located in three groups:

1. type: checking absolute values
2. filters: checking values relative to other values
3. general: standard edits in general form
How the process works

1. The input data are loaded
2. The edits are evaluated
3. Registers are classified into two groups: correct and incorrect
4. Individual data of an invalid record are classified in different status: wrong values, values in wrong edits, values in edits with variables in wrong edits,...
5. Different weights are assigned to individual values depending on the status, the importance of the values, etc.
6. Weights will be used to define the distance between two records (typically between an incorrect and a correct record) when using the donor technique, and to be used in a potential regression analysis.
Flowchart
1 Introduction
   Preface
   Previous experience
   What is TEIDE?
   Implemented methodologies
   What is TEIDE 1?
   What is TEIDE 2?
   A few details

2 On-line demonstration
   Statistical agencies using TEIDE
   Computational experiments

3 Conclusions
Real-world surveys:

- TEIDE was developed and improved by working on real-world surveys at “Instituto Canario de Estadística” (ISTAC) and at “Instituto Andaluz de Estadística” (IAE).
  - IAE: “Encuesta Mundial de Valores”, “Encuesta Social”

We thanks these organizations for supporting TEIDE
### Some benchmarks:

<table>
<thead>
<tr>
<th>Block</th>
<th>Information</th>
<th>Survey</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>HEALTH</td>
<td>HOUSE</td>
</tr>
<tr>
<td>Block</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>#variables</td>
<td>375</td>
<td>176</td>
</tr>
<tr>
<td></td>
<td>#records</td>
<td>5633</td>
<td>7797</td>
</tr>
<tr>
<td></td>
<td>#microdata</td>
<td>2112375</td>
<td>1372272</td>
</tr>
<tr>
<td></td>
<td>#real microdata</td>
<td>1129379</td>
<td>1108119</td>
</tr>
<tr>
<td></td>
<td>#continuous variables</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>#discrete variables</td>
<td>373</td>
<td>164</td>
</tr>
<tr>
<td></td>
<td>#imputable variables</td>
<td>335</td>
<td>128</td>
</tr>
<tr>
<td></td>
<td>#filter edits</td>
<td>283</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>#general edits</td>
<td>33</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Loading–Checking time</td>
<td>105.29</td>
<td>69.43</td>
</tr>
<tr>
<td>Imputation</td>
<td>Total time</td>
<td>3835.47</td>
<td>258.14</td>
</tr>
<tr>
<td></td>
<td>#donor records</td>
<td>4138</td>
<td>5049</td>
</tr>
<tr>
<td></td>
<td>#correctable records</td>
<td>1495</td>
<td>2748</td>
</tr>
<tr>
<td></td>
<td>#non-corrected records</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>#corrected records</td>
<td>1490</td>
<td>2748</td>
</tr>
<tr>
<td></td>
<td>#warning records</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>affected var. average</td>
<td>7.92</td>
<td>7.16</td>
</tr>
<tr>
<td></td>
<td>imputation var. average</td>
<td>5.22</td>
<td>6.44</td>
</tr>
<tr>
<td></td>
<td>wrong-range var. average</td>
<td>1.47</td>
<td>5.93</td>
</tr>
<tr>
<td></td>
<td>wrong-edits var. average</td>
<td>6.73</td>
<td>1.69</td>
</tr>
<tr>
<td></td>
<td>connected var. average</td>
<td>225.13</td>
<td>5.50</td>
</tr>
<tr>
<td></td>
<td>donor record distance average</td>
<td>0.98</td>
<td>1.10</td>
</tr>
<tr>
<td>Overall time</td>
<td></td>
<td>3965.94</td>
<td>341.16</td>
</tr>
</tbody>
</table>
1 Introduction
- Preface
- Previous experience
- What is TEIDE?
- Implemented methodologies
- What is TEIDE 1?
- What is TEIDE 2?
- A few details

2 On-line demonstration
- Statistical agencies using TEIDE
- Computational experiments

3 Conclusions
Finally:

- Download TEIDE 1 from http://www.goma.ull.es
- University of La Laguna (Tenerife) needs economical support to keep working on TEIDE 2, the open-source version.
- If you need help when using TEIDE: Contact me to jjsalaza@ull.es, or convince your boss to visit me in Tenerife.