New Algorithms for the Editing-and-Imputation Problem

Juan-José Salazar-González

DEIOC, University of La Laguna, Tenerife, Spain

jjsalaza@ull.es   http://webpages.ull.es/users/jjsalaza

Joint work with Jorge Riera-Ledesma and Sergio Delgado-Quintero

UNECE Work session on Statistical Data Editing

INE (Madrid) October 20-22, 2003

Work partially funded by "Ministerio de Ciencia y Tecnología"

(TIC2002-00895) and by "Instituto Canario de Estadística"

Outline of this presentation:

1. Introduction to the problem (notation, references, general algorithms)
2. New Mixed Integer Linear Programming Model
3. Algorithms for solving the new model
4. Computational experiments on benchmark instances
5. TEIDE: a new software for categorical data

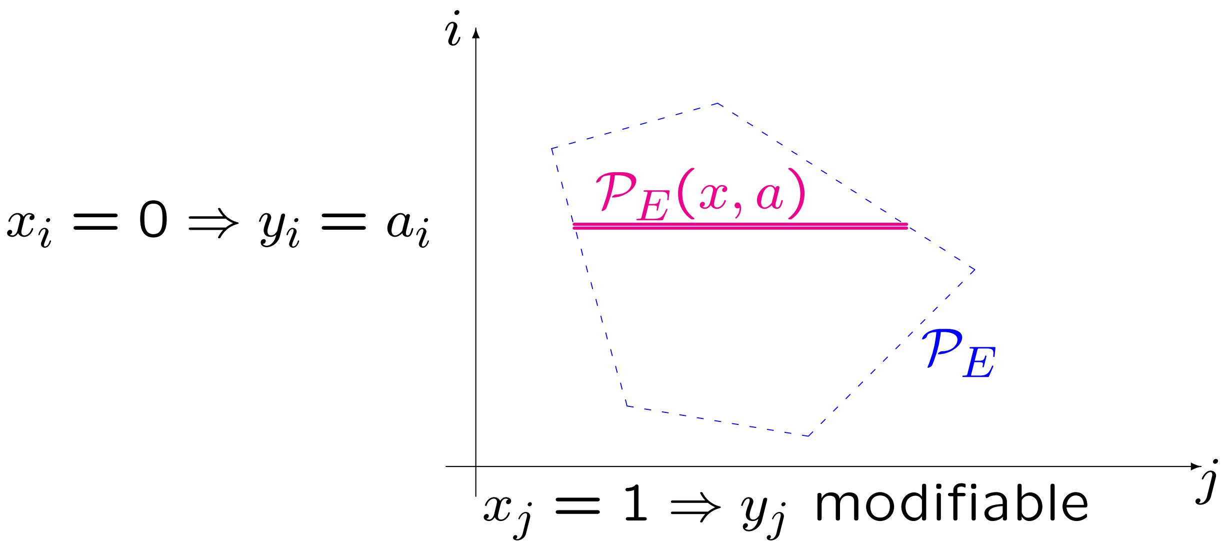
Introduction to the Editing-and-Imputation Problem (EIP):

- Let a be a record with n components indexed in $I := \{1, \ldots, n\}$.
- Let E be a set of rules (named edits) indexed in $J := \{1, \ldots, m\}$.

Let $\mathcal{P}_E$ be the set of all possible records satisfying all edits in E, each one called *valid record*. For a given $x \in \{0, 1\}^n$ and a given a, let

$$\mathcal{P}_E(x, a) := \left\{ y \in \mathcal{P}_E \ : \ y_i = a_i \text{ if } x_i = 0, \text{ for each } i \in I \right\}$$

be a projection of $\mathcal{P}_E$ in the space of the modifiable fields according to x.



Then the EIP is minimize $\left\{ w^T x \ : \ \mathcal{P}_E(x, a) \neq \emptyset \text{ and } x \in \{0, 1\}^n \right\}$, which is a combinatorial optimization problem of type $\mathcal{NP}$-hard in the strong sense.

Previous works:

- I.P. Fellegi, D. Holt, "A systematic approach to automatic edit and imputation", *Journal of the American Statistical Association* 71 (1976) 17–35.
- G. E. Liepins, "A rigourous and systematic approach to automatic data editing and its statistical basis", ORNL/TM -7126, 1980.
- J. Schaffer, "Procedure for solving the data-editing problem with both continuous and discrete data types", *Naval Research Logistics* 34 (1987) 879–890.
- R.S. Garfinkel, A.S. Kunnathur, G.E. Liepins, "Optimal imputation of errorneous data: continuous data, linear constraints", *Operations Research* 34 (1986) 744–751.
- R.S. Garfinkel, A.S. Kunnathur, G.E. Liepins, "Error location for errorneous data: continuous data, linear constraints", *SIAM J. on Scientific and Stat. Computing* 9 (1988) 922–931.
- P.G. McKeown, "A mathematical programming approach to editing of continuous survey data", *SIAM Journal on Scientific and Statistical Computing* 5 (1984) 785–797.
- C.T. Ragsdale, P.G. McKeown, "On solving the continuous data editing problem", *Computers & Operations Research* 23 (1996) 263–273.
- J. Kovar, W.E. Winkler, "Editing economic data", working paper, 2000.
- R. Bruni, A. Sassano, "Logic and optimization techniques for an error free data collecting", working paper, University of Roma, 2001.

Old general algorithm:

Starting from Fellegi and Holt (1976), a commonly used methodology is:

Step 0: Let $K \subseteq E$ be the set of edits not satisfied by the record a. Let x^* be an optimal integer solution of the *Set Covering Problem* (SCP):

$$\text{minimize } \sum_{i \in I} w_i x_i \tag{1}$$

subject to

$$\sum_{i \in I_k} x_i \geq 1 \qquad \text{for all } k \in K \tag{2}$$

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I, \tag{3}$$

where $I_k \subseteq I$ is the subset of fields involved in the edit k.

Step 1: If $\mathcal{P}_E(x^*, a) \neq \emptyset$ then stop (x^* is an optimal EIP solution). Otherwise, find a new violated implicit edit k', add the constraint

$$\sum_{i \in I_{k'}} x_i \geq 1 \tag{4}$$

to the constraint family (2), update x^* with an optimal solution of the new SCP and go to Step 1.

New general algorithm:

Inspired by previous works, we can propose the following general approach:

Step 0: Let $K \subseteq E$ be the set of edits not satisfied by the record a. Let x^* be an optimal integer solution of the *Set Covering Problem* (SCP):

$$\text{minimize} \sum_{i \in I} w_i x_i \tag{5}$$

subject to

$$\sum_{i \in I_k} x_i \geq 1 \qquad \text{for all } k \in K \tag{6}$$

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I, \tag{7}$$

where $I_k \subseteq I$ is the subset of fields involved in the edit k.

Step 1: If $\mathcal{P}_E(x^*, a) \neq \emptyset$ then stop (x^* is an optimal EIP solution). Otherwise, add the constraint

$$\sum_{i \in I : x_i^* = 0} x_i \geq 1 \tag{8}$$

to the constraint family (6), update x^* with an optimal solution of the new SCP and go to Step 1.

New mathematical model for continuous data and linear edits:

- Each component a_i of the given record a is a continuous number in the known interval $[lb_i, ub_i]$, for all $i \in I$.
- Each edit can be written as a finite set of linear inequalities, thus

$$\mathcal{P}_E := \left\{ y \in [lb_1, ub_1] \times \ldots \times [lb_n, ub_n] : \sum_{i \in I} m_{ij} y_i \leq b_j \text{ for all } j \in J \right\}$$

  is a polytope, shortly denoted by $\mathcal{P}_E = \{y \in \mathbb{R}^n : My \leq b, lb \leq y \leq ub\}$.

Then the EIP can be formulated as a Mixed Integer Linear Programming (MILP) problem:

$$\text{minimize} \sum_{i \in I} w_i x_i \tag{9}$$

subject to

$$\sum_{i \in I} m_{ij} y_i \leq b_j \qquad \text{for all } j \in J \tag{10}$$

$$a_i - (a_i - lb_i) x_i \leq y_i \leq a_i + (ub_i - a_i) x_i \qquad \text{for all } i \in I \tag{11}$$

$$x_i \in \{0, 1\} \qquad \text{for all } i \in I. \tag{12}$$

A similar MILP model with double number of variables:

As done by Ragsdale and McKeown (1996), it is possible to write a similar model by considering two 0-1 variables associated to each field i:

$$x_i^- = \begin{cases} 1 & \text{if } y_i < a_i \\ 0 & \text{otherwise} \end{cases} \qquad x_i^+ = \begin{cases} 1 & \text{if } y_i > a_i \\ 0 & \text{otherwise.} \end{cases}$$

Then the EIP is equivalent to:

$$\text{minimize } \sum_{i \in I} w_i (x_i^- + x_i^+)$$

subject to

$$\begin{aligned} \sum_{i \in I} m_{ij} y_i \le b_j & \qquad \text{for all } j \in J \\ a_i - (a_i - lb_i) x_i^- \le y_i \le a_i + (ub_i - a_i) x_i^+ & \qquad \text{for all } i \in I \\ x_i^-, x_i^+ \in \{0, 1\} & \qquad \text{for all } i \in I. \end{aligned}$$

The inequality $x_i^- + x_i^+ \le 1$ is unnecessary due to the objective function.

Still, all the ideas introduced for model (5)–(7) apply also to this extension.

FIRST algorithm for the new MILP model:

At a first glance, the model (9)–(12) can be given to a general-purpose MILP optimizer performing a branch-and-bound scheme, like CPLEX developed by ILOG (www.ilog.com), XPRESS-MP developed by DASHOPTIMIZATION (www.dashoptimization.com), GLPK developed by GNU (www.gnu.edu), or ABACUS with SOPLEX developed by ZIB (www.zib.de).

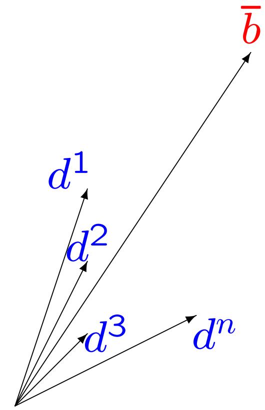
A classical disadvantage is that a LP-based optimizer must internally face ill-conditioned mathematical operations due to constraints (11), leading to numerical problems and wrong solutions. One can try to reduce the number of this bad situations by appropriately turning the tolerance parameters, but it is difficult to find good parameters for most of the EIP instances.

We can help the general-purpose solver by considering, for example, the ad-hoc branching rule: if (x^*, y^*) is an optimal solution of a linear relaxation of the MILP model (9)–(12), then we choose a variable x_i with a non-integer value x_i^* and then we create two subproblems in the branch-decision tree by imposing either $x_i = 1$ or $y_i = a_i$.

Background in Duality Theory:

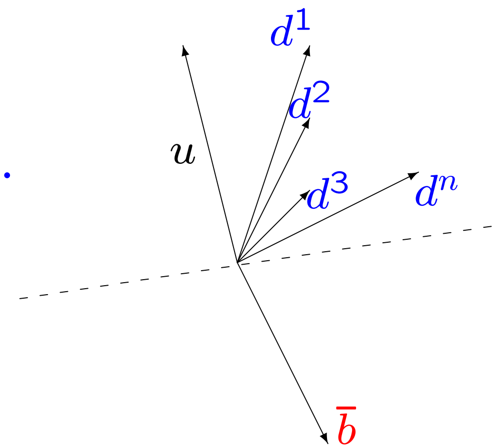
The Farkas' Lemma (1894) says:

Given a set of vectors $d^1, \ldots, d^n$ and $\bar{b}$ in $\mathbb{R}^m$, then



- **either** there are non-negative numbers $y_1, \ldots, y_n$ such that
$$\bar{b} = y_1 d^1 + \cdots y_n d^n,$$

- **or** there is a vector u in $\mathbb{R}^n$ such that
$$d^{1T} u \geq 0, \ \ldots, \ d^{nT} u \geq 0, \ \text{and } \bar{b}^T u < 0.$$



In other words, denoting $\bar{M} := [d^1 \cdots d^n]$:

"The polyhedron $\{y \in \mathbb{R}^n \ : \ \bar{M} y = \bar{b}, y \geq 0\}$ has a solution if and only if all the solutions u of the cone $\{u \in \mathbb{R}^m \ : \ \bar{M}^T u \geq 0\}$ satisfy also $\bar{b}^T u \geq 0$."

[This result was also used by Garfinkel, Kunnathur, Liepins (1986)]

SECOND algorithm for the new MILP model:

By using the Farkas' Lemma, $\mathcal{P}_E(x^*, a) \neq \emptyset$ if and only if

$$\sum_{j \in J} \alpha_j b_j + \sum_{i \in I} \beta_i (a_i + (\mathsf{ub}_i - a_i) x_i^*) - \sum_{i \in I} \gamma_i (a_i - (a_i - \mathsf{lb}_i) x_i^*) \geq 0 \qquad (13)$$

for all the directions of the cone:

$$\mathcal{C}_E := \{(\alpha, \beta, \gamma) : M^T \alpha + \beta - \gamma = 0, \alpha \geq 0, \beta \geq 0, \gamma \geq 0\}.$$

Therefore, a solution x is admissible (or feasible) if and only if it satisfies:

$$\sum_{i \in I} [\beta_i (\mathsf{ub}_i - a_i) + \gamma_i (a_i - \mathsf{lb}_i)] x_i \geq \alpha^T (Ma - b) \qquad (14)$$

for all $(\alpha, \beta, \gamma) \in \mathcal{C}_E$.

Step 0: Let us define a master problem as the set covering problem of Step 0 in the previous algorithms, and let x^* be an optimal solution.

Step 1: If $\mathcal{P}_E(x^*, a) \neq \emptyset$ then stop (x^* is an optimal EIP solution). Otherwise, find an inequality (14) violated by x^*, add it to the master problem, update x^* with an optimal solution of the new master problem and go to Step 1.

Implementing the SECOND algorithm:

Given a solution x^*, the problem in Step 1 of finding a violated inequality (14), if any exists, is called *separation problem* and it is equivalent to

$$\text{minimize} \ \sum_{j \in J} b_j \alpha_j + \sum_{i \in I} (a_i + (\mathsf{ub}_i - a_i) x_i^*) \beta_i - \sum_{i \in I} (a_i - (a_i - \mathsf{lb}_i) x_i^*) \gamma_i$$

subject to

$$M^T \alpha + \beta - \gamma = 0,$$

$$\alpha \geq 0, \beta \geq 0, \gamma \geq 0.$$

If the optimal objective value is negative then the optimal solution $(\alpha^*, \beta^*, \gamma^*)$ defines a violated inequality (14) to be considered in the master problem.

Advantages of this Benders' decomposition approach:

- The cut generation procedure can also be applied when x^* is non-integer.
- Inequality (14) can be strengthened by rounding some coefficients.
- Other families of inequalities (cliques, Gomory,...) can be also added.

Clique inequalities: If $x_{i'} + x_{i''} \geq 1$ for all $i', i'' \in S$ then $\sum_{i \in S} x_i \geq |S| - 1$.

Preliminary computational results:

Algorithm 0: Our new cutting-plane algorithm based on inequalities (8).

Algorithm 1: A general-purpose MILP optimizer on the model (9)–(12).

Algorithm 2: The cutting-plane algorithm described in Garfinkel, Kunnathur and Liepins (1988).

Algorithm 3: The cutting-plane algorithm described in Ragsdale and McKeown (1996).

Algorithm 4: Our branch-and-cut algorithm where only integer solutions x^* are separated, which turns to be a new cutting-plane algorithm based on inequalities (14).

Algorithm 5: Our branch-and-cut algorithm when the separation problem is solved on integer and non-integer solutions x^*.

All implementations done by the same human programmer, using the C++ programming language on a personal computer Pentium 1500 Mhz running Windows XP. CPLEX 8.1 was used as MILP optimizer. Time limit: 1 hour.

Benchmark instances:

Class I: They are the instances used in Ragsdale and McKeown (1996). Hence, $|I| = n = 50$, $|J| = m = 20$, $w_i = 1$ for all $i \in I$, $a_i \in [-100, +100]$, $b_j \in [0, 1000]$; m_{ij} are zero with probability 0.2, in [1,20] with probability 0.24 and in [-20,-1] with probability 0.56. We set $lb_i = -100$ and $ub_i = 100$ for all $i \in I$. The FORTRAN code of the random generator was kindly provided by Cliff Ragsdale.

Class II: They are exactly as before but with $|I| = 100$ and $|J| = 40$. We have considered three families by also considering different intervals $[lb_i, ub_i]$ in $[-10^3, 10^3]$, $[-10^4, 10^4]$ and $[-10^5, 10^5]$.

Class III: They are artificial instances kindly supplied by William Winkler and María García (US Census of Bureau) consisting of 10,994 records with 17 fields and two set of edits. The first set contains 136 edits like

$$l_j \leq \frac{y_{i'}}{y_{i''}} \leq u_j \qquad \text{for some } i', i'' \in I (i' < i'') \text{ and } j \in J,$$

in which each $u_j - l_j$ takes a value between 10^{-1} and 10^7. The second set of edits contains to two balancing edits like

$$y_{i'} + y_{i''} = y_{i'''} \qquad \text{for some } i', i'', i'' \in I.$$

Average results on five instances from Class I:

			Algorithm 1			Algorithm 5			
Failed	*#*	*Obj.*	*Cuts*	*Nodes*	*Time*	(14)	*Iter.*	*Nodes*	*Time*
1-4	5	3.4	3.4	2.0	0.047	9.6	20.6	3.2	0.043
5-8	5	5.0	4.2	8.4	0.066	15.6	35.4	12.4	0.078
9-12	5	5.6	6.6	30.2	0.116	44.6	93.6	33.8	0.206
13-16	5	6.8	4.2	149.6	0.418	70.2	120.2	34.0	0.287
17-20	5	7.4	4.6	289.2	0.631	245.0	550.8	223.4	1.259

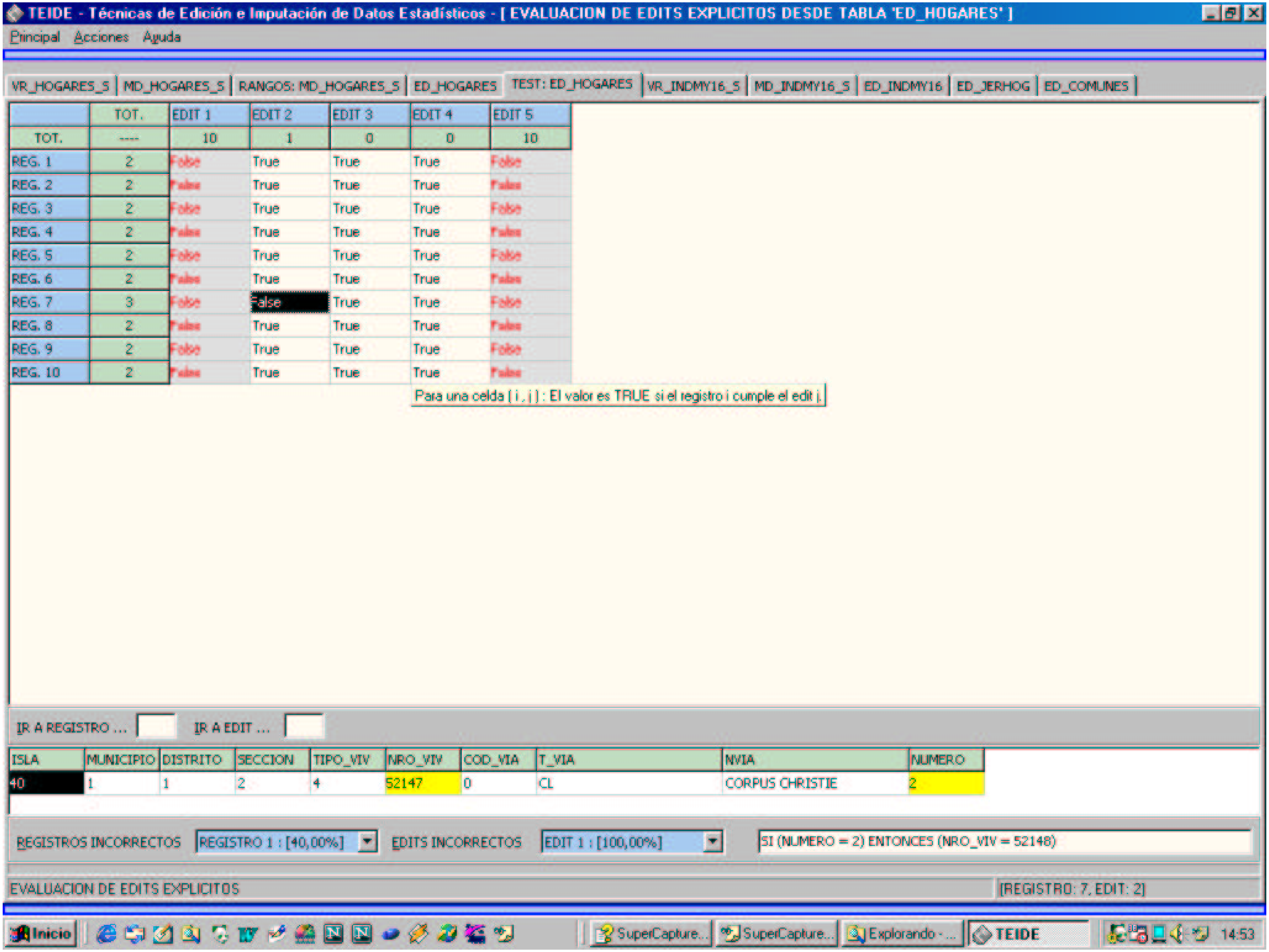
	Algorithm 0			Algorithm 2		Algorithm 3		Algorithm 4	
Failed	*Iter.*	*Time*	*Ok*	*Iter.*	*Time*	*Iter.*	*Time*	(14)	*Time*
1-4		18.3	3	48.0	0.828			4.0	0.029
5-8		-	0	133.0	5.228			9.4	0.131
9-12		6.3	1	228.0	23.869			12.6	0.309
13-16		-	0	666.0	296.684			15.4	0.578
17-20		-	0	684.0	231.087			26.6	1.728

Average results on five instances from Class II:

$[lb_i, ub_i]$	*Failed*	*#*	*Obj.*	*Algorithm 1* Nodes	*Algorithm 1* Time	*Algorithm 1* Ok	*Algorithm 5* (14)	*Algorithm 5* Nodes	*Algorithm 5* Time
	1-8	5	6.4	926.4	6.2	5	1189.4	631.2	6.7
	9-16	5	9.2	6297.4	27.7	5	1641.4	661.6	9.4
$[-10^3, 10^3]$	17-24	5	10.0	9822.2	47.1	5	17816.0	4438.0	115.6
	25-32	5	12.0	8752.8	53.1	5	14348.6	4170.2	88.1
	33-40	5	12.4	6336.0	44.6	5	17275.0	3511.2	120.9
	1-8	5	3.6	1389.2	17.7	5	138.8	122.6	1.0
	9-16	5	5.0	25644.8	187.4	5	2254.6	1159.6	13.8
$[-10^4, 10^4]$	17-24	5	5.6	149298.5	1595.9	4	13319.8	3199.4	92.0
	25-32	5	5.8	93463.0	1011.3	4	19524.0	4104.2	127.2
	33-40	5	6.8	232025.0	3063.8	4	87244.0	15334.6	1452.4
	1-8	5	3.6	6501.8	54.0	5	100.4	67.8	0.6
	9-16	5	4.8	96563.5	1366.0	4	1741.4	1930.4	14.4
$[-10^5, 10^5]$	17-24	5	6.4	-	-	0	2210.2	2536.2	17.0
	25-32	5	6.6	-	-	0	3868.6	4063.8	31.0
	33-40	5	7.6	-	-	0	7054.0	7621.0	57.2

Average results on instances from Class III:

			Algorithm 1			*Algorithm 5*				
Failed	#	*Obj.*	*Cuts*	*Nodes*	*Time*	*Clique*	(14)	*Iter.*	*Nodes*	*Time*
1-15	223	1.75	16.8	0.229	0.008	12.135	0.135	1.188	0.000	0.004
16-30	5384	2.91	69.5	0.673	0.016	15.996	0.297	1.331	0.001	0.005
31-45	4281	4.10	72.7	1.342	0.026	19.050	0.308	1.314	0.000	0.005
46-60	1018	5.28	70.7	3.949	0.037	36.357	0.315	1.356	0.000	0.006
61-75	87	6.20	62.0	8.034	0.046	62.207	0.287	1.471	0.000	0.005


TEIDE - Técnicas de Edición e Imputación de Datos Estadísticos - [EVALUACION DE EDITS EXPLICITOS DESDE TABLA 'ED_HOGARES']
Principal Acciones Ayuda
VR_HOGARES_5 | MD_HOGARES_5 | RANGOS: MD_HOGARES_5 | ED_HOGARES | TEST: ED_HOGARES | VR_INDMY16_5 | MD_INDMY16_5 | ED_INDMY16 | ED_JERHOG | ED_COMUNES
| TOT. | EDIT 1 | EDIT 2 | EDIT 3 | EDIT 4 | EDIT 5
TOT. | ---- | 10 | 1 | 0 | 0 | 10
REG. 1 | 2 | False | True | True | True | False
REG. 2 | 2 | False | True | True | True | False
REG. 3 | 2 | False | True | True | True | False
REG. 4 | 2 | False | True | True | True | False
REG. 5 | 2 | False | True | True | True | False
REG. 6 | 2 | False | True | True | True | False
REG. 7 | 3 | False | False | True | True | False
REG. 8 | 2 | False | True | True | True | False
REG. 9 | 2 | False | True | True | True | False
REG. 10 | 2 | False | True | True | True | False
Para una celda [i , j] : El valor es TRUE si el registro i cumple el edit j.
IR A REGISTRO ... IR A EDIT ...
ISLA | MUNICIPIO | DISTRITO | SECCION | TIPO_VIV | NRO_VIV | COD_VIA | T_VIA | NVIA | NUMERO
40 | 1 | 1 | 2 | 4 | 52147 | 0 | CL | CORPUS CHRISTIE | 2
REGISTROS INCORRECTOS REGISTRO 1 : [40,00%]
EDITS INCORRECTOS EDIT 1 : [100,00%]
SI (NUMERO = 2) ENTONCES (NRO_VIV = 52148)
EVALUACION DE EDITS EXPLICITOS
[REGISTRO: 7, EDIT: 2]
Inicio
SuperCapture...
SuperCapture...
Explorando - ...
TEIDE
14:53