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

You are here...

Limitations

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:

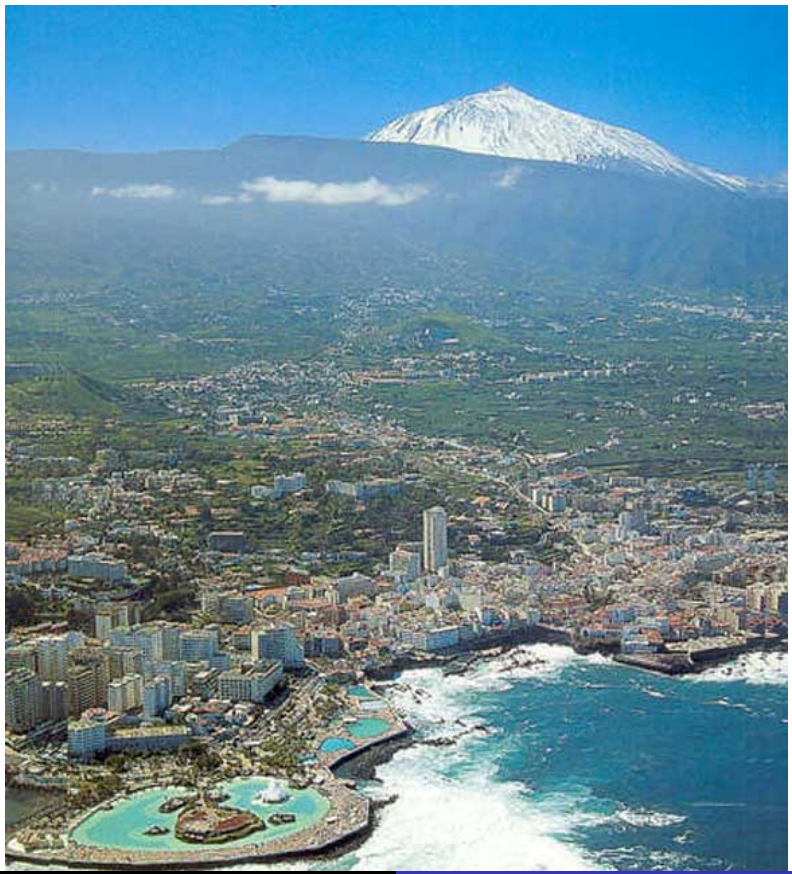
- Finding suppressions in SDC is like finding errors in SDE.
- Finding the expected attacker value of a suppression is like finding the imputation of an error.

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. **T**écnicas de **E**dición e **I**mputación de **D**atos **E**stadísticos 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 $(x_1 = 2 * x_4 + x_2)$
   - Logical $(if\ (x_1 = a)\ then\ (x_3 \leq b))$
   - Mixed $(if\ (x_4 * x_2 = 6)\ then\ (x_1 = 1))$

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 syntaxis 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 syntaxis 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`

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

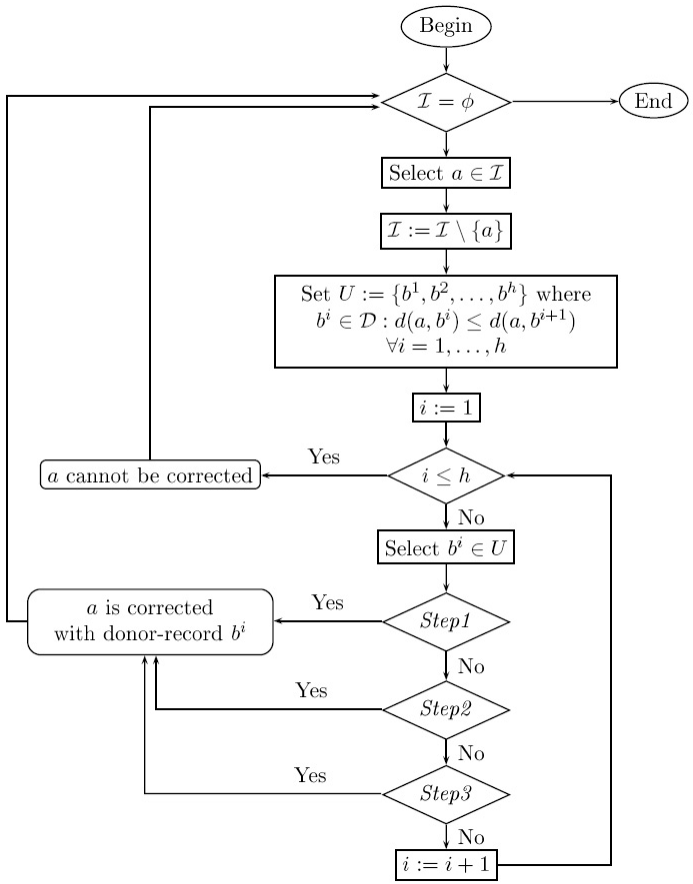
Edits are located in three groups:

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


Begin
$\mathcal{I} = \phi$
End
Select $a \in \mathcal{I}$
$\mathcal{I} := \mathcal{I} \setminus \{a\}$
Set $U := \{b^1, b^2, \ldots, b^h\}$ where $b^i \in \mathcal{D} : d(a, b^i) \leq d(a, b^{i+1})$ $\forall i = 1, \ldots, h$
$i := 1$
$i \leq h$
Yes
a cannot be corrected
No
Select $b^i \in U$
Step1
Yes
a is corrected with donor-record b^i
No
Step2
Yes
No
Step3
Yes
No
$i := i + 1$

You are here...

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).
- ISTAC: “Encuesta de Condiciones de Vida de Hogares e Individuos Canarios”, “Encuesta de Salud", .$^{\text{En}}$cuesta de Turismo", Çesta de la Compra”, “Encuesta de Tecnología en Hogares”.
- IAE: “Encuesta Mundial de Valores”, “Encuesta Social”

We thanks these organizations for supporting TEIDE

Some benchmarks:

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
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
	connected var. average	225.13	5.50	195.63
	donor record distance average	0.98	1.10	1.35
Overall time		3965.94	341.16	5899.28

You are here...

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.