ISSN: 1696-2281
eISSN: 2013–8830
SORT 35 (1) January-June (2011)

SORT

Statistics and Operations Research Transactions

Sponsoring institutions

Universitat Politècnica de Catalunya
Universitat de Barcelona
Universitat de Girona
Universitat Autònoma de Barcelona
Institut d'Estadística de Catalunya

Supporting institution

Spanish Region of the International Biometric Society

Generalitat
de Catalunya
Institut d'Estadística
de Catalunya

SORT

Volume 35
Number 1
January-June 2011
ISSN: 1696-2281
eISSN: 2013-8830

Articles

Selected article from *XII Conferencia Española de Biometría 2009*

Information for authors and subscribers

SORT 35 (1) January-June 2011, 3-20

Extending controlled tabular adjustment for non-additive tabular data with negative protection levels

Jordi Castro[*]

Universitat Politècnica de Catalunya

Abstract

Minimum distance controlled tabular adjustment (CTA) is a recent perturbative methodology for the protection of tabular data. An implementation of CTA was recently used by Eurostat for the protection of European Union level structural business and animal production statistics. The real-world instances to be solved forced the classical CTA model to be extended with two features: first, to deal with non-additive tables; second, and most important, to consider negative protection levels. The latter extension means a significant modification of the classical CTA mixed integer linear model. We present and compare new models for these extensions. Computational results are reported using a set of real-world instances, and two state-of-the-art commercial solvers (CPLEX and Xpress).

MSC: 90C11, 90C06, 90B99, 62Q05

Keywords: Mathematical programming, mixed integer linear optimization, statistical disclosure control, perturbative methods, controlled tabular adjustment, official statistics

1. Introduction

Tabular data protection is one of the two disciplines of the statistical disclosure control field (microdata protection being the second one). The interested reader is addressed to the recent research monographs Willenborg and de Waal (2000); Domingo-Ferrer and Franconi (2006); Domingo-Ferrer and Franconi (2008) for an overview of this field. Controlled tabular adjustment (CTA) and other minimum distance related variants were suggested in Dandekar and Cox (2002) and Castro (2006) as a replacement to other

[*] Dept. of Statistics and Operations Research. Universitat Politècnica de Catalunya. Jordi Girona 1–3, 08034 Barcelona. jordi.castro@upc.edu. http://www-eio.upc.es/˜jcastro
Received: April 2010
Accepted: November 2010

Table 1: *(a) Sizes of optimization problems associated to cell suppression (CSP), controlled rounding (CRP) and CTA. (b) Figures for a particular table of 4000 cells, 1000 sensitive cells, and 2500 linear relations.*

Problem	constraints	continuous	binary
CSP/CRP	$2(m+2n)s$	$2ns$	n
CTA	$m+4s$	$2n$	s

(a)

Problem	constraints	continuous	binary
CSP/CRP	21,000,000	8,000,000	4,000
CTA	6,500	8,000	1,000

(b)

computationally more expensive approaches for tabular data protection. CTA can be seen as a method for generating a safe synthetic table, which is as close as possible to the original table. This is obtained by solving the following optimization problem: given a non-safe table, with a set of sensitive cells to be protected, find the closest safe table to the original one (according to some distance) by adding the minimum amount of perturbations. Some of the good properties of CTA are:

- It can be applied to any table or set of linked tables. Even for complex and large tables a solution can be obtained in reasonable time (likely suboptimal, but with an acceptable optimality gap).
- From a computational point of view, the size of the resulting optimization problem is by far lower than for other well-known protection methods, such as the cell suppression problem (CSP) (Castro (2007a)) and the controlled rounding problem (CRP) (Salazar-González (2006)). For a table of n cells, s of them being sensitive, and m table linear relations, Table 1*(a)* shows the dimensions of the optimization problem formulated by CSP, CRP and CTA (number of constraints, and number of continuous and binary variables). For example, the particular figures for a table of 4000 cells, 1000 sensitive cells, and 2500 linear relations are provided in Table 1*(b)*, clearly showing the different order of magnitude between the optimization problems.
- State-of-the-art solvers, such as CPLEX (IBM ILOG CPLEX (2009)) or Xpress (FICO Dash Xpress (2008)), can be applied to the solution of CTA (at least for medium size instances). Other approaches like CSP or CRP require specialized solution methods, either optimal or heuristic, even for small instances. For very large-scale instances, it is possible to develop specialized, hopefully more efficient, procedures for CTA. Some preliminary work has already been started (Castro and Baena (2008), González and Castro (2009)), but they are beyond the scope of this work.

- Either L_1, L_∞ or Euclidean L_2 distances can be used in the objective function of CTA. L_2 distances provide mixed integer quadratic problems, which are more difficult to be solved, but reduce the largest deviations. L_1 provides simpler optimization problems, and it is currently mostly used by National Statistical Institutes. All the models in this paper use L_1.
- The particular model of CTA with the L_1 distance does not guarantee integrality of the perturbations (i.e., they can be fractional values); models with other distances (L_2 or L_∞) neither guarantee integrality. Indeed, it is possible to obtain tables where the perturbations are fractional (e.g., three-dimensional tables are modeled as a multicommodity flow problem (Castro (2005, 2007b)), which is known not to provide integral flows). However, in most tables tested with the L_1 distance, the solution provided was integer without imposing integrality of perturbations (however, we do not claim the matrices were totally unimodular, which is sufficient for guaranteeing integrality). Even if perturbations were not integer, they would still be valid for magnitude tables.
- Previous empirical testing (Castro and Giessing (2006)) showed the quality of the solution (measured as number of cells with large significant deviations) provided by CTA was comparable, even higher, than that obtained with CSP. Other quality criteria (Cox, Kelly and Patil (2004)) can also be easily added to the CTA formulation.

A package implementing CTA (Castro, González and Baena (2009)) has recently been incorporated within a wider scheme for the protection of structural business statistics disseminated by Eurostat (project coordinated by Statistics Netherlands, with the participation of Destatis and Universitat Politècnica de Catalunya) (Giessing, Hundepool and Castro (2009)). When applying the same scheme to the protection of animal production statistics of the European Union two unforeseen features of CTA were required: it should deal with non-additive tables, and it should cope with negative protection levels. While the former is a simple extension, the latter significantly changes the optimization model; even worse, the solution space of the models with negative protection levels increases (as shown in Section 4), and it may make harder finding an optimal or good solution. Non-additivity may result when dealing with externally obtained tables, with empty or approximate cells. Negative protection levels can be used to deal with correlated tables. More details will be provided at the beginning of sections 3 and 4. In this paper we discuss several models for the general CTA problem with either positive and negative protection levels, and either additive or non-additive tables. The computational results show which is the most effective variant to be used in practice for real-world instances. The most efficient model turned out to be as efficient as the standard CTA model, being much more general: it can deal with either additive or non-additive tables, and with positive and negative protection levels.

The structure of the paper is as follows. Section 2 outlines the standard CTA formulation, which is the basis for the extensions of subsequent sections. Sections 3 and 4 show the new models to deal with non-additive tables and negative protection levels. Section 5 reports the computational results obtained with the several resulting models in the solution of a set of real-world instances.

2. The standard CTA model

Any CTA instance, either with one table or a number of tables, can be represented by the following parameters:

- A set of cells $a_i, i \in \mathscr{N} = \{1,\dots,n\}$, that satisfy some linear relations $Aa = b$ (a being the vector of a_i's), and a vector $w \in \mathbb{R}^n$ of positive weights for the deviations of cell values.
- A lower and upper bound for each cell $i \in \mathscr{N}$, respectively l_{x_i} and u_{x_i}, which are considered to be known by any attacker. If no previous knowledge is assumed for cell i, then $l_{x_i} = 0$ ($l_{x_i} = -\infty$ if $a \geq 0$ is not required) and $u_{x_i} = +\infty$ can be used.
- A set $\mathscr{S} = \{i_1, i_2, \dots, i_s\} \subseteq \mathscr{N}$ of indices of confidential or sensitive cells.
- A lower and upper protection level for each confidential cell $i \in \mathscr{S}$, respectively lpl_i and upl_i, such that the released values satisfy either $x_i \geq a_i + upl_i$ or $x_i \leq a_i - lpl_i$.

CTA attempts to find the closest safe values $x_i, i \in \mathscr{N}$, according to some distance L, that makes the released table safe. This involves the solution of the following optimization problem:

$$\begin{array}{rl} \min\limits_{x} & ||x-a||_L \\ \text{subject to} & Ax = b \\ & l_x \leq x \leq u_x \\ & x_i \leq a_i - lpl_i \text{ or } x_i \geq a_i + upl_i \quad i \in \mathscr{S}. \end{array} \tag{1}$$

Problem (1) can also be formulated in terms of deviations from the current cell values. Defining $z = x - a$, $l_z = l_x - a \leq 0$, and $u_z = u_x - a \geq 0$, we obtain

$$\begin{array}{rl} \min\limits_{z} & ||z||_L \\ \text{subject to} & Az = 0 \\ & l_z \leq z \leq u_z \\ & z_i \leq -lpl_i \text{ or } z_i \geq upl_i \quad i \in \mathscr{S}. \end{array} \tag{2}$$

Using the L_1 distance weighted by w, and introducing variables $z^+, z^- \in \mathbb{R}^n$ so that $z = z^+ - z^-$ and $|z| = z^+ + z^-$, and binary variables $y \in \{0,1\}^s$ the final MILP model for CTA is:

$$\begin{aligned}
\min_{z^+,z^-,y} \quad & \sum_{i=1}^{n} w_i(z_i^+ + z_i^-) && \text{(3a)}\\
\text{subject to} \quad & A(z^+ - z^-) = 0 && \text{(3b)}\\
& 0 \le z_i^+ \le u_{z_i}, \quad 0 \le z_i^- \le -l_{z_i} \;\; i \in \mathscr{N} \setminus \mathscr{S} && \text{(3c)}\\
& y \in \{0,1\}^s && \text{(3d)}\\
& \left.\begin{array}{ll} upl_i y_i & \le z_i^+ \le u_{z_i} y_i \\ lpl_i(1-y_i) & \le z_i^- \le -l_{z_i}(1-y_i) \end{array}\right\} i \in \mathscr{S} && \text{(3e)}
\end{aligned}$$

Constraints (3b) impose feasibility of the published perturbed table. Constraints (3c) guarantee perturbations are within allowed bounds. Constraints (3d)–(3e) force the new table to be safe. When $y_i = 1$ the constraints mean $upl_i \le z_i^+ \le u_{z_i}$ and $z_i^- = 0$, thus the protection sense is “upper”; when $y_i = 0$ we get $z_i^+ = 0$ and $lpl_i \le z_i^- \le -l_{z_i}$, thus the protection sense is “lower”.

3. Non-additive tables

In some instances the original cell values do not satisfy $Aa = b$. This is mainly due to missing or approximate cell values of externally provided tables, which may require the application of cell imputation techniques. This is specially relevant for data managed by Eurostat, where the sources are different countries of the European Union. In particular, this requirement was necessary for the protection of animal production statistics (i.e., milk production) at the European Union and state members levels. Tables already protected (i.e., they contained missing information) for each member state were received. The protection of this set of tables, together with the European Union totals, can be accomplished by first estimating values for the missing information, although they result in non-additive tables, and using RCTA to make the resulting tables both safe and additive. Some details about the overall procedure can be found in Giessing, Hundepool and Castro (2009).

If the table is non-additive, i.e., $Aa \neq b$, then the constraints (3b) of the CTA model have to be replaced by

$$A(z^+ - z^-) = b - Aa. \tag{4}$$

Indeed, note that a deviation satisfying (4) makes the resulting table feasible:

$$A(a + z^+ - z^-) = Aa + A(z^+ - z^-) = Aa + (b - Aa) = b.$$

If the original table is already additive, then $b - Aa = 0$, and therefore (3b) and (4) are equivalent. Since (4) is more general, it should be preferred in any CTA model. Note the complexity of (3) is the same either considering (3b) or (4).

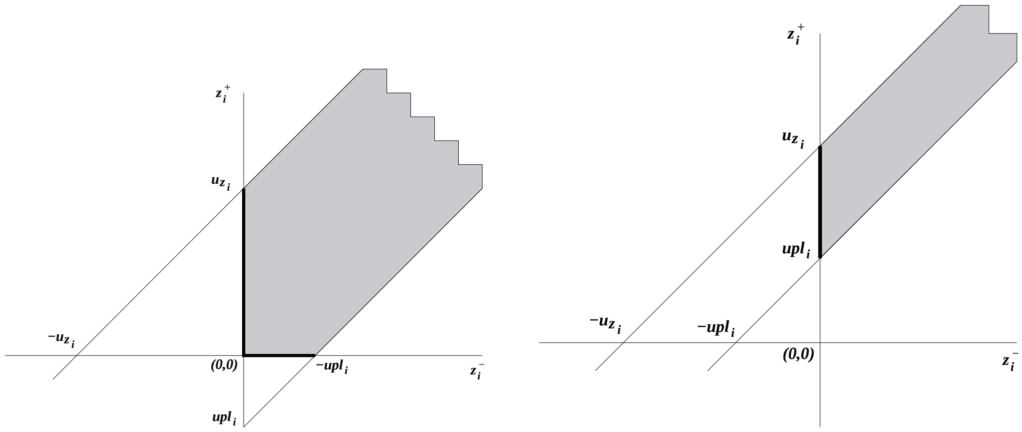
4. Negative protection levels

Negative protection levels may be required when protecting correlated tables. Protection levels lpl_i and upl_i for cell a_i preclude values of the interval $[a_i - lpl_i, a_i + upl_i]$ for this cell in the released table. Let us refer to this interval as the "protection interval". If the protection levels are positive then $a_i \in [a_i - lpl_i, a_i + upl_i]$, which is the usual situation. However, if this table is correlated with another that has been previously protected and released, we may need a protection interval that does not include a_i (for instance, to avoid that the ratios between both released tables are close to their real values). Of course if a_i is not in the protection interval, it may be released with no change, and then it could be (wrongly) assumed it is no longer a confidential cell, and that it does not require protection levels. However, because of the deviations of other cells and the preservation of the constraints (4), a positive deviation of a_i may be required in a solution, in which case the protection interval has to be considered. This issue of negative protection levels is directly related with the non-additivity of the previous section. In particular, for the real case of the European Union animal production statistics project (i.e., milk production), the presence of non-additive tables (whose values were estimated) may mean that the protection intervals have to be shifted, which may result formally in negative protection levels. Additional details can be found in Giessing, Hundepool and Castro (2009).

According to the signs of the lower and upper protection levels, there are four possible combinations that should be addressed by the new CTA model. Note that the MILP model (3) used, for instance, in Castro (2006) and Dandekar and Cox (2002) is only valid for one case, when protection levels are nonnegative. On the other hand, the generic formulation (2) is valid for the four cases, but it is not in the form of a mathematical programming problem. For instance, for a cell $a_i = 10$ with lower and upper protection levels lpl_i and upl_i, the four cases according to signs imposed by the constraints of (2) are:

- If $lpl_i = 3$ and $upl_i = 2$, then $z_i \leq -3$ or $z_i \geq 2$, i.e., the protection interval is $[7, 12]$.
- If $lpl_i = 3$ and $upl_i = -2$, then $z_i \leq -3$ or $z_i \geq -2$, i.e., the protection interval is $[7, 8]$.
- If $lpl_i = -2$ and $upl_i = 3$, then $z_i \leq 2$ or $z_i \geq 3$, i.e., the protection interval is $[12, 13]$.
- If $lpl_i = -2$ and $upl_i = -3$, then $z_i \leq 2$ or $z_i \geq -3$, i.e., any value can be released for this cell (there is no protection interval).

If the constraints (3e) were applied when protection levels are negative, then some components of z^+ or z^- would be negative, and the objective function (3a) would no longer represent the absolute value. This happens because in (3e) variables z^+ and z^- are associated to upper and lower protection deviations, instead of being auxiliary variables to model the L_1 distance.



(a) *(b)*

Figure 1: *In grey, feasible set Ω^i for $y_i = 1$, when either $upl_i \le 0$ (figure (a)) or $upl_i \ge 0$ (figure (b)).*

Let us consider the model (2), and let us introduce $z^+, z^- \in \mathbb{R}^n$ such that $z = z^+ - z^-$ and $|z| = z^+ + z^-$. Then, considering the table may be non-additive, (2) can be written as

$$\begin{array}{rl} \min\limits_{z^+,z^-} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{subject to} & A(z^+ - z^-) = b - Aa \\ & l_z \le z^+ - z^- \le u_z \\ & z_i^+ - z_i^- \le -lpl_i \text{ or } z_i^+ - z_i^- \ge upl_i \quad i \in \mathscr{S} \\ & (z^+, z^-) \ge 0. \end{array} \tag{5}$$

Introducing binary variables $y \in \{0,1\}^s$, (5) can be recast as the following MILP model:

$$\begin{array}{rl} \min\limits_{z^+,z^-,y} & \sum\limits_{i=1}^{n} w_i(z_i^+ + z_i^-) \\ \text{subject to} & (z^+, z^-, y) \in \Omega = \Omega^A \cap (\cap_{i\in\mathscr{N}} \Omega^{0_i}) \cap (\cap_{i\in\mathscr{S}} \Omega^i), \end{array} \tag{6}$$

where Ω^A, Ω^{0_i} and Ω^i are defined as

$$\Omega^A = \left\{ (z^+, z^-) : A(z^+ - z^-) = b - Aa \right\}, \tag{7}$$

$$\Omega^{0_i} = \left\{ (z_i^+, z_i^-) : l_{z_i} \le z_i^+ - z_i^- \le u_{z_i}, (z_i^+, z_i^-) \ge 0 \right\} \; i \in \mathscr{N}, \tag{8}$$

$$\begin{aligned} \Omega^i = \big\{ & (z_i^+, z_i^-, y_i) : z_i^+ - z_i^- \ge upl_i y_i + l_{z_i}(1 - y_i), \\ & z_i^+ - z_i^- \le -lpl_i(1 - y_i) + u_{z_i} y_i, (z_i^+, z_i^-) \ge 0, y_i \in \{0,1\} \big\} \; i \in \mathscr{S}. \end{aligned} \tag{9}$$

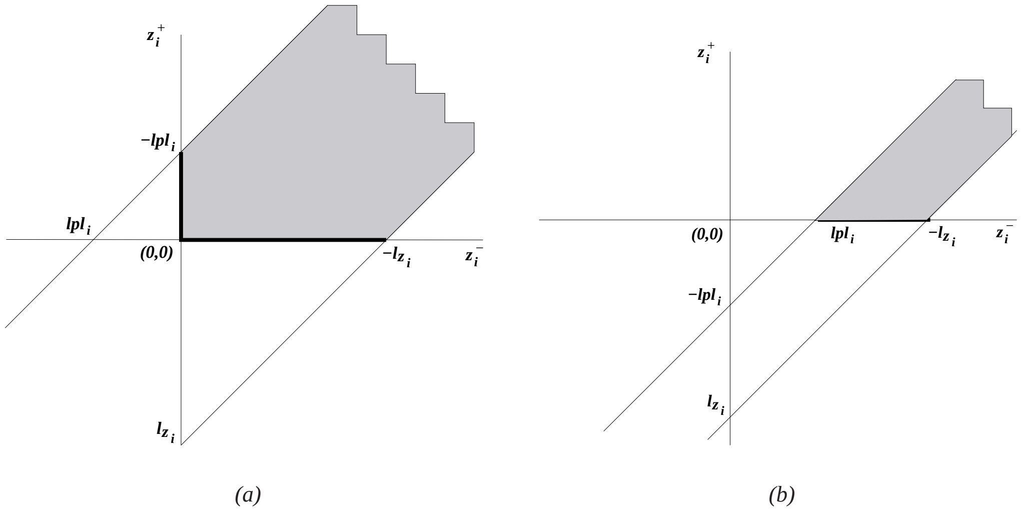


Figure 2: *In grey, feasible set Ω^i for $y_i = 0$, when either $lpl_i \le 0$ (figure (a)) or $lpl_i \ge 0$ (figure* (b)*).*

If $y_i = 1$, Ω^i reduces to

$$\left\{(z_i^+, z_i^-) : upl_i \le z_i^+ - z_i^- \le u_{z_i}, (z_i^+, z_i^-) \ge 0\right\} \tag{10}$$

i.e., the protection sense is “upper”. If $y_i = 0$, Ω^i is made up of points

$$\left\{(z_i^+, z_i^-) : l_{z_i} \le z_i^+ - z_i^- \le -lpl_i, (z_i^+, z_i^-) \ge 0\right\}, \tag{11}$$

i.e., the protection sense is “lower”. (10) and (11) define the feasible sets on the (z_i^+, z_i^-) space for the deviations of sensitive cells, depending on they are, respectively, upper or lower protected. The feasible set (10) is shown in Figure 1, considering two different cases: either $upl_i \le 0$ – Figure 1*(a)* – or $upl_i \ge 0$ – Figure 1*(b)*. Similarly, Figure 2 shows the feasible set (11) for the two cases $lpl_i \le 0$ – Figure 2*(a)* – and $lpl_i \ge 0$ – Figure 2*(b)*. Note that when $lpl_i = 0$ and $upl_i = 0$ both figures *(a)* and *(b)* of Figures 1 and 2 coincide.

From the objective function of (6), since $w_i > 0$, we have that in an optimal solution either $z_i^+ > 0$ or $z_i^- > 0$, but not both. Therefore, the optimal sets of Figures 1 and 2 are restricted to the thick segments on the axes. When lpl_i and upl_i are nonnegative, once y_i is fixed, the optimal sets are convex and we know which component will be zero in the optimal solution: $z_i^- = 0$ if $y_i = 1$ (Figure 1*(b)*), and $z_i^+ = 0$ if $y_i = 0$ (Figure 2*(b)*). Therefore we may write an alternative formulation for Ω^i when $lpl_i \ge 0$ and $upl_i \ge 0$:

$$\begin{aligned}\Omega_1^i = \big\{ (z_i^+, z_i^-, y_i) : {} & upl_i y_i \le z_i^+ \le u_{z_i} y_i, \\ & lpl_i(1-y_i) \le z_i^- \le -l_{z_i}(1-y_i), y_i \in \{0,1\}\big\} \; i \in \mathscr{S}. \end{aligned} \tag{12}$$

Note that constraints in Ω_1^i are equal to constraints (3e) of the standard CTA model. Next Proposition 1 shows that formulation (12) is stronger than (9). Moreover, denoting by $LR(\Omega)$ the linear relaxation of the set Ω (i.e., the set obtained by replacing conditions $y_i \in \{0,1\}$ in Ω by $0 \leq y_i \leq 1$ in $LR(\Omega)$), the proposition also shows that the linear relaxation of (12) is included in that of (9), and therefore any branch-and-bound based procedure is in theory more efficient with formulation Ω_1^i.

Proposition 1 *Given the two sets defined in (9) and (12), if $lpl_i \geq 0$ and $upl_i \geq 0$, then*

(i) $\Omega_1^i \subset \Omega^i$*, for all* $i \in \mathscr{S}$*;*

(ii) $LR(\Omega_1^i) \subset LR(\Omega^i)$*, for all* $i \in \mathscr{S}$*.*

Proof

(i) The proof is immediate just looking at Figures 1*(b)* and 2*(b)*.

(ii) We first show that $LR(\Omega_1^i) \subseteq LR(\Omega^i)$. From (12), any point (z_i^+, z_i^-, y_i) in $LR(\Omega_1^i)$ satisfies

$$\begin{aligned} upl_i y_i \leq \quad & z_i^+ && \leq u_{z_i} y_i, \\ l_{z_i}(1-y_i) \leq \quad & -z_i^- && \leq -lpl_i(1-y_i). \end{aligned}$$

Adding the two previous inequalities we obtain $upl_i y_i + l_{z_i}(1-y_i) \leq z_i^+ - z_i^- \leq u_{z_i} y_i - lpl_i(1-y_i)$, and thus, from (9), $(z_i^+, z_i^-, y_i) \in LR(\Omega^i)$. Finally we show that $LR(\Omega_1^i) \neq LR(\Omega^i)$ by noting that, for instance, when $y_i = 0$ points in $LR(\Omega_1^i)$ are of the form $(0, z_i^-, 0)$ (i.e., the thick line of Figure 2*(b)*), while points in $LR(\Omega^i)$ are of the form $(z_i^+, z_i^-, 0)$ (i.e., the shadowed region of Figure 2*(b)*). □

Similarly, the thick lines of Figure 3 show the subset of Ω^{0_i} in an optimal solution. Such a subset is nonconvex, and it can be improved by adding two new groups of constraints:

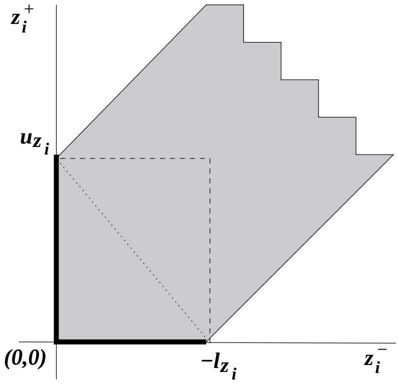


Figure 3: Strengthened formulations for Ω^{0_i}, represented by the shadowed region. Additional constraints are shown by the dashed and dotted lines.

- First, we may add upper bounds for z_i^+ and z_i^-. These are represented by the dashed line of Figure 3. The new set

$$\Omega_1^{0_i} = \left\{(z_i^+, z_i^-) : 0 \le z_i^+ \le u_{z_i}, 0 \le z_i^- \le -l_{z_i}\right\} \tag{13}$$

  is bounded, and there is no need to include the now redundant constraints $l_{z_i} \le z_i^+ - z_i^- \le u_{z_i}$, $i \in \mathscr{N}$. Note that (13) only imposes bounds on variables, but no constraint; this can significantly improve the performance of a solver.

- Second, looking at Figure 3 it is clear that the convex hull of points in the optimal set is the triangle of vertices $(0,0), (-l_{z_i}, 0), (0, u_{z_i})$. The convex hull is formulated by the new set

$$\Omega_2^{0_i} = \left\{(z_i^+, z_i^-) : z_i^+ \le u_{z_i} + \frac{u_{z_i}}{l_{z_i}} z_i^-, (z_i^+, z_i^-) \ge 0.\right\} \tag{14}$$

  The new constraint $z_i^+ \le u_{z_i} + \frac{u_{z_i}}{l_{z_i}} z_i^-$ corresponds to the dotted line of Figure 3. Although it reduces the feasible region, it complicates the formulation by adding an extra constraint for each cell $i \in \mathscr{N}$, which could significantly increase the computational time.

The following proposition 2 states the previous relations between sets Ω^{0i}, $\Omega_1^{0_i}$ and $\Omega_2^{0_i}$.

Proposition 2 *Given the sets Ω^{0i}, $\Omega_1^{0_i}$ and $\Omega_2^{0_i}$ respectively defined in (8), (13)and (14), then $\Omega_2^{0_i} \subset \Omega_1^{0_i} \subset \Omega^{0i}$.*

Proof The proof is immediate from Figure 3. □

4.1. Models considered

Combining the alternative formulations for Ω^i and Ω^{0_i} of previous section, (i.e., either Ω^i or Ω_1^i, and either Ω^{0i}, $\Omega_1^{0_i}$ or $\Omega_2^{0_i}$) in (6), it is possible to obtain different optimization models. We note that the alternative formulation Ω_1^i for Ω^i can only be used if $lpl_i \ge 0$ and $upl_i \ge 0$, whereas the alternative formulations $\Omega_1^{0_i}$ and $\Omega_2^{0_i}$ for Ω^{0_i} are always valid.

We have considered eight different models, which are tested in the computational results of Section 5. The objective function is the same for the eight models, and corresponds to that of (6); the models only differ in the representation of the feasible set. The first group of four models considers the formulation Ω^i for any $i \in \mathscr{S}$, independently of the sign of lpl_i and upl_i (i.e., even when $lpl_i \ge 0$ and $upl_i \ge 0$ formulation Ω^i is used). These four models will be denoted as the *new* models, and their feasible sets are respectively formulated as:

$$\begin{aligned}
\Omega_{new_1} &= \Omega^A \cap (\cap_{i\in\mathcal{N}}\Omega_1^{0_i}) \cap (\cap_{i\in\mathcal{S}}\Omega^i), & (15)\\
\Omega_{new_2} &= \Omega^A \cap \Big(\cap_{i\in\mathcal{N}}(\Omega_1^{0_i}\cap\Omega^{0_i})\Big) \cap (\cap_{i\in\mathcal{S}}\Omega^i), & (16)\\
\Omega_{new_3} &= \Omega^A \cap \Big(\cap_{i\in\mathcal{N}}(\Omega_1^{0_i}\cap\Omega_2^{0_i}\Big) \cap (\cap_{i\in\mathcal{S}}\Omega^i), & (17)\\
\Omega_{new_4} &= \Omega^A \cap \Big(\cap_{i\in\mathcal{N}}(\Omega^{0_i}\cap\Omega_1^{0_i}\cap\Omega_2^{0_i}\Big) \cap (\cap_{i\in\mathcal{S}}\Omega^i). & (18)
\end{aligned}$$

The second group of four models uses Ω^i for sensitive cells $i \in \mathcal{S}$ with either $upl_i < 0$ or $lpl_i < 0$, and Ω_1^i when $upl_i \geq 0$ and $lpl_i \geq 0$. They are thus a hybrid between the standard CTA model of (3) and the general model for negative protection levels of (6). They will be referred as the *hybrid* models. Making a partition of the set of sensitives cells $\mathcal{S} = \mathcal{S}^- \cup \mathcal{S}^+$, where $\mathcal{S}^- = \{i \in \mathcal{S} : lpl_i < 0 \text{ or } upl_i < 0\}$ and $\mathcal{S}^+ = \{i \in \mathcal{S} : lpl_i \geq 0 \text{ and } upl_i \geq 0\}$, the feasible sets of the four hybrid models are:

$$\begin{aligned}
\Omega_{hyb_1} &= \Omega^A \cap (\cap_{i\in\mathcal{N}}\Omega_1^{0_i}) \cap (\cap_{i\in\mathcal{S}^+}\Omega_1^i) \cap (\cap_{i\in\mathcal{S}^-}\Omega^i), & (19)\\
\Omega_{hyb_2} &= \Omega^A \cap \Big(\cap_{i\in\mathcal{N}}(\Omega_1^{0_i}\cap\Omega^{0_i})\Big) \cap (\cap_{i\in\mathcal{S}^+}\Omega_1^i) \cap (\cap_{i\in\mathcal{S}^-}\Omega^i), & (20)\\
\Omega_{hyb_3} &= \Omega^A \cap \Big(\cap_{i\in\mathcal{N}}(\Omega_1^{0_i}\cap\Omega_2^{0_i}\Big) \cap (\cap_{i\in\mathcal{S}^+}\Omega_1^i) \cap (\cap_{i\in\mathcal{S}^-}\Omega^i), & (21)\\
\Omega_{hyb_4} &= \Omega^A \cap \Big(\cap_{i\in\mathcal{N}}(\Omega^{0_i}\cap\Omega_1^{0_i}\cap\Omega_2^{0_i}\Big) \cap (\cap_{i\in\mathcal{S}^+}\Omega_1^i) \cap (\cap_{i\in\mathcal{S}^-}\Omega^i). & (22)
\end{aligned}$$

5. Computational results

The eight optimization models resulting from the feasible sets defined by (15)–(22) and the objective function of (6) have been implemented using the AMPL modelling system (Fourer, Gay and Kernighan (2002)). A set of real-world instances have been solved with the eight models, using both the MILP solvers of CPLEX 12.1 and Xpress Optimizer 19.00.00. The particular values of w in these real-world instances were specifically computed (i.e., they were neither 1, nor the cell value). All the runs have been performed on a Linux Dell Precision T5400 workstation with 16GB of memory and four Intel Xeon E5440 2.83 GHz processors, without exploitation of parallelism capabilities (to fairly compare CPLEX and Xpress solution times, since our CPLEX version allows multithreading whereas the Xpress version do not). A MILP optimality gap of 0 was set for all the executions. The MILP optimality gap is defined as

$$gap = \frac{|best - lb|}{1 + |best|} \cdot 100\%, \tag{23}$$

best being the best current solution, and lb the best current lower bound. A zero optimality gap is impractical with real-world instances as the ones considered in this work, since it provides prohibitively large executions. However, it was used to test the strength of each formulation.

Feasibility and integrality tolerances were also reduced for both solvers; they were set, respectively, to 10^{-8} and 0 for CPLEX and to 10^{-8} and 10^{-8} for Xpress (since it does not allow integrality tolerances smaller than the feasibility tolerance). Such a reduction is required to avoid solutions with underprotected cells. Indeed, (9) and (12) impose, among other constraints,

$$z_i^+ - z_i^- \leq -lpl_i(1-y_i) + u_{z_i}y_i, \qquad z_i^+ \leq u_{z_i}\, y_i.$$

In practical tables u_{z_i} and l_{z_i} may be very large, e.g., $u_{z_i} = l_{z_i} = M$. If, because of the feasibility and integrality tolerance, we get a solution $y_i = \epsilon$ instead of $y_i = 0$, then the above constraints would be

$$z_i^+ - z_i^- \leq -lpl_i(1-\epsilon) + M\epsilon \neq -lpl_i, \qquad z_i^+ \leq M\epsilon \neq 0.$$

Therefore, sensitive cell i would result underprotected. Decreasing the feasibility tolerance, we make the above ϵ value smaller, but the problem becomes much harder and the probability of the problem being reported as infeasible – when it is feasible – is increased. A better option is to avoid big M values for cell deviations, but this means the real cell bounds (lower and upper bounds) should be small. In this work we set a bound $M = 10^8$ for cell deviations (i.e., if the real bound is greater than M, then it is replaced by M; otherwise the real bound is used). However, even with such a bound on the deviations and with the above small feasibility and integrality tolerances, some solutions reported unprotected cells, as shown in below tables.

Table 2 shows the dimensions of the real-world instances considered, which were generated in Statistics Germany from data provided by Eurostat. Columns n, s, m and "N.coef" report, respectively, the number of cells, sensitive cells, linear relations of the table, and nonzero coefficients of matrix A. The nine instances can be grouped in small instances (the first three), medium size instances (the middle three), and large instances (the last three). The medium and large instances can be considered difficult since they

Table 2: *Dimensions of the test instances.*

Instance	n	s	m	N.coef
APS-Jan	87	5	35	177
APS-Feb	87	5	35	177
APS-Mar	87	5	35	177
sbs-E	1430	382	991	4680
sbs-C	4212	1135	2580	13806
dposrel	9568	1492	3956	22698
sbs-D$_a$	28288	7142	13360	87022
sbs-D$_b$	28288	7131	13360	87022
balofpay-eus-p1	39060	2483	37818	175965

Table 3: *Results for each model and solver (three smaller instances).*

Instance	CPLEX				Xpress			
model	CPU	f^*	B&B	n.u.	CPU	f^*	B&B	n.u.
APS-Jan								
new$_1$	0.004	7.12	0	0	0	7.12	1	0
new$_2$	0.004	7.12	0	0	0	7.12	1	0
new$_3$	0.004	7.12	0	0	0	7.12	1	0
new$_4$	0	7.12	0	0	0	7.12	1	0
hyb$_1$	0	7.12	0	0	0	7.12	1	0
hyb$_2$	0.004	7.12	0	0	0	7.12	1	0
hyb$_3$	0.004	7.12	0	0	0	7.12	1	0
hyb$_4$	0.004	7.12	0	0	0	7.12	1	0
APS-Feb								
new$_1$	0.008	66.85	6	0	0	66.85	15	0
new$_2$	0.008	66.85	6	0	0	66.85	15	0
new$_3$	0.004	66.85	11	0	0	66.85	15	0
new$_4$	0.004	66.85	11	0	0	66.85	15	0
hyb$_1$	0.008	66.85	6	0	0	66.85	3	0
hyb$_2$	0.004	66.85	6	0	0	66.85	3	0
hyb$_3$	0.004	66.85	7	0	0	66.85	3	0
hyb$_4$	0.004	66.85	7	0	0	66.85	3	0
APS-Mar								
new$_1$	0.008	11.90	1	0	0	11.90	1	0
new$_2$	0.004	11.90	1	0	0	11.90	1	0
new$_3$	0.004	11.90	3	0	0	11.90	1	0
new$_4$	0.004	11.90	3	0	0	11.90	1	0
hyb$_1$	0.004	11.90	0	0	0	11.90	1	0
hyb$_2$	0.004	11.90	0	0	0	11.90	1	0
hyb$_3$	0.004	11.90	0	0	0	11.90	1	0
hyb$_4$	0	11.90	0	0	0	11.90	1	0

have a complex structure, and a significant number of cells, constraints and sensitive cells. These nine instances are related to data from structural business statistics, balance of payment, and animal production statistics of the European Union.

The results for each model and solver, for each group of three instances, i.e., small, medium size and large, are respectively reported in Tables 3–5. Columns “CPU”, f^*, “B&B” and “n.u.” provide, respectively, the CPU solution time, best objective function reached, number of branch-and-bound nodes explored, and number of underprotected cells in the solution. A time limit of 7200 seconds was set in all the executions. When this time limit is reached, the CPU time column shows the optimality gap (23) of the solution obtained within the time limit. We provide results with both CPLEX and Xpress since they are the two solvers mainly used in the statistical disclosure control community.

Table 4: *Results for each model and solver (three medium size instances).*

Instance	CPLEX				Xpress			
model	CPU	f^*	B&B	n.u.	CPU	f^*	B&B	n.u.
sbs-E								
new_1	42.86	107442.27	7406	0		[1]		
new_2		[1]				[1]		
new_3	364.71	107720.37	107090	0		[1]		
new_4	167.49	107439.65	37401	0		[1]		
hyb_1	14.26	107442.27	1056	0		[1]		
hyb_2	12.73	107439.65	1086	0		[1]		
hyb_3	10.36	107853.98	770	0	[2] (1.05%)	121084.67	3014871	0
hyb_4	9.48	107853.26	885	0		[1]		
sbs-C								
new_1	[2] (0.07%)	313562.69	305971	0		[1]		
new_2	[2] (0.07%)	313655.95	213097	0		[1]		
new_3	[2] (46%)	314547.38	161825	0		[1]		
new_4	[2] (1.3%)	313742.96	192901	0		[1]		
hyb_1	58.70	331425.16	525	0		[1]		
hyb_2	52.69	315160.90	518	0		[1]		
hyb_3	904.37	324572.49	103510	0		[1]		
hyb_4	[2] (0.004%)	314001.24	1301687	0		[1]		
dposrel								
new_1	10.2	7807.98	1533	62	8	7808.28	961	0
new_2	9.9	7807.98	1422	62	10	7808.28	915	0
new_3	18.0	7807.98	1723	62	8	7813.72	517	0
new_4	18.8	7807.99	1943	63	8	7813.72	517	0
hyb_1	8.9	7808.28	1231	1	6	7808.28	299	0
hyb_2	8.5	7808.28	1238	1	5	7808.29	361	0
hyb_3	13.6	7808.28	1939	1	6	7813.72	311	1
hyb_4	13.7	7808.28	2047	1	6	7813.72	311	1

[1] No feasible solution found, problem reported as infeasible

[2] Time limit reached

However, our purpose is not to compare the two different solvers, but the models and to show the difficulties found by the optimization solvers. From Tables 3–5 the following observations can be made:

- Both CPLEX and Xpress, with the eight different models, successfully solved the very small instances of Table 3 in less than 1 second, exploring very few branch-and-bound nodes.
- The medium size and large instances of Tables 4–5 are difficult for state-of-the-art solvers. For some instances and models, CPLEX and Xpress were not able to

find either an optimal solution (executions marked with a [2] in Tables 4–5), or a feasible solution within the 7200 seconds time limit (executions marked with a [4] in Table 5). In some CPLEX executions the optimization process even failed by numerical errors of the solver (runs marked with a [3] in Table 5).

- For some combinations instance–model the optimization problems are reported as infeasible (when they are feasible) due to the small feasibility tolerances used. These executions are marked with a [1] in Tables 4–5. However, if the feasibility

Table 5: *Results for each model and solver (three larger instances).*

Instance	CPLEX				Xpress			
model	CPU	f^*	B&B nodes	n.u.	CPU	f^*	B&B nodes	n.u.
sbs-D$_a$								
new$_1$	[2](20%)	414666.45	26096	0		[1]		
new$_2$	[2](22%)	417332.53	20699	0		[1]		
new$_3$		[3]				[1]		
new$_4$	[2](33%)	417841.08	22207	0		[1]		
hyb$_1$		[1]				[1]		
hyb$_2$		[1]				[1]		
hyb$_3$		[4]				[1]		
hyb$_4$		[4]				[1]		
sbs-D$_b$								
new$_1$	[2](22%)	408432.48	29318	0		[1]		
new$_2$	[2](56%)	767929.98	16906	0		[1]		
new$_3$	[2](31%)	416436.74	19107	0		[1]		
new$_4$		[3]				[1]		
hyb$_1$		[4]				[1]		
hyb$_2$		[1]				[1]		
hyb$_3$		[4]				[1]		
hyb$_4$		[4]				[1]		
balofpay-eus-p1								
new$_1$		[3]			[2](88%)	5366.63	6407	0
new$_2$		[3]			[2](88%)	5366.63	6507	0
new$_3$		[3]			[2](88%)	7300.04	5351	0
new$_4$		[3]			[2](88%)	7300.04	5281	0
hyb$_1$		[3]			[2](54%)	4708.11	5727	0
hyb$_2$		[3]			[2](56%)	4554.76	9690	0
hyb$_3$		[3]			[2](55%)	5303.8	1672	0
hyb$_4$		[3]				[1]		

[1] No feasible solution found, problem reported as infeasible

[2] Time limit reached

[3] Unrecoverable failure: singular basis

[4] Time limit reached with no integer solution

tolerance is increased, then we obtain bad solutions, with a significant number of underprotected cells. This undesirable effect due to large feasibility tolerances even happens for small instances; for instance, four out of the five sensitive cells of Table 3 would have been underprotected in the optimal solution if a feasibility tolerance of 10^{-5} had been used. Even with the tight feasibility tolerances considered, we see that executions of instance “dposrel” of Table 4 provided 63 underprotected cells for the *new* models; this value was reduced to one underprotected cell when the *hybrid* model was used.

- The additional constraints (14) in models new_3, new_4, hyb_3 and hyb_4 may significantly increase the solution time. For instance, model new_1 of instance “sbs-E” with CPLEX takes 42.86 seconds, while models new_3 and new_4 take 364.71 and 167.49 seconds; similarly, for CPLEX and instance “dposrel”, models new_3 and new_4, and hyb_3 and hyb_4, require a 100% and a 50% more time than models new_1 and new_2, and hyb_1 and hyb_2, respectively. However, as suggested by Proposition 2 the number of branch-and-bound nodes may be reduced: this is observed in *new* models of instance “dposrel” with Xpress, and *hybrid* models of instance “sbs-E” with CPLEX, both of Table 4. Therefore, constraints (14) could be of help in some situations.
- In general, the *hybrid* model is preferred, since it is more efficient. This is consistent with Proposition 1. For instance, in Table 4 for “sbs-E” and CPLEX, the four executions with the *hybrid* models are much faster than with the *new* variants. This is also observed in instance “balofpay-eus-p1” and Xpress, where the *hybrid* models provided better solutions than the *new* models within the time limit. However, in some cases, when the *hybrid* models have difficulties, the *new* ones can be an alternative, as shown for instance sbs-D_a and CPLEX in Table 5.

6. Conclusions

From the computational and theoretical results with the several models tested, it can be concluded that the *hybrid* approach is in general more efficient than the *new* models for the solution of CTA instances with either positive or negative protection levels. It has also been shown that both types of models might have difficulties when exposed to real-world and complex CTA instances, even using the best today optimization solvers. This motivates further development on optimization methods for difficult CTA instances. Some steps have been done along these lines using, e.g., cutting plane or Benders decomposition approaches (Castro and Baena (2008)), and heuristic block coordinate decompositions (González and Castro (2009)). However, there is not yet a definitive approach for any CTA instance. This is part of the further work to be done in the statistical disclosure control field.

7. Acknowledgments

This work has been supported by grants MICINN MTM2009-08747 and SGR-2009-1122, and by Eurostat framework contract 22100.2006.002-226.532.

References

Castro, J. (2005). Quadratic interior-point methods in statistical disclosure control. *Computational Management Science*, 2(2), 107–121.

Castro, J. (2006). Minimum-distance controlled perturbation methods for large-scale tabular data protection. *European Journal of Operational Research*, 171, 39–52.

Castro, J. (2007a). A shortest-paths heuristic for statistical data protection in positive tables. *INFORMS Journal on Computing*, 19(4), 520–533.

Castro, J. (2007b). An interior-point approach for primal block-angular problems. *Computational Optimization and Applications*, 36, 195–219.

Castro, J. and Baena, D. (2008). Using a mathematical programming modeling language for optimal CTA. *Lecture Notes in Computer Science*, 5262, 1–12.

Castro, J. and Giessing, S. (2006). Testing variants of minimum distance controlled tabular adjustment. In *Monographs of Official Statistics. Work session on Statistical Data Confidentiality*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 2006, 333–343. ISBN 92-79-01108-1.

Castro, J., González, J. A. and Baena, D. (2009). User's and programmer's manual of the RCTA package. Technical Report DR 2009/01, Dept. of Statistics and Operations Research, Universitat Politècnica de Catalunya.

Cox, L. H., Kelly, J. P. and Patil, R. (2004). Balancing quality and confidentiality for multivariate tabular data. *Lecture Notes in Computer Science*, 3050, 87–98.

Dandekar, R. A., and Cox, L. H. (2002). Synthetic tabular data: an alternative to complementary cell suppression. Manuscript, Energy Information Administration, U.S. Department of Energy.

Domingo-Ferrer, J. and Franconi, L. (eds.) (2006). *Lecture Notes in Computer Science. Privacy in Statistical Databases*, 4302, Springer, Berlin.

Domingo-Ferrer, J. and Saigin, Y. (eds.) (2008). *Lecture Notes in Computer Science. Privacy in Statistical Databases*, 5262, Springer, Berlin.

Fair Isaac Corporation Dash Xpress (2008). *Xpress Optimizer. Reference Manual, release 19.0.0*.

Fourer, R., Gay, D. M. and Kernighan, D. W. (2002). *AMPL: A Modeling Language for Mathematical Programming*, Duxbury Press.

Giessing, S., Hundepool, A. and Castro, J. (2009). Rounding methods for protecting EU-aggregates. In *Worksession on statistical data confidentiality. Eurostat methodologies and working papers*, Eurostat-Office for Official Publications of the European Communities, Luxembourg, 255–264.

González, J. A. and Castro, J. (2011). A heuristic block coordinate descent approach for controlled tabular adjustment. *Computers & Operations Research*, 38, 1826–1835.

IBM ILOG CPLEX (2009), *User's Manual for CPLEX v12.1*.

Salazar-González, J. J. (2006). Controlled rounding and cell perturbation: statistical disclosure limitation methods for tabular data. *Mathematical Programming*, 105, 583–603.

Willenborg, L. and de Waal, T. (2000). *Lecture Notes in Statistics. Elements of Statistical Disclosure Control*, 155, Springer, New York.

SORT 35 (1) January-June 2011, 21-38

The choice of type of input-output table revisited: moving towards the use of supply-use tables in impact analysis

José M. Rueda-Cantuche[1,*]

European Commission, Joint Research Centre-IPTS and Pablo de Olavide University at Seville

Abstract

The construction of symmetric input-output tables (SIOTs) is a controversial issue as regards the choice of model to construct both product-by-product and industry-by-industry SIOTs, especially the former ones. However, there has been little attention paid so far by the UN and the Eurostat Systems of National Accounts on the choice of type of SIOT to carry out impact analyses let alone other input-output applications. Concerning the price and quantity models in input-output analysis, this paper identifies severe problems in the correct interpretation of the meaning of their results and proposes the use of supply and use tables instead of SIOTs to solve these problems.

MSC: 91B82, 91B66

Keywords: Statistical methods, economic indices and measures, multisectoral models.

1. Introduction

Typical research questions that can be addressed by input-output analysis are as follows. What is the impact on employment of an increase in households' consumption of renewable energies? Or what would be the effect on fuel prices of an increase in the labour costs of the electricity industry? Many input-output practitioners would claim that they could easily answer these questions as long as they could dispose of the so-called symmetric input-output tables (SIOTs). However, very few authors reflect on the issue

[*] The views expressed in this paper are those of the authors and should not be attributed to the European Commission or its services. This work is supported by the EU funded FP7 Project: "World Input-Output Database: Construction and Applications", Grant agreement no.: 225281

[1] European Commission, Joint Research Centre, Institute for Prospective Technological Studies (IPTS) and Pablo de Olavide University at Seville. Phone: +34 954488363. E-mail: Jose.Rueda-Cantuche@ec.europa.eu

Received: July 2010

Accepted: March 2011

that in both examples the impact drivers and the resulting effects are referring to different issues. On the one hand, households may increase their consumption of bio-fuels (of a single product or group of products) while the impacts actually refer to the number of jobs created in a certain industry. On the other hand, labour costs have increased in the electricity industry while the price effects should refer to a single product (e.g. fossil fuel). Thus, we believe that the main unnoticed shortcoming underlying the use of SIOTs to address these types of research questions is precisely its symmetry, in the sense that they are defined either on a product-by-product or on an industry-by-industry basis. Moreover, although the choice of type of SIOT is playing increasingly a relevant role in the most recent systems of national accounts, they still provide unclear guidelines on the type of table to be used for what type of analysis. There is no clear structure or even clear recommendations. As it will be shown in this paper, the so-called supply and use tables solve efficiently this matter since they are defined on a product-by-industry basis rather than on a product or on an industry basis only. Therefore, we will eventually recommend exploring new possibilities in order to find suitable supply-use based input-output techniques to give a proper answer to the type of questions raised at the beginning of this paragraph.

Accordingly, Section 2 will introduce the input-output framework; the next section will address how the issue of the choice of type of SIOT is insufficiently dealt with by the United Nations and European Systems of National Accounts. Section 4 reviews the most relevant input-output applications, namely the quantity and the price models; and the last Section concludes with some recommendations on the benefits of using supply and use tables rather than SIOTs in impact input-output analysis.

2. The input-output framework

Following Rueda-Cantuche *et al.* (2009), an input-output framework revolves around the so-called supply and use tables. They can be seen as the output mix of industries and the industries' use of inputs, respectively. On the one hand, the supply table comprises an intermediate matrix of goods and services (rows) produced by industries (columns), plus additional column vectors including imports, distribution margins (trade and transport) and net taxes on products, all of which make the total supply of products of an economy. On the other hand, the use table represents domestically produced and imported intermediate and final uses. They may be valued at basic and at purchasers' prices. There are several additional column vectors that show the usual final demand categories, i.e. final consumption, investment and exports; and additional rows, which eventually represent the different components of the gross value added, e.g. labour costs, capital use, other net taxes on production and net operating surplus (see Tables 1 and 2).

Note that the valuation of the supply and use tables is not coincident. The supply table is measured at basic prices, which means excluding trade and transport margins and net taxes on products. To the contrary, the use table is measured at purchasers' prices,

Table 1: *Simplified overview of a supply table (Rueda-Cantuche* et al.*, 2009).*

	INDUSTRIES (NACE)	OUTPUT OF INDUSTRIES (NACE)							IMPORTS				VALUATION		
	PRODUCTS (CPA)	Agriculture	Industry	Construction	Trade	Private services	Government services	Total	Intra EU imports cif	Extra EU imports cif	Imports cif	Total supply at basic prices	Trade and transport margins	Taxes less subsidies on products	Total supply at purcha-sers' prices
No		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Products of agriculture	Production matrix (V^T)							Imports cif				Valuation items		
2	Products of industry														
3	Construction work														
4	Trade														
5	Private services														
6	Government services														
7	Total														
8	Cif/ fob adjustments on imports														
9	Direct purchases abroad by residents														
10	Output at basic prices	Total output of industries at basc prices							Total imports				Total		

Table 2: *Simplified overview of a use table (Rueda-Cantuche* et al.*, 2009).*

	INDUSTRIES (NACE)	OUTPUT OF INDUSTRIES (NACE)							FINAL USES									
	PRODUCTS (CPA)	Agriculture	Industry	Construction	Trade	Private services	Government services	Total	Final consumption expenditure by households	Final consumption expenditure by non-profit organisations	Final consumption expenditure by government	Gross fixed capital formation	Changes in valuables	Changes in inventories	Exports intra EU fob	Exports extra EU fob	Total	Total use at purchasers' prices
No		1	2	3	4	5	6	7	7	8	9	10	11	12	13	14	15	16
1	Products of agriculture	Intermediate uses (U)							Final uses (Y)									
2	Products of industry																	
3	Construction work																	
4	Trade																	
5	Private services																	
6	Government services																	
7	Total	Total intermediate consumption							Total final uses of goods and services									
8	Cif/ fob adjustments on exports																	
9	Direct purchases abroad by residents																	
10	Domestic purchases. by non-residents																	
11	Total	Total intermediates							Total final uses									
12	Compensation of employees	Value added (W)																
13	Other net taxes on production																	
14	Consumption of fixed capital																	
15	Operating surplus, net																	
16	Total	Total value added at basic prices																
17	Output at basic prices	Total output of industries at basic prices																

which means at the price producers and/or consumers pay goods and services for final use or intermediate inputs (including trade and transport margins and taxes less subsidies on products). As stated by Eurostat (2008), basic prices are the preferable valuation concept in the supply and use framework in the sense that it provides a more homogeneous valuation. Thus, for analytical purposes a valuation as much homogeneous as possible will be required as the input-output relations are to be interpreted as technical coefficients.

The construction of SIOTs has suffered from controversial contributions in the literature. On the one hand, a product-by-product table describes the technological relations between products (Eurostat, 2008). The intermediate matrix describes a kind of recipe of how to produce commodities in terms of the amounts used of others, irrespective of the producing industry. On the other hand, industry-by-industry tables depict inter-industry relations. The intermediate matrix would describe on an industry basis, the use of commodities of the other industries (Eurostat, 2008).

Independently of the purpose of the analysis, both types of SIOTs have their own advantages and disadvantages. On the one hand, the product-by-product tables are more homogeneous in their description of the transactions being one of the most commonly used tables in input-output analysis (productivity, comparison of costs structures, employment effects, energy policy...) and have a clear input structure in terms of products for intermediate uses and value added for the compensation of labour and capital for homogenous branches. However, product-by-product tables require labour intensive compilation tasks; they must be based on analytical assumptions that take final results away from actual market transactions and observations, and hence, they make more difficult the integration of other statistical sources and the reporting on the transformation procedure. On the other hand, industry-by-industry tables are much closer to statistical sources; they allow for an easier comparability with other statistical databases; they are less labour intensive to compile, being based on pragmatic assumptions rather than on analytical hypotheses. Nevertheless, the larger the secondary activities in the supply table are the more difficult it becomes to identify homogeneous cost structures in an industry-by-industry table.

In practice, most of the countries worldwide compile product-by-product tables although there are some hardly negligible countries like Denmark, the Netherlands, Norway, Canada and Finland that compile industry-by-industry SIOTs. Nevertheless, one can always shift from one type to another as it is shown in Table 3.

Basically, the choice of the type of SIOT is related to the treatment of secondary products (Rueda-Cantuche and ten Raa, 2009). There are two main approaches to eliminating secondary production from industries in order to get homogenous branches of production in a product-by-product SIOT. Both of them can be derived from combining the information on input structures depicted by the use table at basic prices with the supply table so that all the secondary production (including the inputs used to produce them) are re-allocated either to the industry for which the product is a primary output (product technology, Model A) or to the main product of the industry that actually produces it (industry technology, Model B). The transformed use table is what is referred to as an input-output table (UN, 2009, par. 28.47). It follows that in deriving a product-by-product matrix in the simplest possible way, the final demand of the use table remains unchanged. By contrast, the demand for intermediate uses and labour and capital inputs are determined by the nature of the products made (UN, 2009, par. 28.48).

There are other possible technology assumptions available in the literature, that were reviewed by Viet (1994) and by ten Raa and Rueda-Cantuche (2003), who also provided their advantages and disadvantages from a theoretical approach (see also Kop Jansen and ten Raa, 1990). For more details, the interested reader could check the above references.

In deriving an industry-by-industry SIOT in the simplest way, the key issue is reallocating items between rows rather than between columns (as in product-by-product SIOTs). Contrarily to the product-by-product SIOTs, final uses will have to change thus indicating now the intermediate and final demand associated to the industry supplying the products rather than to the products themselves. Recall that the use tables have industries in columns and products in rows and we aim to construct a SIOT with industries both in rows and columns. Concerning the value added components, they remain unchanged because the level of the industry outputs will not be altered by the methods used for the construction of the SIOT.

It is assumed that as the level of the product output changes into that of the industry output, the pattern of sales will however remain the same. This is called a sales structure approach and only two approaches may be identified: the fixed industry sales structure assumption (Model C), where the industry deliveries are independent of the products delivered, and the fixed product sales structure (Model D), where they are instead independent of the producing industry. Rueda-Cantuche and ten Raa (2009) identified Model C as the most suitable from an axiomatic point of view.

For reading Table 3, let us define a use matrix, $\mathbf{U} = (u_{ij})\ i, j = 1, \ldots, n$ of products i consumed by industry j, and a supply matrix $\mathbf{V}^{\mathsf{T}} = (v_{ij})\ i, j = 1, \ldots, n$ where product i is produced by industry j, which is actually the transposition of the so-called make matrix $\mathbf{V}$. Models A, B, C and D can be easily formalized on the basis of supply and use matrices as it is shown in Table 3, where we provide bridges matrices that can be used to shift from one model to another. The matrices in the main diagonal refer to the mathematical expressions of the technical coefficient matrices of each model. Eventually, SIOTs can be calculated by post-multiplying the $\mathbf{A}$ matrices depicted in Table 3 with a diagonalised matrix of product outputs (for Models A and B) or of industry outputs (for Models C and D). Simple matrix algebra can be used by the reader to trace proofs.

Following the Eurostat Manual's (2008, p.349) notation and denoting as ˆ the diagonalization whether by suppression of the off-diagonal elements of a square matrix or by placement of the elements of a vector; we have denoted $\mathbf{g}$ as the column vector of industry output; $\mathbf{q}$ as the column vector of product output; $\mathbf{C} = \mathbf{V}^{\mathsf{T}}\hat{\mathbf{g}}^{-1}$ as the product-mix matrix with share of each product in industry outputs (supply table); $\mathbf{D} = \mathbf{V}\hat{\mathbf{q}}^{-1}$ as the market shares matrix with contribution of each industry to the product output (supply table); and $\mathbf{Z} = \mathbf{U}\hat{\mathbf{g}}^{-1}$ as the inputs requirements for products per unit of output of an industry (use table).

Product-by-product SIOTs (mainly using Model A or a slightly modified version) are the most common type of SIOT compiled by many European Union countries. Furthermore, Model B implies a mix of input structures that makes the use of product-by-product SIOTs inconsistent with technically oriented input-output analysis. Some European Union countries compile industry-by-industry SIOTs. They usually apply Model D (fixed product sales structure) for the transformation of supply and use tables into input-output tables. Model D is clearly preferred, due to the unrealistic feature of the alternative assumption of fixed industry sales structure.

3. The choice of type of input-output table in the UN and European systems of national accounts

The choice of technology assumption in the construction of product-by-product SIOTs has played a relevant role in the various systems of national accounts and handbooks/manuals published by the United Nations (UN) and Eurostat. To the contrary, the choice of type of SIOTs (product-by-product or industry-by-industry) has been almost fully neglected. In this section, we will explore the treatment of this issue by the two latest systems of national accounts published by the UN and Eurostat together with their respective handbooks or manuals.

3.1. SNA93, UN Handbook of IO Compilation (1999) and SNA08

Essentially, the SNA93 (UN, 1993) states that only product-by-product tables will be described in detailed since they are often proved as most useful (par. 15.150) but however the SNA93 does not provide any justification for this assortment and simply ignores industry-by-industry tables.

It was not until the publication of the UN Handbook of Input-Output Compilation and Analysis (UN, 1999) when industry-by-industry tables received a more detailed treatment, although still not too far reaching. After providing the definitions of product and industry SIOTs (par. 4.41), the UN Handbook asserts that industry-by-industry SIOTs are much less useful than product-by-product SIOTs because an industry might represent a group of establishments, part of which may be artificially created by mathematical methods (e.g. extrapolation) and therefore, does not reflect any “realistic” picture of the economy. Concerning IO modeling, the UN Handbook (par. 4.60) also states that industry-by-industry tables are of almost no interest to analysts since final demand is, rarely, in terms of industry outputs.

With an increasing interest for industry-by-industry SIOTs, the new System of National Accounts-SNA08 (UN, 2009) now includes one section specifically for these

Table 3: *Bridge matrices for technical coefficients to switch between different types of SIOTs.*

To: From:	MODEL A Product-by-product Product technology based	MODEL B Product-by-product Industry technology based	MODEL C Industry-by-industry Fixed industry sales structure	MODEL D Industry-by-industry Fixed product sales structure
Model A	$\mathbf{A_A\,(U, V) = Z\,C^{-1}}$	$\mathbf{A_B\,(U, V) = A_A\,CD}$	$\mathbf{A_C\,(U, V) = C^{-1}\,A_A\,C}$	$\mathbf{A_D\,(U, V) = DA_A\,C}$
Model B	$\mathbf{A_A\,(U, V) = A_B\,D^{-1}\,C^{-1}}$	$\mathbf{A_B\,(U, V) = Z\,D}$	$\mathbf{A_C\,(U, V) = C^{-1}\,A_B\,D^{-1}}$	$\mathbf{A_D\,(U, V) = DA_B\,D^{-1}}$
Model C	$\mathbf{A_A\,(U, V) = CA_C\,C^{-1}}$	$\mathbf{A_B\,(U, V) = CA_C\,D}$	$\mathbf{A_C\,(U, V) = C^{-1}\,Z}$	$\mathbf{A_D\,(U, V) = DCA_C}$
Model D	$\mathbf{A_A\,(U, V) = D^{-1}\,A_D\,C^{-1}}$	$\mathbf{A_B\,(U, V) = D^{-1}\,A_D\,D}$	$\mathbf{A_C\,(U, V) = C^{-1}\,D^{-1}\,A_D}$	$\mathbf{A_D\,(U, V) = D\,Z}$

Legend
A = Technical coefficients matrix
$\mathbf{V}^{T}$ = Supply matrix
U = Use matrix
e = Column vector of ones
Z = Inputs requirements for products per unit of output of an industry (use table)
C = Product-mix matrix with share of each product in output of an industry (supply table)
D = Market shares matrix with contribution of each industry to the output of a product (supply table)

kinds of tables (pars. 28.57 to 28.63). As to the choice of type of SIOTs, the SNA08 states that both product-by-product and industry-by-industry SIOTs serve different analytical functions (price consistency, labour market, technology, inter-industry relations...). It is also interesting to remark that in one of the annexes (par. A4.21), the SNA08 recognizes a change of emphasis from product-by-product SIOTs to industry-by-industry ones.

3.2. ESA95, the Eurostat Manual of Supply, Use and IO Tables (2008) and the ESA08 (draft version)

Unfortunately, to the knowledge of the authors, neither the ESA95 nor the draft version of the European System of Accounts – ESA08 (Eurostat, 2009) – mentions explicitly the issue of the choice of type of SIOTs. The ESA95 just offers a flexible approach to compile industry-by-industry SIOTs or product-by-product SIOTs according to the objective of economic analysis. As in the SNA93, it is recommended to compile the latter tables although industry-by-industry tables are also accepted if the industries are close to homogenous units of production (Eurostat, 2008; p.31). Nevertheless, the Eurostat Manual (2008) considerably deals with this issue in its chapter 11.

Following the Eurostat Manual (2008, p. 301), (...) "*product-by-product input-output tables are theoretically more homogeneous in their description of the transactions than industry-by-industry tables, since a single element of the latter can refer to products that are characteristic in other industries. This supports the assumption that in practice product-by-product tables generally are better suited for many types of input-output analysis. For example, it seems more feasible to use product-by-product input-output tables for productivity analysis or the analysis of new technologies in the economy. On the other hand, industry-by-industry input-output tables are possibly the better option if the economic impact of a major tax reform is studied on the basis of input-output data* (...)". Similarly to the UN Systems of National Accounts (SNA93 and SNA08) , there is also here only a general remark on the suitability of the type of SIOT, which cannot be considered as a clear guidance on which types of tables are to be used for what type of analysis.

Broadly speaking, very little secondary output reported in the supply table would lead to fade away the distinction between products and industries. So, a relatively low level of secondary activities reported in the European Union supply tables may well suggest, as one can read in the Eurostat Manual (2008, p. 309), that the difference between product-by-product SIOTs and industry-by-industry SIOTs is relatively small, and consequently both transformations can be regarded as valid options for impact analysis. However, (...) "*it must be noticed that secondary activities vary considerably across sectors even the general level is low* (...)" (Eurostat, 2008; p. 309).

The Eurostat Manual (2008, p. 340) eventually argues that "*the type of tables that best fulfils the standard quality criteria is the industry-by-industry table based on the assumption of fixed product sales structures and the product-by-product SIOT based* *on the product technology assumption. These types of tables reflect the accumulated experience and current practice of those countries most permanently involved in the compilation of SIOTs*". Focusing on the these two models (Models A and D) to construct product-by-product tables and industry-by-industry tables, respectively, the Eurostat Manual defines a set of quality features of both types of SIOTs (p. 340-341):

Transparency

Industry-by-industry SIOTs provide more transparency than product-by-product SIOTs because the fixed product sales structure assumption can be derived from the supply and use tables without too much effort and in such a way that negatives do not appear. Conversely, the product technology assumption is usually applied in a complex context requiring a balancing procedure to treat the negative elements that may arise and thus, causing less transparency.

Comparability

Industry-by-industry SIOTs guarantee more comparability with national accounts data since they are closer to statistical sources, survey results and actual observations. To the contrary, product-by-product tables have been compiled in an analytical step which creates less comparability with the sources but at the same time guarantees more comparability across nations.

Inputs

Product-by-product SIOTs have a clear input structure in terms of products for intermediate use and value added for the compensation of labour and capital for homogenous branches. However, in industry-by-industry SIOTs, mixed bundles of goods and services rather than homogeneous products are reported for intermediate and final uses.

Resources and timeliness

The compilation of product-by-product tables based on the product technology assumption requires more resources and balancing efforts due to the treatment of the negatives that may appear. Consequently, publication may be delayed. However, industry-by-industry tables can be directly derived from supply and use tables with less resource intensive efforts.

Analytical potential

The Eurostat Manual (2008, p. 341) states that "*industry-by-industry tables are well suited for specific analytical purposes which are related to industries (tax reform, impact analysis, fiscal policy, monetary policy, etc.)*" while product-by-product tables "*are well suited for many other specific analytical purposes which are related to* *homogeneous production units (productivity, comparison of cost structures, employment effects, energy policy, environmental policy, etc.)*" Although useful, this distinction just enumerates possible applications without a clear guidance on which types of tables are to be used for what type of analysis, which will hopefully be provided by this deliverable.

To cut a long story short, the choice of type of SIOT is not a relevant issue in the two most recent ESAs (1995 and 2008) although the Eurostat Manual (2008) gives much more insight into the matter than any of the UN documents. However, we still think that a deeper and clearer connection between standard input-output applications and the use of product-by-product and/or industry-by-industry tables is needed.

4. The relevance of the applications: the quantity and the price models in input-output analysis

4.1. The quantity and price models in input-output analysis

The main purpose of this section is to present briefly the theoretical background of the two most commonly and broadly used models in input-output analysis, i.e. the quantity and the price models. It will follow a discussion on the choice of type of SIOT for each type of model together with some guidelines.

Dietzenbacher (1997) considered the following SIOT in money terms (say, euros) for period 0:

$\mathbf{X}_0$	$\mathbf{f}_0$	$\mathbf{x}_0$
$\mathbf{v}_0^\mathsf{T}$	—	$\mathbf{v}_0^\mathsf{T}\mathbf{e}$
$\mathbf{x}_0^\mathsf{T}$	$\mathbf{e}^\mathsf{T}\mathbf{f}_0$	

$\mathbf{X}_0$ is the $n \times n$ matrix of intermediate uses; its typical element x_{ij}^0 denotes the value (in euros) of the deliveries from industry (product) i to industry (product) j, which will depend on the type of SIOT used. Dietzenbacher (1997) did not however distinguish in his paper between the two types of SIOTs referring implicitly all the time to industry-by-industry tables. The column vector $\mathbf{f}_0$ can be interpreted as sectoral (product) final demands including private and government consumption, investments and net exports[1]. The row vector $\mathbf{v}_0^\mathsf{T}$ gives the value added in each industry (product or homogenous branch), containing, for instance, payments for the labour and capital primary factors. The value of each industry (product) output is given by the elements of the vector $\mathbf{x}_0$

[1] Dietzenbacher (1997) made this assumption without loss of generality and for the sake of notational convenience.

while **e** denotes the *n*-dimension column vector of ones. Column-wise, a SIOT depicts input structures and row-wise, output structures. Since the total value of outputs equals the total value of inputs, for each industry (product), the following sets of accounting equations are obtained:

$$\mathbf{x}_0 = \mathbf{X}_0\,\mathbf{e} + \mathbf{f}_0 \tag{1}$$

$$\mathbf{x}_0^{\mathsf{T}} = \mathbf{e}^{\mathsf{T}}\,\mathbf{X}_0 + \mathbf{v}_0^{\mathsf{T}} \tag{2}$$

It follows that the input coefficients are defined as the industry (product) *i*'s input into industry (product) *j* as a fraction of the purchaser's output (x_j^0). They are obtained as $a_{ij}^0 = x_{ij}^0/x_j^0$, or in matrix terms, as $\mathbf{A}_0 = \mathbf{X}_0\,\hat{\mathbf{x}}_0^{-1}$ where $\hat{\mathbf{x}}_0$ denotes a diagonal matrix. Then, equation (1) may be written as:

$$\mathbf{x}_0 = \mathbf{A}_0\,\mathbf{x}_0 + \mathbf{f}_0 \tag{3}$$

In a similar way, the output coefficients denote the industry (product) *i*'s delivery to industry (product) *j* as a fraction of the seller's output (x_i^0). They are obtained as $b_{ij}^0 = x_{ij}^0/x_i^0$ or, in matrix terms, as $\mathbf{B}_0 = \hat{\mathbf{x}}_0^{-1}\,\mathbf{X}_0$. Subsequently, equation (2) may be rewritten as

$$\mathbf{x}_0^{\mathsf{T}} = \mathbf{x}_0^{\mathsf{T}}\,\mathbf{B}_0 + \mathbf{v}_0^{\mathsf{T}} \tag{4}$$

From the accounting equations (3) and (4), it is usual to obtain the so-called *Leontief quantity model* and the *Ghosh price model*, respectively. However, we must include also two other types of models that are not so often treated in the input-output literature but that deserve to be mentioned for the sake of comprehensiveness.

Quantity models

Equation (3) rests on the assumption of fixed technical coefficients being the new industry (product) output vector $(\mathbf{x}_1)$ required for an exogenously specified new final demand vector $(\mathbf{f}_1)$ such that,

$$\mathbf{x}_1 = (\mathbf{I} - \mathbf{A}_0)^{-1}\,\mathbf{f}_1 \tag{5}$$

Given a shock in the physical amounts consumed by final users of a product (or of the bundle of products produced by a certain industry, both primarily and secondarily produced), then the effect on the total output value of the industry (product) output is given by $\mathbf{x}_1$. Notice that in this *Leontief quantity model* there is no change in prices.

Furthermore, equation (5) can also be expressed as a ratio per unit of output value of the period 0 as[2],

$$\hat{\mathbf{x}}_0^{-1}\,\mathbf{x}_1 = \hat{\mathbf{x}}_0^{-1}\,(\mathbf{I}-\mathbf{A}_0)^{-1}\hat{\mathbf{x}}_0\hat{\mathbf{x}}_0^{-1}\,\mathbf{f}_1 = (\mathbf{I}-\mathbf{B}_0)^{-1}\hat{\mathbf{x}}_0^{-1}\,\mathbf{f}_1 \tag{6}$$

which gives the variation rate of the quantities produced to meet the new final demand. That is, the new output total value $(\mathbf{x}_1)$ results from the multiplication of old prices $(\mathbf{p}_0)$ by the new quantities demanded $(\mathbf{q}_1)$ such as,

$$\mathbf{x}_1 = \hat{\mathbf{p}}_0\,\mathbf{q}_1 \tag{7}$$

whilst the old output values result from the amounts consumed valued at prices of period 0, as

$$\hat{\mathbf{x}}_0^{-1} = (\hat{\mathbf{p}}_0\,\hat{\mathbf{q}}_0)^{-1} = \hat{\mathbf{q}}_0^{-1}\hat{\mathbf{p}}_0^{-1} \tag{8}$$

Then, by replacing the right-hand side (RHS) of equation (6) by equations (7) and (8), it is straightforward that,

$$\hat{\mathbf{q}}_0^{-1}\,\hat{\mathbf{p}}_0^{-1}\,\hat{\mathbf{p}}_0\,\mathbf{q}_1 = \hat{\mathbf{q}}_0^{-1}\,\mathbf{q}_1 = (\mathbf{I}-\mathbf{B}_0)^{-1}\,\hat{\mathbf{x}}_0^{-1}\,\mathbf{f}_1 \tag{9}$$

which is the so-called *Ghosh quantity model* (Dietzenbacher, 1997). A change in the final demand shares over the total output value of period 0 caused by variations in the quantities demanded will lead to changes in the quantities produced.

Price models Equation (4) is based on the assumption of fixed output coefficients. For a new value added vector $(\mathbf{v}_2^{\mathsf{T}})$, the new total output values are calculated by,

$$\mathbf{x}_2^{\mathsf{T}} = \mathbf{v}_2^{\mathsf{T}}\,(\mathbf{I}-\mathbf{B}_0)^{-1} \tag{10}$$

Given a price change in any of the primary factors used (generally speaking, capital and labour), then the effect on the output value of the industry (product) output is given by $\mathbf{x}_2$. Notice that in this *Ghosh price model* there is no change in quantities consumed of primary inputs and of goods and services.

Moreover, equation (10) can also be expressed as a ratio per unit of output value of the period 0 as,

$$\mathbf{x}_2^{\mathsf{T}}\,\hat{\mathbf{x}}_0^{-1} = \mathbf{v}_2^{\mathsf{T}}\,\hat{\mathbf{x}}_0^{-1}\,\hat{\mathbf{x}}_0\,(\mathbf{I}-\mathbf{B}_0)^{-1}\,\hat{\mathbf{x}}_0^{-1} = \mathbf{v}_2^{\mathsf{T}}\,\hat{\mathbf{x}}_0^{-1}\,(\mathbf{I}-\mathbf{A}_0)^{-1} \tag{11}$$

[2] The relationship between the Leontief and the Ghosh inverses can be found in Miller and Blair (2009, p. 548).

which gives the price variation of products generated by the variation in the prices of primary factors. That is, the new output total value ($\mathbf{x}_2$) results from the multiplication of old quantities produced ($\mathbf{q}_0$) by the new prices ($\mathbf{p}_2$) such as,

$$\mathbf{x}_2^{\mathsf{T}} = \mathbf{p}_2^{\mathsf{T}}\,\hat{\mathbf{q}}_0 \tag{12}$$

while the old output values result from the amounts consumed valued at prices of period 0, as

$$\hat{\mathbf{x}}_0^{-1} = (\hat{\mathbf{p}}_0\,\hat{\mathbf{q}}_0)^{-1} = \hat{\mathbf{q}}_0^{-1}\,\hat{\mathbf{p}}_0^{-1} \tag{13}$$

Therefore, by replacing the RHS of equation (11) by equations (12) and (13), it is easy to obtain that,

$$\mathbf{p}_2^{\mathsf{T}}\,\hat{\mathbf{q}}_0\,\hat{\mathbf{q}}_0^{-1}\,\hat{\mathbf{p}}_0^{-1} = \mathbf{p}_2^{\mathsf{T}}\,\hat{\mathbf{p}}_0^{-1} = \mathbf{v}_2^{\mathsf{T}}\,\hat{\mathbf{x}}_0^{-1}\,(\mathbf{I} - \mathbf{A}_0)^{-1} \tag{14}$$

which is the so-called *Leontief price model* or *supply-driven model* (Dietzenbacher, 1997). A change in value added shares over the total output value of period 0 caused by variations in the prices of primary inputs will lead to changes in product prices.

4.2. The relationship between the models and the choice of type of input-output table

Quantity models

The Ghosh and Leontief quantity models are demand driven models. They both measure the effects on the output (in physical and monetary values, respectively) of a change in final demand. To that purpose, the use of product-by-product tables would imply to assume a shock in the final demand of a specific product irrespectively of the industry that actually produced it. For instance, for an increase in the households' purchase of electric cars against fuel based vehicles one would need a product-by-product table in order to quantify the effects on the quantities of energy inputs supplied to meet such new demand. Furthermore, if greenhouse gas direct emissions are available on a product basis, the total effects on the environment can be easily calculated with a product-by-product table by multiplying the new output value $\mathbf{x}_1$ (from equation 5) by the emission levels per product output. Nevertheless, emission coefficients are mostly available on an industry basis, which then makes product-by-product tables unsuitable. Furthermore, if one eventually uses an industry-by-industry table the calculated effects would be caused instead by a change in the final demand of the bundle of goods and services produced by a specific industry, which is not necessarily that of a specific commodity. All in all, in the case of environmental analysis, the kind of data available and the objective of the analysis definitely play a major role in the choice of type of SIOT to be used.

Input-output analysis is also applied to labour market analyses through the calculation of employment multipliers under the Leontief quantity model. Due to the fact that employment data are usually recorded by firms and therefore grouped by industries, industry-by-industry tables may be more appropriate than product-by-product tables. It is not very likely to find employment data related to products. Moreover, one must bear in mind that the effects on employment thus calculated using industry-by-industry tables will be caused by a change in the final demand of a mixed bundle of goods and services produced by a certain sector, which does not necessarily be a single specific commodity.

The input-output quantity models are used to evaluate the effects of introducing a new product technology as well. Provided that the new technology refers to a single product and that it can be easily subtracted from its mother branch, the Leontief and Ghosh quantity models would allow for evaluating the effects on the output value (and physical amounts produced) of the other competing products. At this respect, product-by-product tables seems to be more suitable than industry-by-industry tables, where each industry produces more than one single product. Clearly, the new demand for a new product (e.g. electric cars) will drive a set of direct and indirect effects on the other product outputs.

The calculation of value added and income (wages and salaries) multipliers are also a matter of interest in the input-output literature. It is quite intuitive that the compensation of employees and the value added are clearly linked to industries rather than to products or homogenous branches. Industry-by-industry tables keep a direct link to the original statistical sources. Bearing this in mind, industry-by-industry tables are in this case also preferable to product-by-product tables although the IO literature admit several impact analyses on the basis of value added/income related to homogenous branches of activities.

As a summarizing remark, the IO quantity models are driven by changes in the amounts of goods and services consumed or demanded. The use of product-by-product tables is preferable since the shock can be easily assigned to a single product and the output effects can also be related to homogenous branches of activities. To the contrary, the use of industry-by-industry tables in this context would lead to measure the effects of a variation in the demanded quantity of a mixed bundle of goods and services produced by a certain industry on the industry output values and amounts of (mixed) goods and services produced. The choice favours clearly product-by-product tables almost in all cases. However, the Leontief quantity model is extensively used to account for many different kinds of multiplier effects, e.g. environmental, employment, income... that needs data that are almost solely available on an industry basis. To some extent, this justifies the use of industry-by-industry tables in some situations. Therefore, it seems to be a clear trade-off. Either one assumes that the additional data (environmental, employment, income...) is on a product basis and uses product-by-product tables to measure the effects on the output value (also in physical terms) of changes in final demand of single products, or one assumes that the additional data is on an industry basis and uses industry-by-industry tables, although being aware that the derived effects on total output values are referred not to single products but to a mixed bundle of goods and services produced by a certain industry.

Price models

The Ghosh and Leontief price models measure the effects of variations in the prices of primary inputs on the output value and on the prices of goods and services, respectively. The amount of factor inputs used remains unchanged and so the amounts of goods and services produced. These models are seen as supply-side driven models preferably to be used in cases of shortage of supply or excess of demand. Variations in salaries and wages per hour, in profit rates, in fixed capital use rates or in net tax rates[3] on production will generate changes in prices of goods and services and output value that could be quantified through the price IO models. As a result, industry-by-industry tables seems to be more suitable for these kind of analyses since initial changes are referred to the different components of the value added, which are directly linked to the surveyed firms data and/or groups of firms (industries) data. Indeed, statistical data on labour costs are referred to workers employed in industries and not in homogenous branches of activity. Environmentally oriented fiscal policies (excluding taxes on products) on taxes and subsidies on production (e.g. environmental tax) are commonly referred to the carbon emissions generated at the level of industries rather than to homogenous branches[4]. Moreover, profit rates are also related to firms and industries rather than to products.

Nevertheless, the price changes obtained through the IO price models using an industry-by-industry SIOT are not reflecting single product price variations but variations in the prices of a mixed bundle of goods and services produced by an industry. Hence, there is a clear trade-off again at this respect. Either one assumes that changes in primary inputs occur in homogenous branches and uses product-by-product tables to calculate single product price changes or one assumes that the price variations of primary factors occur in industries and uses industry-by-industry tables to obtain mixed product price changes. The choice is eventually up to the user.

Supply-use tables

Two major trade-offs have been identified concerning the choice of type of SIOT to be used in impact analysis. The main difficulty underlying the two trade-offs is referred to the symmetry of the SIOTs. They are defined as product-by-product or industry-by-industry type. Hence, if one is interested in estimating, for instance, the effects of an increase in the labour costs of the electricity sector (industry) on the prices of fuels

[3] Generally speaking, the taxes less subsidies on production included in the value added at basic prices are those that are not payable per unit of some good or service produced or transacted (ESA95).

[4] The ESA95 (4.22) includes taxes on pollution resulting from production activities as “other taxes on production” (D29); although, they may actually appear to be taxes on products (e.g. energy products). These pollution taxes consist of taxes levied on the emission or discharge into the environment of noxious gases, liquids or other harmful substances.

(product), then the choice of type of SIOT would lead to provide two different answers with neither of them being the correct one. On the one hand, if we use product-by-product tables we will be assigning the increase of labour costs to a homogenous branch of activity and not to the electricity sector and on the other hand, if we use industry-by-industry tables, the price effects will correspond to a mixed basket of goods and services of the fuel producing industry rather than to fuel.

To solve this issue, supply-use tables are clearly the best choice since they are defined on a product-by-industry basis rather than on a product or industry basis. However, there has been very little research on the application of supply and use tables to impact analysis. To the knowledge of the authors, the single contributions at this respect can be found in ten Raa and Rueda-Cantuche (2007) and in Rueda-Cantuche and Amores (2010). The former authors proved that employment and output multipliers (from the Leontief quantity model) can be derived from supply and use data by regressing employment (output) by industries on the net output[5] by products. Therefore, a change in the net output of products (implicitly a change in the final demand) will cause a variation in the employment (output) of industries. The interested reader may find more details in the cited paper. The latter contribution relates to environmental input-output impact analysis and applied the same concept to carbon dioxide emissions in Denmark. This line of research can be further extended methodologically to include time series of multiregional supply-use systems. So far it has been applied only to a single-country for one year only.

5. Conclusions and recommendations

This section summarizes the main conclusions and recommendations that can be drawn from the paper.

The construction of symmetric input-output tables (SIOTs) is a controversial issue in the input-output literature as regard the choice of model to construct both product-by-product and industry-by-industry SIOTs, especially the former ones. However, there has been so far little attention paid on the choice of type of SIOT to carry out impact analyses let alone other input-output applications. The UN and Eurostat systems of national accounts just simply refer to this issue vaguely and basically recommend nothing except that the purpose of the analysis will determine the choice of type to be used. Moreover, there are no explicit guidelines for the user to make the correct choice accordingly with its own purpose.

In empirical research, it depends on the objectives of the analysis which type of table is best suited for economic analysis. Particularly in impact analyses, questions

[5] Ten Raa and Rueda-Cantuche (2007) defined net output as the difference between the intermediate parts of the supply and use matrices, which incidentally makes the final demand vector if one sums the elements of the net output matrix over columns.

like, for example, what fuel price effects would generate an increase in the labour costs of the electricity industry cannot really be answered by input-output price models as it is generally thought. Moreover, this is even independent of the type of SIOT used. Either one assumes that changes in primary costs (labour) occur in homogeneous branches rather than in industries and therefore uses product-by-product tables or one assumes that the price changes of primary factors effectively occur in industries and thus, uses industry-by-industry tables. Nonetheless, the corresponding reported price effects will be those of the fuel industry rather than those of the fuel product itself.

As regard input-output quantity models there is also a trade-off in the case of impact analyses related to environment, employment... or any economic dimension for which data is mainly available on an industry basis. Either one assumes that the additional data external to the input-output system (employment, emissions...) is on a product basis and uses a product-by-product table to evaluate the total effects of a change in the amount of the final demand consumed of a single product (like e.g. bio-fuels) or one assumes that the additional data is on an industry basis and uses industry-by-industry tables. Nevertheless, the derived total effects on employment, emissions... will correspond to a change in the output of a mixed bundle of goods and services produced by a certain industry rather than to changes in single product outputs.

Two major trade-offs have been identified concerning the choice of type of SIOT to be used in input-output impact analyses. The main shortcoming underlying this issue is related to the symmetry of SIOTs. They are defined as either product-by-product or industry-by-industry type. To solve this matter efficiently, supply and use tables are clearly the best choice since they are defined on a product-by-industry basis rather than solely on a product or industry basis. It is therefore advisable to follow the lines of the pioneering works of ten Raa and Rueda-Cantuche (2007) and Rueda-Cantuche and Amores (2010) and continue exploring the use of supply and use tables in the calculation of input-output impact multipliers of any kind. These authors currently propose to use econometric techniques to estimate unbiased and consistent input-output effects of any kind (emissions, employment, income...) from Model A and rectangular supply and use tables. This new approach opens up the door to further research with the other three models (B, C and D) and to provide possibly the first reliable inference based results in input-output analysis (including hypotheses tests, confidence intervals...). Of course, one can always come back to standard input-output analysis bearing in mind the methodological trade-offs addressed in this paper.

6. References

Armstrong, A.G. (1975). Technology assumptions in the construction of United Kingdom input–output tables, in: R.I.G. Allen and W.F. Gossling (eds) *Estimating and Updating Input–Output Coefficients* (London, Input–Output Publishing).

Dietzenbacher, E. (1997). In vindication of the Ghosh model : a reinterpretation as a price model, *Journal of Regional Science*, 37, 629–651.

Eurostat (1995). *European system of accounts—ESA* (Luxembourg, Office for Official Publications of the European Communities).

Eurostat (2008). *Eurostat manual of supply, use and input-output tables*, Methodologies and working papers, (Luxembourg, Office for Official Publications of the European Communities)

Eurostat (2009). *European system of accounts 2008 —ESA08* (Luxembourg, Office for Official Publications of the European Communities). Draft version: restricted.

Gigantes, T. (1970). The representation of technology in input–output systems, in: A.P. Carter and A. Brôdy (eds) *Contributions to Input–Output Analysis* (Amsterdam, North-Holland).

Konijn, P.J.A. (1994). *The make and use of commodities by industries*. PhD Thesis (Enschede, The Netherlands, University of Twente).

Konijn, P.J.A. and Steenge, A. E. (1995). Compilation of input–output data from the national accounts, *Economic Systems Research*, 7, 31–45.

Kop Jansen, P.S.M. and ten Raa, T. (1990). The choice of model in the construction of input-output coefficients matrices, *International Economic Review*, 31, 213–227.

Miller, R.E. and Blair. P.D. (2009). *Input-Output Analysis. Foundations and Extensions*. (Cambridge, Cambridge University Press).

Office of Statistical Standards (1974). *Input–output tables for 1970* (Tokyo, Institute for Dissemination of Government Data).

Rueda-Cantuche, J.M. and Amores, A.F. (2010). Consistent and unbiased carbon dioxide emission multipliers: performance of Danish emission reductions via external trade, *Ecological Economics*, 69, 988–998.

Rueda-Cantuche, J.M., Beutel, J., Neuwahl, F., Mongelli, I. and Loeschel, A. (2009). A symmetric input-output table for EU27: latest progress, *Economic Systems Research*, 21, 59–79.

Rueda-Cantuche, J.M. and ten Raa, T. (2009). The choice of model in the construction of industry coefficients matrices, *Economic Systems Research*, 21, 363–376.

ten Raa, Th., Chakraborty, D. and Small, J.A. (1984). An alternative treatment of secondary products in input–output analysis, *Review of Economics and Statistics*, 66, 88–97.

ten Raa, T. and Rueda-Cantuche, J.M. (2003). The construction of input-output coefficients matrices in an axiomatic context: some further considerations, *Economic Systems Research*, 15, 439–455.

ten Raa, T. and Rueda-Cantuche, J.M. (2007). Stochastic analysis of input-output multipliers on the basis of use and make matrices, *Review of Income and Wealth*, 53, 2, 318–334.

United Nations (1968). *A system of national accounts*, Studies in Methods Series F, nr. 2, rev. 3. (New York, United Nations).

United Nations (1973). *Input–Output tables and analysis*, Studies in Methods Series F, nr. 14, rev. 1. (New York, United Nations).

United Nations (1993). *Revised system of national accounts*, Studies in Methods Series F, no. 2, rev. 4 (New York, United Nations).

United Nations (1999). *Handbook of input-output table compilation and analysis*, Studies in Methods, Handbook of National Accounts, Series F, no. 74 (New York, United Nations).

United Nations (2009). *System of national accounts 2008*, (New York, European Commission, International Monetary Fund, Organization for Economic Co-operation and Development, United Nations and World Bank): http://unstats.un.org/unsd/nationalaccount/sna2008.asp

Viet, V.Q. (1994). Practices in input–output table compilation, *Regional Science and Urban Economics*, 24, 27–54.

SORT 35 (1) January-June 2011, 39-62

Imputation of numerical data under linear edit restrictions

Wieger Coutinho[1], Ton de Waal[2] and Marco Remmerswaal[3]

Abstract

A common problem faced by statistical offices is that data may be missing from collected data sets. The typical way to overcome this problem is to impute the missing data. The problem of imputing missing data is complicated by the fact that statistical data often have to satisfy certain edit rules, which for numerical data usually take the form of linear restrictions. Standard imputation methods generally do not take such edit restrictions into account. In the present article we describe two general approaches for imputation of missing numerical data that do take the edit restrictions into account. The first approach imputes the missing values by means of an imputation method and afterwards adjusts the imputed values so they satisfy the edit restrictions. The second approach sequentially imputes the missing data. It uses Fourier-Motzkin elimination to determine appropriate intervals for each variable to be imputed. Both approaches are not based on a specific imputation model, but allow one to specify an imputation model. To illustrate the two approaches we assume that the data are approximately multivariately normally distributed. To assess the performance of the imputation approaches an evaluation study is carried out.

MSC: 62-02, 62-07, 90C90

Keywords: Fourier-Motzkin elimination, imputation, linear edit restrictions, linear programming.

1. Introduction

National statistical institutes (NSIs) publish figures on many aspects of society. To this end, these NSIs collect data on persons, households, enterprises, public bodies, etc. A major problem that has to be faced is that data may be missing from the collected data sets. Some units that are selected for data collection cannot be contacted or may refuse to

[1] Tarwekamp 172, 2592 XN The Hague, The Netherlands

[2] Statistics Netherlands, PO Box 24500, 2490 HA The Hague, The Netherlands (tel. +31 70 337 4930, e-mail: t.dewaal@cbs.nl or tonwaal@planet.nl).

[3] Rijswijk University of Professional Technical Education, Lange Kleiweg 80, 2288 GK Rijswijk, The Netherlands
Received: June 2009
Accepted: November 2010

respond altogether. This is called unit non-response. Unit non-response is not considered in this article. For many records, i.e. the data of individual respondents, data on some of the items may be missing. Persons may, for instance, refuse to provide information on their income or on their sexual habits, while at the same time giving answers to other, less sensitive questions on the questionnaire. Enterprises may not provide answers to certain questions, because they may consider it too complicated or too time-consuming to answer these specific questions. Missing items of otherwise responding units is called item non-response. Whenever we refer to missing data in this article we will mean item non-response.

Missing data is a well-known problem that has to be faced by basically all institutes that collect data on persons or enterprises. In the statistical literature ample attention is hence paid to missing data. The most common solution to handle missing data in data sets is imputation, where missing values are estimated and filled in. An important problem of imputation is to preserve the statistical distribution of the data set. This is a complicated problem, especially for high-dimensional data. For more on this aspect of imputation and on imputation in general we refer to Kalton and Kasprzyk (1986), Rubin (1987), Kovar and Whitridge (1995), Schafer (1997), Little and Rubin (2002), and Longford (2005).

At NSIs the imputation problem is further complicated owing to the existence of constraints in the form of edit restrictions, or edits for short, that have to be satisfied by the data. Examples of such edits are that the profit and the costs of an enterprise have to sum up to its turnover, and that the turnover of an enterprise should be at least zero. Records that do not satisfy these edits are inconsistent, and are hence considered incorrect. While imputing a record, we aim to take these edits into account, and thus ensure that the final, imputed record satisfies all edits. The imputation problem at NSIs is hence given by: impute the missing data in the data set under consideration in such a way that the statistical distribution of the data is preserved as well as possible subject to the condition that all edits are satisfied by the imputed data.

For academic statisticians the wish of NSIs to let the data satisfy specified edits may be difficult to understand. Statistically speaking there is indeed hardly a reason to let a data set satisfy edits. However, as Pannekoek and De Waal (2005) explain, NSIs have the responsibility to supply data for many different, both academic and non-academic, users in society. For the majority of these users, inconsistent data are incomprehensible. They may reject the data as being an invalid source or make adjustments themselves. This hampers the unifying role of NSIs in providing data that are undisputed by different parties such as policy makers in government, opposition, trade unions, employer organizations, etc. As mentioned by Särndal and Lundström (2005, p. 176): “Whatever the imputation method used, the completed data should be subjected to the usual checks for internal consistency. All imputed values should undergo the editing checks normally carried out for the survey”.

Simple sequential imputation of the missing data, where edits involving fields that have to be imputed subsequently are not taken into account while imputing a field, may lead to inconsistencies. Consider, for example, a record where the values of two variables, *x* and *y*, are missing. Assume these variables have to satisfy three edits saying that *x* is at least 50, *y* is at most 100, and *y* is greater than or equal to *x*. Now, if *x* is imputed first without taking the edits involving *y* into account, one might impute the value 150 for *x*. The resulting set of edits for *y*, i.e. *y* is at most 100 and *y* is greater than or equal to 150, cannot be satisfied. Conversely, if *y* is imputed first without taking the edits involving *x* into account, one might impute the value 40 for *y*. The resulting set of edits for *x*, i.e. *x* is at least 50 and 40 is greater than or equal to *x*, cannot be satisfied.

In this article we develop two general approaches for imputation of missing numerical data that ensure that edits are satisfied, while at the same time allowing one to specify a statistical imputation model. Despite the fact that much research on imputation techniques has been carried out, imputation under edits is still a rather neglected area. As far as we are aware, apart from some research at NSIs (see, e.g., Tempelman, 2007) hardly any research on general approaches to imputation under edit restrictions has been carried out. An exception is imputation based on a truncated multivariate normal model (see, e.g., Geweke, 1991, and Tempelman, 2007). Imputation based on a truncated multivariate normal model can take the edit restrictions we consider in this article into account. Using this model has two drawbacks, however. First of all, the truncated multivariate model is computationally very demanding and complex to implement in a software program. Second, it is obviously only suited for data that (approximately) follow a truncated multivariate normal distribution, not for data that follow other distributions. Some software packages developed by NSIs, such as GEIS (Kovar and Whitridge, 1990), SPEER (Winkler and Draper, 1997), SLICE (De Waal, 2001) and Banff (Banff Support Team, 2008), also ensure that edits are satisfied after imputation. However, these packages only apply relatively simple imputation models, whereas our approaches allow more complicated imputation models.

Both approaches we describe in this article allow one to separate the imputation model from how the edits are handled. In other words, the two approaches described are not based on a specific imputation model, but allow one to specify an imputation model. For both approaches a broad class of imputation models can be applied.

To illustrate the two approaches we will assume in this article that the data are approximately multivariately normally distributed. In fact, in our calculations we will treat the unknown distribution of the data as being a multivariate normal distribution exactly. For data that have to satisfy edits defined by linear inequalities this is surely incorrect, because at best the data could follow a truncated normal distribution but never a regular normal distribution. Our simplification makes it relatively easy to determine marginal and conditional distributions, which are needed for one of the two imputation approaches examined in this article. We only use the (approximate) multivariate normal model to illustrate how our general approaches can actually be applied in practice. We have selected the (approximate) multivariate normal model for computational convenience. We certainly do not want to suggest that this model is the most appropriate one for the data sets we have used in our evaluation study. Another computationally convenient choice would have been to use hot-deck imputation instead of the (approximate) multivariate normal model.

In order to estimate the parameters of the multivariate normal distribution, we have used the EM algorithm. As starting values for the EM algorithm we have used the observed means and covariance matrix of the complete cases. Our implementation of the EM algorithm is based on Schafer (1997).

The remainder of this article is organised as follows. Section 2 first discusses the kind of linear edits on which we will focus in this article. Section 3 describes an adjustment approach where imputed records are later adjusted so they satisfy the specified edits. A second imputation approach is described in Section 5. A fundamental role in this approach is played by Fourier-Motzkin elimination. We refer to this imputation approach as the FM approach. The Fourier-Motzkin elimination technique itself is explained in Section 4. Section 6 illustrates the FM approach by means of an example. An evaluation study and its results for the (approximate) multivariate normal model are described in Section 7. In that section we compare the results of the adjustment approach with the FM approach for the multivariate normal imputation model. Finally, Section 8 concludes the article with a short discussion.

2. Linear edit restrictions

In this article we focus on linear edits for numerical data. Linear edits are either linear equations or linear inequalities. We denote the number of continuous variables by n, and the variables themselves by x_i $(i = 1,\dots,n)$. We assume that edit j $(j = 1,\dots,J)$ can be written in either of the two following forms:

$$a_{1j}x_1 + \ldots + a_{nj}x_n + b_j = 0, \tag{1}$$

or

$$a_{1j}x_1 + \ldots + a_{nj}x_n + b_j \geq 0. \tag{2}$$

Here the a_{ij} and the b_j are certain constants, which define the edit.

Edits of type (1) are referred to as balance edits. An example of such an edit is

$$T = P + C, \tag{3}$$

where T is the turnover of an enterprise, P its profit, and C its costs. Edit (3) expresses that the profit and the costs of an enterprise should sum up to its turnover. A record not satisfying this edit is obviously incorrect. Edit (3) can be written in the form (1) as $T - P - C = 0$.

Edits of type (2) are referred to as inequality edits. An example is

$$T \geq 0, \tag{4}$$

expressing that the turnover of an enterprise should be non-negative. An inequality edit such as (4), expressing that the value of a variable should be non-negative, is also referred to as a non-negativity edit.

3. An adjustment approach

A straightforward approach to let imputed values satisfy specified edits is to use an adjustment approach consisting of two steps. In the first step the missing data are imputed without taking the edits (1) and (2) into account. These missing data can, for instance, be imputed by assuming that the data follow a multivariate normal distribution, and use a standard imputation method for this situation (see, e.g., Little and Rubin, 2002, and Schafer, 1997). As already mentioned, in this article we illustrate our approaches by indeed assuming that the data follow a multivariate normal distribution, and impute the missing data of a record by drawing values from the appropriate estimated conditional distribution for the missing data given the observed values. We refer to this as the first imputation step.

We denote the values after the first imputation step for the record under consideration by $x_{\text{first},i}$ $(i = 1,\dots,n)$. In the second step, the adjustment step, the final values in the record under consideration, $x_{\text{final},i}$ $(i = 1,\dots,n)$, are determined by minimising the objective function

$$\sum_i w_{\text{adj},i} \left| x_{\text{first},i} - x_{\text{final},i} \right| \tag{5}$$

subject to the condition that the values $x_{\text{final},i}$ $(i = 1,\dots,n)$ satisfy all edits (1) and (2) and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value. The latter condition means that only the values *imputed* in the first imputation step may be modified. In (5) the $x_{\text{first},i}$ $(i = 1,\dots,n)$ are known values and the $x_{\text{final},i}$ $(i = 1,\dots,n)$ are the unknowns. The $w_{\text{adj},i}$ $(i = 1,\dots,n)$ are non-negative adjustment weights, reflecting how serious one considers a change of a unit in variable x_i to be.

The adjustment weights $w_{\text{adj},i}$ $(i = 1,\dots,n)$ can be calculated in many ways. In our application we have set $w_{\text{adj},i} = 1/\overline{x}_{\text{first},i}$, where $\overline{x}_{\text{first},i}$ is the average value of the i-th variable. In this way, the objective function (5) takes the relative deviation between $x_{\text{first},i}$ and $x_{\text{final},i}$ rather than the absolute deviation into account. All the weights $w_{\text{adj},i} = 1/\overline{x}_{\text{first},i}$ $(i = 1\dots,n)$ were indeed non-negative.

The problem of minimising the objective function (5) subject to the condition that the values $x_{\text{final},i}$ $(i = 1,\dots,n)$ satisfy all edits (1) and (2) can be formulated as a linear programming problem by introducing additional variables u_i $(i = 1,\dots,n)$ and adding the constraints

$$u_i \geq x_{\text{first},i} - x_{\text{final},i} \tag{6}$$

and

$$u_i \geq x_{\text{final},i} - x_{\text{first},i}. \tag{7}$$

It is easy to see that the problem of minimising the objective function

$$\sum_i w_{\text{adj},i} u_i \tag{8}$$

subject to (6), (7), the condition that the values $x_{\text{final},i}$ $(i = 1,\dots,n)$ satisfy all edits (1) and (2) and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value yields the same optimal value for the objective function (8) and the same optimal values for $x_{\text{final},i}$ $(i = 1,\dots,n)$ as minimising (5) subject to the condition that the values $x_{\text{final},i}$ $(i = 1,\dots,n)$ satisfy all edits (1) and (2), and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value (see also Chvátal, 1983).

In the problem of minimising (8) subject to (6), (7), the condition that the values $x_{\text{final},i}$ $(i = 1,\dots,n)$ satisfy all edits (1) and (2), and the condition that for all variables x_i that were observed $x_{\text{final},i}$ equals the corresponding observed value, the $x_{\text{final},i}$ and u_i $(i = 1,\dots,n)$ are the unknowns. This linear programming problem can, for instance, be solved by means of the well-known simplex algorithm, an interior-point algorithm (see, e.g., Chvátal, 1983, and Nemhauser and Wolsey, 1988) or a generalized reduced gradient method (see, e.g., Lasdon et al., 1978).

The adjustment approach is quite a general and logical approach. In the first step one can apply the imputation method and imputation model that are best from a statistical point of view for the data under consideration. In the second step the imputed values are (hopefully only slightly) adjusted so they satisfy the specified edits.

The main strength of the adjustment approach is its simplicity: one does not need to implement complicated algorithms in a computer program or buy special-purpose software. Standard software, such as Excel, suffices to implement the adjustment approach. In our application we have indeed used the solver offered by Excel. To be precise: we have used the generalized reduced gradient method as implemented in the GRG2 code of the Excel solver (see Lasdon et al., 1978, and Fylstra, 1998). We have used the GRG2 code of Excel instead of the implementation of the simplex algorithm in Excel as we noted that the GRG2 method as implemented in Excel resulted in a larger number of records not “lying on the boundary of the feasible region defined by the edits”.

In our evaluation study we pay special attention to the number of records “lying on the boundary of the feasible region defined by the edits”. In this article we define a record to “lie on the boundary of the feasible region defined by the edits” if at least one of the inequality edits is satisfied with equality. We are aware that this is a ambiguous definition, and also one that differs from the usual definition of “lying on the boundary” as used in the theory of linear programming. Namely, our definition of “lying on the boundary of the feasible region defined by the edits” is dependent on how the edits are stated, rather than only on the shape of the feasible region. For instance, an edit given by “$x = y + z$” can also be expressed as two inequality edits: “$x \leq y + z$” and “$x \geq y + z$”. In the latter case, *all* records will lie on the boundary of the feasible region defined by the edits after imputation according to our definition. In our definition we implicitly assume that edits are stated as balance edits instead of (pairs of) inequality edits whenever possible. In all practical situations occurring at statistical offices we have encountered so far this was always the case.

The reason why we pay special attention to the number of records on the boundary of the feasible region defined by the edits is that, when the adjustment approach is applied, a record that does not satisfy the edits after the first imputation step, will often be adjusted in such a way that the final, adjusted, record lies on the boundary of the feasible region defined by the edits.

In the next three sections we describe our second imputation approach, the FM approach. We begin the description of the FM approach by explaining Fourier-Motzkin elimination.

4. Eliminating variables by means of Fourier-Motzkin elimination

Fourier-Motzkin elimination (see, e.g., Duffin, 1974, and De Waal and Coutinho, 2005) is a technique to project a set of linear constraints involving m variables onto a set of linear constraints involving $m-1$ variables. The original set of constraints involving m variables can be satisfied if and only if the corresponding, projected set of constraints involving $m-1$ variables can be satisfied. The standard version of Fourier-Motzkin elimination handles only inequalities as constraints. We use an extended version of Fourier-Motzkin elimination that can also handle equations. In our application of Fourier-Motzkin elimination the constraints are defined by the edits.

In order to eliminate a variable x_r from the set of current edits by means of Fourier-Motzkin elimination, we start by copying all edits not involving this variable from the set of current edits to a new set of edits Ψ.

If variable x_r occurs in an equation, we express x_r in terms of the other variables. Say, x_r occurs in edit s of type (1), we then write x_r as

$$x_r = -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} x_i\right) \tag{9}$$

Expression (9) is used to eliminate x_r from the other edits involving x_r. These other edits are hereby transformed into new edits, not involving x_r, that are logically implied by the old ones. These new edits are added to our new set of edits Ψ. Note that if the original edits are consistent, i.e. can be satisfied by certain values u_i $(i = 1,\ldots,m)$, then the new edits are also consistent as they can be satisfied by u_i $(i = 1,\ldots,m; i \neq r)$. Conversely, note that if the new edits are consistent, say they can be satisfied by values v_i $(i = 1,\ldots,m; i \neq r)$, then the original edits are also consistent as they can be satisfied by the values v_i $(i = 1,\ldots,m)$ where v_r is defined by filling v_i $(i = 1,\ldots,m; i \neq r)$ into (9).

If x_r does not occur in an equality but only in inequalities, we consider all pairs of edits (2) involving x_r. Suppose we consider the pair consisting of edit s and edit t. We first check whether the coefficients of x_r in those inequalities have opposite signs, i.e. we check whether $a_{rs} \times a_{rt} < 0$. If this is not the case, we do not consider this particular combination (s,t) anymore. If the coefficients of x_r do have opposite signs, one of the edits, say edit s, can be written as an upper bound on x_r, i.e. as

$$x_r \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i\neq r} a_{is}x_i\right), \tag{10}$$

and the other edit, edit t, as a lower bound on x_r, i.e. as

$$x_r \geq -\frac{1}{a_{rt}}\left(b_t + \sum_{i\neq r} a_{it}x_i\right). \tag{11}$$

Edits (10) and (11) can be combined into

$$-\frac{1}{a_{rt}}\left(b_t + \sum_{i\neq r} a_{it}x_i\right) \leq x_r \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i\neq r} a_{is}x_i\right),$$

which yields an implied edit not involving x_r given by

$$-\frac{1}{a_{rt}}\left(b_t + \sum_{i\neq r} a_{it}x_i\right) \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i\neq r} a_{is}x_i\right). \tag{12}$$

The implied edit (12) is added to our new set of edits Ψ. After all possible pairs of edits involving x_r have been considered and all implied edits given by (12) have been generated and added to Ψ, we delete the original edits involving x_r that we started with. In this way we obtain a new set of edits Ψ not involving variable x_r. This set of edits Ψ may be empty. This occurs, for instance, when all current edits are inequalities involving x_r and the coefficients of x_r in all those inequalities have the same sign. Note that if the original edits are consistent, say they can be satisfied by certain values u_i $(i = 1,\ldots,m)$, then the new edits are also consistent as they can be satisfied by u_i

$(i = 1,\ldots,m; i \neq r)$. This is by definition also true if the new set of edits is empty. Conversely, note that if the new edits are consistent, say they can be satisfied by certain values v_i $(i = 1,\ldots,m; i \neq r)$, then the minimum of the right-hand sides of (12) for the v_i $(i = 1,\ldots,m; i \neq r)$ is larger than, or equal to, the maximum of the left-hand sides of (12) for the v_i $(i = 1,\ldots,m; i \neq r)$. This implies that we can find a value v_r such that

$$-\frac{1}{a_{rt}}\left(b_t + \sum_{i \neq r} a_{it} v_i\right) \leq v_r \leq -\frac{1}{a_{rs}}\left(b_s + \sum_{i \neq r} a_{is} v_i\right) \qquad \text{for all pairs } s \text{ and } t,$$

which in turn implies that the original edits are consistent. We have demonstrated the main property of Fourier-Motzkin elimination: a set of edits is consistent if and only if the set of edits after elimination of a variable is consistent. Note that as one only has to consider pairs of edits, the number of implied edits is obviously finite. We illustrate Fourier-Motzkin elimination by means of the example below.

Example: Suppose there are four variables, T (turnover), P (profit), C (costs), and N (number of employees), and that the edits are given by (3), (4),

$$P \leq 0.5T, \tag{13}$$

$$-0.1T \leq P, \tag{14}$$

$$T \leq 550N. \tag{15}$$

If we eliminate variable P, we use equation (3) to express P in terms of T and C. That is, we use $P = T - C$. After Fourier-Motzkin elimination, we obtain the edits (4), (15),

$$T - C \leq 0.5T, \qquad (\text{equivalently: } 0.5T \leq C) \tag{16}$$

and

$$-0.1T \leq T - C \qquad (\text{equivalently: } C \leq 1.1T). \tag{17}$$

The main property of Fourier-Motzkin elimination says that the original set of edits (3), (4), and (13) to (15) for T, P, C and N can be satisfied if and only if the set of edits (4), and (15) to (17) for T, C and N can be satisfied.

This was an example of Fourier-Motzkin elimination if the variable to be eliminated is involved in an equation. We now use the resulting set of edits (4), and (15) to (17) for variables T, C and N to give an example of the elimination of a variable involved in inequalities only. If we eliminate variable C from edits (4), and (15) to (17), we first copy the edits not involving C, i.e. edits (4) and (15). Moreover, we can combine edits (16) and (17) to obtain

$$0.5T \leq 1.1T, \tag{18}$$

which is equivalent to (4). So, eliminating C from (4), and (15) to (17) leads to edits (4) and (15). The main property of Fourier-Motzkin elimination says that the set of edits (4), and (15) to (17) for T, C and N can be satisfied if and only if edits (4) and (15) for T and N can be satisfied. Combining the two results we have found, we conclude that the original set of edits (3), (4), and (13) to (15) for T, P, C and N can be satisfied if and only if edits (4) and (15) for T and N can be satisfied. □

5. An imputation approach based on Fourier-Motzkin elimination

The FM approach consists of the following steps:

0. Assume a statistical imputation model for the data, and – if necessary for the model – estimate the model parameters.

We order the variables to be imputed from the variable with the most missing values to the variable with the least missing values. If two of more variables have the same number of missing values, we order them in an arbitrary way. For each record to be imputed, we apply Steps 1 to 5 below. We repeat this process until all records have been imputed.

1. Fill in the values of the non-missing data into the edits. This leads to a set of edits $E(0)$ involving only the variables to be imputed for the record under consideration.
2. Use Fourier-Motzkin elimination to eliminate the variables to be imputed for the record under consideration from set of edits $E(0)$ in the fixed order described above until only one variable remains. The set of edits after the i-th variable to be imputed has been eliminated is denoted by $E(i)$. The final set of edits defines a feasible interval for the remaining variable. Set k equal to the number of variables to be imputed for the record under consideration.
3. Draw a value for the k-th variable to be imputed.
4. If the drawn value lies inside the feasible interval $E(k-1)$, accept it and go to Step 5. If it lies outside the feasible interval, reject it and return to Step 3.
5. If $k = 1$, all variables have been imputed and we stop. Otherwise, we fill in the drawn value for the selected variable k into the edits in $E(k-2)$. This defines a feasible interval for the $(k-1)$-th eliminated variable. We update k by $k := k-1$, and go to Step 3.

Note that the theory developed in Section 4 implies that if the record to be imputed can be imputed consistently, the feasible interval determined in Step 2 or 5 is never empty.

In Step 0 one can either assume an implicitly defined statistical imputation model, for instance when one wants to apply hot-deck imputation, or an explicitly defined imputation model, such as the multivariate normal model like we do in this article. In both cases we suggest to draw a value for the variable to be imputed from the conditional distribution of the selected variable given all known values, either observed or already imputed ones.

If the feasible interval determined in Step 2 has width 0, there is only one feasible value for the variable under consideration. In this case it is not necessary to draw a value in Step 3. Instead we immediately impute the only feasible value. In some other cases the width of the feasible interval determined in Step 2 may be rather small. In those cases many values may need to be drawn before a value inside the feasible interval is drawn. We therefore set a limit, N_{draw}, on the number of times that a value for a particular variable may be drawn. If this limit is reached, and no value inside the feasible has been drawn, the last value drawn is set to the nearest value of the feasible interval. By means of N_{draw} one can indirectly control the number of imputed records on the boundary of the feasible region defined by the edits. If N_{draw} is set to a low value, relatively many imputed records will be on this boundary; if N_{draw} is set to a high value, relatively few imputed records will be on the boundary.

The variables are imputed in reverse order of elimination. Since we have ordered the variables to be imputed from the variable with the most missing values to the variable with the least missing values before applying Steps 1 to 5 of the above algorithm, the variables are imputed in order of increasing number of missing values. That is, the variable with the least missing values is imputed first and the variable with the most missing values last.

As mentioned before, to illustrate our approaches we assume in this article that the data are multivariately normally distributed, and we use the EM algorithm to estimate the model parameters.

It is well known that in the worst case Fourier-Motzkin elimination can be computationally very expensive. However, the imputation problems arising in practice at statistical offices only have a limited number of variables and edits. The largest problems we are aware of have a few hundreds of variables and slightly more than 100 edits. For realistic problems of this limited size, Fourier-Motzkin elimination is generally sufficiently fast. In fact, it has been shown for the related – but computationally much more demanding – error localization problem of the same size in terms of variables and edits that in practical cases arising at statistical offices the computational performance of Fourier-Motzkin elimination is generally acceptable (see De Waal and Coutinho, 2005, and De Waal, 2005). In our application of Fourier-Motzkin elimination in this article to small imputation problems, the computing time of Fourier-Motzkin elimination was negligible, i.e. close to 0 seconds, for all runs. Once the parameters of the multivariate normal distribution had been determined by means of the EM algorithm, imputing the missing values took only took a few seconds for the entire data sets. Moreover, in the imputation process, the bulk of the computing time for the FM approach was spent on drawing values from the multivariate normal distribution rather than on Fourier-Motzkin elimination

The main reason for developing the FM approach is the fact that promising results have been obtained by so-called sequential imputation methods. Sequential imputation methods are a well-known class of imputation methods, see, e.g., Van Buuren and Oudshoorn (1999 and 2000), Raghunatan et al. (2001) and Rubin (2003). These imputation methods sequentially impute the variables and allow a separate imputation model to be specified for each variable. By imputing all variables containing missing data in turn and iteratively repeating this process several times, the statistical distribution of the imputed data generally converges to an unspecified multivariate distribution. The main strength of sequential imputation is its flexibility: rather than using one multivariate imputation model for all variables simultaneously, which is generally computationally demanding and complex to handle, one can specify a different imputation model for each variable. Sequential imputation methods can be extended to ensure that they satisfy edits. In principle, the FM approach can be implemented as a sequential imputation approach that allows such an extension, although in our illustration we assume a multivariate normal distribution as imputation model rather than separate imputation models for the variables to be imputed (see Tempelman, 2007, and Pannekoek, Shlomo and De Waal, 2008, for other extensions of sequential imputation to ensure that edits are satisfied).

Of course, the adjustment approach may also be used in a sequential imputation approach, namely one may first use a sequential imputation approach and later adjust the imputed values so they satisfy the edits. A fundamental difference between this approach and the FM approach is that in the adjustment approach the imputed values are adjusted simultaneously afterwards, whereas in the FM approach each separate imputed value is immediately adjusted in order to ensure that all edits can be satisfied. Immediately adjusting each imputed value in order to ensure that all edits can be satisfied might improve (or deteriorate) the statistical results as subsequent imputed values may depend on previously imputed values (see also Section 7.3).

6. Illustration of the FM approach

In this section we illustrate the FM approach by means of an example. In our example, we assume that we are given a data set with some missing values, that there are four variables, T, P, C and N, and that the edits are given by (3), (4) and (13) to (15).

We focus on Steps 1 to 5 of the approach for a specific record. We assume that the data follow a multivariate normal distribution, and assume that the model parameters, means $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, estimated in Step 0 of our approach are given by

$$\boldsymbol{\mu} = (1000, 200, 500, 4)$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 13500 & 3000 & 10500 & 60 \\ 3000 & 2500 & 500 & 10 \\ 10500 & 500 & 10000 & 50 \\ 60 & 10 & 50 & 1 \end{pmatrix}.$$

Here the first column/row corresponds to T, the second column/row to P, the third column/row to C, and the fourth column/row to N.

Now, suppose that for a certain record in our data set we have $N = 5$, and that the values for T, P and C are missing. We first fill in the observed value for N into the edits (3), (4) and (13) to (15) (Step 1 of our approach). We obtain (3), (4), (13), (14) and

$$T \leq 2750, \tag{19}$$

Now, we sequentially eliminate the variables for which the values are missing from the edits. We start by eliminating P from (3), (4), (13), (14) and (19). This leads to the edits (4), (16), (17) and (19). Edits (4), (16), (17) and (19) have to be satisfied by C and T.

We next eliminate variable C, and obtain edits (4), (18) and (19). Edit (18) is equivalent to (4). The edits that have to be satisfied by T are hence given by (4) and (19). The feasible interval for T is therefore given by [0, 2750]. We have now completed Step 2 of our approach.

To impute T, we determine the distribution of T, conditional on the value for variable N. The distribution of T turns out to be N(1060, 9900), the normal distribution with mean 1060 and variance 9900. We draw values from this distribution until we draw a value inside the feasible interval (Steps 3 and 4 of the approach). Suppose we draw the value 1200.

We fill in the imputed value for T into the edits for C and T, i.e. edits (4), (16), (17) and (19) (Step 5 of the approach). We obtain

$$1200 \geq 0,$$

$$600 \leq C,$$

$$C \leq 1320,$$

$$1200 \leq 2750.$$

The feasible interval for C is hence given by [600, 1320]. We determine the distribution of C, conditional on the values for variables N and T. This distribution turns out to be

N(656.11, 18181.18). We draw values from this distribution until we draw a value inside the feasible interval (Steps 3 and 4 of the approach). Suppose we draw the value 700.

We fill in the imputed values for C and T into the edits that have to be satisfied by C, T and P, i.e. edits (3), (4), (13), (14) and (19) (Step 5 of the approach). We obtain

$$1200 = P + 700,$$

$$1200 \geq 0,$$

$$P \leq 600,$$

$$-120 \leq P,$$

$$1200 \leq 2750.$$

There is only one feasible value for P, namely 500. The imputed record we obtain is given by $T = 1200$, $C = 700$, $P = 500$, and $N = 5$.

7. Evaluation study

7.1. Evaluation data

For our evaluation study we have used three data sets: a data set with actually observed data from a business survey, data set R_{all}, the same data set but without balance edits, data set R_{ineq}, and a data set with synthetic data, data set S. The data sets R_{all} and R_{ineq} contain raising weights. These raising weights differ across different (strata of) records, and are used in some of our evaluation measures. In data S all raising weights were set to 1. The main characteristics of these data sets are presented in Table 1.

Table 1: *The characteristics of the evaluation data sets.*

	Data set R_{all}	Data set R_{ineq}	Data set S
Total number of records	3,096	3,096	500
Number of records with missing values	544	469	490
Total number of variables	8	7	10
Total number of edits	14	12	16
Number of balance edits	1	0	3
Total number of inequality edits	13	12	13
Number of non-negativity edits	8	7	9

The actual values for data set R_{all}, and hence also for data set R_{ineq}, are all known. In the completely observed data set values were deleted by a third party, using a mechanism unknown to us. Data set R_{ineq} was constructed in order to examine the effects of balance edits on the results. In fact, we have “removed” the balance edit from data set R_{all} in two different ways. First of all, we have only “removed” the balance edit, i.e. did not explicitly demand that after imputation the balance edit holds true for all records, but have left all involved variables in the data set. As a consequence, the estimated covariance matrix will be singular and the balance edit will be automatically satisfied by the imputed data, if the parameters of the normal distribution are estimated by means of the EM algorithm using the complete cases to obtain a first estimate for the model parameters as we do in our application. We refer the interested reader to Chapter 4 in Tempelman (2007) for a proof. The evaluation results should hence be the same as for the case where all edits are used, apart from some minor differences due to the stochastic nature of the approaches used. This is confirmed by our evaluation study (results not reported in this article). Second, we have removed one of the variables, R_4, involved in the balance edit and its associated non-negativity edit from R_{all}. R_{ineq} is the resulting data set. This data set obviously does not have to satisfy any balance edit. The removed variable R_4 does not occur in any of the other edits apart from its associated non-negativity edit.

Data set S is indirectly based on an observed business survey and its corresponding edits. This observed data was used to estimate the parameters of a multivariate normal model by means of the EM algorithm. Next, data set S was generated by drawing from the estimated multivariate normal model. If a drawn vector did not satisfy all specified edits it was rejected, else it was accepted. In this way 500 vectors were generated. Missing values were generated by randomly deleting for each variable a specified number of values. The number of values deleted was (much) higher than in the actually observed business survey in order to evaluate the performance of our imputation approaches for a very complicated situation.

For all three data sets we have two versions available: a version with missing values and a version with complete records. The former version is imputed. The resulting data set is then compared to the version with complete records, which we consider as a data set with the true values.

The numbers of missing values and (unweighted) means of the 8, respectively 7, variables of data set R_{all} and data set R_{ineq} are given in Table 2 and those of the 10

Table 2: *The numbers of missing values and the means of the variables of data sets R_{all} and R_{ineq}.*

Variable	Number of missing values	Mean
R_1	76	11,574.83
R_2	79	777.56
R_3	130	8,978.70
R_4	147	1,034.07
R_5	68	10,012.77
R_6	67	169.24
R_7	73	209.86
R_8	0	37.41

Table 3: *The numbers of missing values and the means of the variables of data set S.*

Variable	Number of missing values	Mean
R_1	120	97.77
S_2	180	175,018.30
S_3	240	731.03
S_4	120	175,749.33
S_5	180	154,286.53
S_6	180	7,522.34
S_7	180	8,519.65
S_8	180	1,277.04
S_9	120	171,605.57
S_{10}	120	4,143.76

variables of data set S in Table 3. The means are taken over all observations in the complete versions of the data sets.

Variable R_8 in data sets R_{all} and R_{ineq} does not contain any missing values and is only used as auxiliary variable.

7.2. Evaluation measures

To measure the performance of our imputation approaches we use several evaluation measures, The first measure we use is the d_{L1} measure proposed by Chambers (2003). This d_{L1} measure is the average distance between the imputed and true values defined as

$$d_{L1} = \frac{\sum_{k \in M} w_k \, |\hat{y}_k - y_k^*|}{\sum_{k \in M} w_k},$$

where $\hat{y}_k$ is the imputed value in record k of the variable under consideration, y_k^* the corresponding true value, M denotes the set of n_{imp} records with imputed values for variable y and w_k is the raising weight for record k.

The second measure we use is the m_1 measure, which has also been proposed by Chambers (2003). This measure, which measures the preservation of the first moment of the empirical distribution of the true values, is defined as

$$m_1 = \left| \sum_{k \in M} w_k (\hat{y}_k - y_k^*) / \sum_{k \in M} w_k \right| .$$

That m_1 measures the preservation of the first moment of the empirical distribution of the true values becomes clear if we rewrite m_1 as

$$m_1 = \left| \sum_{k \in D} w_k (\hat{y}_k - y_k^*) / \sum_{k \in D} w_k \right| \times \left(\frac{\sum_{k \in D} w_k}{\sum_{k \in M} w_k} \right),$$

where D denotes the entire data set. The quantity $\sum_{k\in D} w_k y_k^* / \sum_{k\in D} w_k$ is an estimate for the population mean. So, m_1 is the deviation of the first moment of the empirical distribution from the first moment of the true distribution times the constant factor $\sum_{k\in D} w_k / \sum_{k\in M} w_k$.

The third measure is the *rdm* (relative difference in means) measure. This measure has been used in an evaluation study by Pannekoek and De Waal (2005), and is defined as

$$rdm = \frac{\sum_{k\in M} \hat{y}_k - \sum_{k\in M} y_k^*}{\sum_{k\in M} y_k^*}.$$

Smaller absolute values of the above three measures indicate better imputation performance.

To remain consistent with the literature, in particular with the previously published papers by Chambers (2003) and Pannekoek and De Waal (2005), we have not made an attempt to make the d_{L1} and the m_1 measures comparable across variables.

The three evaluation measures described so far all measure the deviation of the imputed values from the true values. The next three evaluation measures measure statistical aspects, such as the preservation of the empirical distribution and the preservation of standard errors.

The first of these measures is the percent difference between the standard deviation (STD) of the mean of the imputations to the standard deviation of the mean of the true values:

$$100\frac{(STD_{\text{imp}} - STD_{\text{true}})}{STD_{\text{true}}}$$

A smaller absolute value for the percent difference between the standard deviation of the mean of the imputations to the standard deviation of the mean of the true values indicates better performance.

Another evaluation measure is a sign test using paired data. This sign test can be carried out by creating a new variable that is defined as the difference between the original value and the imputed value. The test with the null hypothesis that the median of the difference is equal to zero is equivalent to the test that the medians of the original and imputed values are equal. The sign statistic is defined as

$$S = (n^+ - n^-)/2$$

where n^+ is the number of values greater than 0 and n^- the number of values less than 0. A small p-value means that we reject the null hypothesis of equal medians. We will interpret this as: a larger p-value indicates better performance.

The next evaluation measure we use is the Kolmogorov-Smirnov non-parametric test statistic (*K-S*). This statistic is used to compare the empirical distribution of the original values to the empirical distribution of the imputed values (also proposed by Chambers, 2003). For unweighted data, the empirical distribution of the original values is defined as: $F_{y^*}(t) = \sum_{i \in M} I(y_i^* \leq t)/n_{\text{imp}}$, where n_{imp} is again the number of imputed values and I the indicator function. $F_{\hat{y}}(t)$ is defined similarly. The *K-S* is defined as:

$$K\text{-}S = \max_j (|F_{y^*}(t_j) - F_{\hat{y}}(t_j)|),$$

where the $\{t_j\}$ values are the $2n_{\text{imp}}$ jointly ordered original and imputed values of y. A smaller value for *K-S* indicates better performance.

The final evaluation measure we consider is the number of records on the boundary of the feasible region defined by the edits. Records lying on the boundary of the feasible region defined by the edits are outliers in some sense. Too many outliers in this sense, and any other sense, could make the imputed data less suited for certain statistical analyses. The number of records on the boundary defined by the edits should preferably be close to the actual number of records of the true data on the boundary of the feasible region defined by the edits. This evaluation measure is a bit less important than the others, which measure the statistical quality of the imputed values more directly.

We use the measures in a relative way, namely to compare the adjustment approach to the FM approach. The measures are neither necessarily appropriate nor sufficient to measure the impact of imputation on the quality of survey estimates in general. For an actual production process it depends on the intended use of the data which of the evaluation measures is considered more important.

7.3. Evaluation results

Both imputation approaches described in this article are of a stochastic nature as they depend on drawing vectors from a probability distribution. To reduce the effects of the stochastic nature of our approaches we have repeated each evaluation experiment 10 times, and have calculated the average of these 10 experiments. Unless stated otherwise the value of N_{draw} for the FM approach (see Section 5) is set to 160 in our experiments. The value of 160 for N_{draw} is based on a limited explorative, trial-and-error search, aiming to find an optimal trade-off between the quality of the imputations and the required computing time. The results for data set R_{all} are presented in Table 4 for the adjustment approach and Table 5 for the FM approach.

Table 4: *Evaluation results for the adjustment approach on data set* R_{all}.

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	2069.21	1145.83	0.15	0.02	0.57	0.03
R_2	226.91	108.27	0.17	0.22	0.90	0.00
R_3	303.79	285.46	−0.21	−0.13	0.12	0.01
R_4	283.05	263.84	1.79	3.71	0.00	0.01
R_5	16.11	13.21	−0.01	0.00	0.00	0.38
R_6	41.00	40.61	2.65	0.10	0.00	0.03
R_7	86.37	75.14	1.42	1.87	0.07	0.00

Table 5: *Evaluation results for the FM approach on data set* R_{all}.

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	3108.97	2633.30	0.34	−0.02	0.00	0.00
R_2	290.66	235.89	0.33	0.11	0.00	0.00
R_3	169.68	130.85	−0.04	0.00	0.02	0.17
R_4	183.83	152.04	0.40	0.16	0.00	0.02
R_5	68.29	61.31	0.01	0.00	0.13	0.74
R_6	27.37	26.87	1.83	−0.40	0.00	0.00
R_7	95.44	92.48	2.17	0.87	0.00	0.00

Variable R_8 does not have any missing values, so no evaluation results for R_8 are presented in Tables 4 and 5. The results for data set R_{ineq} are presented in Table 6 for the adjustment approach and Table 7 for the FM approach.

Table 6: *Evaluation results for the adjustment approach on data set* R_{ineq}

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	1868.22	256.14	−0.25	−0.36	0.01	0.10
R_2	205.16	34.67	−0.37	−0.42	0.00	0.00
R_3	1490.74	1451.99	−0.99	−0.93	0.00	0.00
R_5	1227.87	541.04	−0.49	−0.44	0.00	0.00
R_6	2783.81	2783.81	592.50	58.43	0.00	0.00
R_7	14.40	12.03	−0.54	−0.47	0.00	0.36

Table 7: *Evaluation results for the FM approach on data set* R_{ineq}

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	3105.74	2719.82	0.33	−0.01	0.00	0.00
R_2	278.66	225.06	0.30	0.10	0.00	0.00
R_3	359.48	277.00	−0.09	0.00	0.00	0.01
R_5	1844.58	1762.83	0.14	−0.01	0.00	0.00
R_6	27.07	26.66	1.78	−0.39	0.00	0.00
R_7	85.50	82.26	1.80	0.60	0.00	0.00

Table 8: *Evaluation results for the adjustment approach on data set S*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	13943.12	13916.90	142.57	863.15	0.20	0.00
S_2	17440.92	8066.39	0.05	0.17	0.10	0.06
S_3	9941.38	9767.14	13.14	68.89	0.00	0.00
S_4	32672.09	31633.86	0.19	0.37	0.00	0.11
S_5	11404.99	5274.79	−0.04	−0.02	0.00	0.35
S_6	2221.02	1430.56	0.18	0.37	0.00	0.00
S_7	3472.59	1405.63	0.16	0.51	0.00	0.01
S_8	5062.49	4818.50	3.63	11.52	0.00	0.00
S_9	5715.68	3569.85	0.02	0.00	0.87	0.95
S_{10}	28261.21	28064.01	7.22	20.89	0.00	0.00

Table 9: *Evaluation results for the FM approach on data set S.*

Variable	d_{L1}	m_1	*rdm*	*percent difference*	*sign*	*K-S*
R_1	62.39	50.19	0.51	5.81	0.01	0.00
S_2	6754.16	2204.84	−0.01	−0.01	0.00	0.11
S_3	3413.06	3268.38	4.40	28.46	0.00	0.00
S_4	4594.46	3229.51	0.02	0.00	0.13	0.66
S_5	35442.70	28136.41	−0.19	0.01	0.56	0.00
S_6	3600.36	2597.57	−0.33	0.59	0.00	0.00
S_7	15202.73	10779.74	1.21	8.49	0.79	0.00
S_8	21984.15	21247.71	16.01	81.38	0.13	0.00
S_9	3959.69	1940.22	0.01	0.00	0.87	0.95
S_{10}	2223.89	1289.30	0.33	3.76	0.20	0.05

The results for data set S are presented in Table 8 for the adjustment approach and Table 9 for the FM approach.

It is hard to draw conclusions from Tables 4 to 9. For some variables the adjustment approach leads to better results than the FM approach. For other variables the opposite happens. This is not very surprising as both approaches rely on the same statistical model for drawing imputation values, which fails to capture all distributional aspects of the data. In order to draw some conclusions we examine how often one approach leads to better results than the other, where “better” is defined as “closer to zero” for all evaluation measures considered in Tables 4 to 9 except for the sign test. For the sign test “better” is defined in the opposite way, i.e. the larger the p-value, the better the performance. For data set R_{all}, the results for the adjustment approach in Table 4 are in 19 cases better than those for the FM approach in Table 5. The opposite happens in 16 cases. For data set R_{ineq}, the results for the adjustment approach in Table 6 are in 13 cases better than those for the FM approach in Table 7. The opposite happens in 15 cases. For data set S, the results for the adjustment approach in Table 8 are in 20 cases better than those for the FM approach in Table 9. The opposite happens in 31 cases. From this we conclude that for data sets R_{all} and R_{ineq} the results for the six evaluation measures of the adjustment approach are comparable to the results for the FM approach. The inclusion or exclusion of the balance edit in R_{all}, respectively R_{ineq} does not seem to affect the results much. For the more complicated data set S the FM approach leads to slightly better results than the adjustment approach. This is probably caused by the fact that in the FM approach the values imputed cannot be too far from their true values as each separately imputed value is at worst on the boundary of its feasible interval. This imputed value is later used as predictor in order to impute other missing values. In the adjustment approach the values imputed in the first step may be far from their true values. For the complicated data set S, this is apparently not, or in any case to an insufficient extent, corrected in the adjustment step.

In Table 10 the average number of records on the boundary of the feasible region over 10 evaluation experiments for the adjustment approach and the FM approach on data sets R_{all}, R_{ineq}, and S are presented. For the FM approach we show the results for three different values of N_{draw}, namely the values 1, 160 and 1000. The value of N_{draw} used is mentioned between brackets. The results for the six evaluation measures considered before for $N_{\text{draw}} = 1$ and $N_{\text{draw}} = 1000$ (not presented here) are comparable to the results presented in Tables 5, 7, and 9, where $N_{\text{draw}} = 160$. In Table 10 we also present the number of records on the boundary of the feasible region for the complete versions of the three mentioned data sets. In almost all cases records of these data sets lie on the boundary of the feasible region because a variable that has to satisfy a non-negativity edit attains the value zero.

Table 10: *(Average) number of records on the boundary of the feasible region defined by the edits.*

	Average number for FM approach (1)	Average number for FM approach (160)	Average number for FM approach (1000)	Average number for the adjustment approach	Actual number for complete data
Data set R_{all}	499.4	468.2	468.0	499.8	495
Data set R_{ineq}	435.8	397.4	397.0	394.1	424
Data set S	200.5	186.6	186.8	185.5	2

Table 10 shows that the result for data set R_{ineq} for the FM approach is closer to the actual number of records on the boundary of the feasible region defined by the edits for the complete data than the adjustment approach for any of the three values of N_{draw}. For data set R_{all} it depends of the value of N_{draw} which approach leads to a result that is the closest to the actual number of records on the boundary for the complete data. For data set S the results of the adjustment approach are slightly closer to the actual number of records on the boundary for the complete data than the FM approach for any of the three values of N_{draw}. The difference between the results for the adjustment approach and the FM approach for $N_{\text{draw}} = 160$ are, however, negligible.

Table 10 also shows the effect of the parameter N_{draw} of the FM approach: the higher N_{draw}, the less records will generally lie on the boundary of the feasible region. By means of N_{draw} one can indirectly control the number of records on the boundary of the feasible region.

If one wants, for the FM approach, the number of imputed records on the boundary of the feasible region defined by the edits to be close to the actual number of records on the boundary for the complete data, one should choose N_{draw} between 1 and 160 for data sets R_{all} and R_{ineq}. Data set S appears to be too complicated for both the adjustment and the FM approach. The number of imputed records on the boundary of the feasible region is too high for both approaches. By increasing the value of N_{draw} the number of records on the boundary decreases only slowly for the FM approach. Increasing the value of N_{draw} also leads to an increase of the computing time, however. So, although one can influence the number of records on the boundary of the feasible region by changing the value of N_{draw}, the effect of changing the value of N_{draw} is limited, in any case for complicated data sets such as S. The drawback of the adjustment approach noted in Section 3 that the number of records on the boundary of the feasible region for this approach is for a substantial part determined by the first imputation step does not appear to be a major disadvantage in comparison to the FM approach – at least not for our evaluation data – as the results of the adjustment approach are not clearly worse than those of the FM approach in this respect.

8. Discussion

In this article we have described two imputation approaches that lead to imputed data that satisfy specified edits. The main aim of the article was to describe the two general frameworks, which are basically independent of the imputation method or imputation model actually applied. To illustrate how these approaches work in practice we have used a multivariate normal imputation model.

For the data sets in our evaluation study we conclude that, for the multivariate normal imputation model, for 2 of the 3 data sets (R_{all} and R_{ineq}) the FM approach leads to comparable evaluation results as the adjustment approach. For the other data set (data set S) the FM approach leads to (slightly) better than the adjustment approach (see Tables 8 and 9). The FM approach seems to have a built-in mechanism to protect itself from imputing very wrong values. Such a mechanism seems to be lacking from the adjustment approach. Our study is, however, very limited and more research is necessary before we can draw any definite conclusions.

In our application of the adjustment approach we have used a linear objective function. The main reason for using a linear objective function is that this is easy to implement in a software program. The results of the adjustment approach may possibly be improved by using a quadratic objective function instead of our linear one. In any case, for statisticians, minimising a quadratic objective function is more natural and often more logical than minimising a linear objective function.

The FM approach has the advantage that one can, indirectly, control the number of records on the boundary of the feasible region defined by the edits. The price that has to be paid for this is that the algorithm is more complicated than for the adjustment approach. Moreover, the effect of this indirect control over the number of records on the boundary of the feasible region seems limited. From a purely practical point of view, the adjustment approach may therefore be a better choice in many cases.

For data set S, far too many records lie on the boundary of the feasible region for both the adjustment approach and the FM approach. For almost all records on the boundary one or more non-negativity edit is satisfied with equality, i.e. the value of the involved variable equals zero. The fact that far too many non-negativity edits are satisfied with equality strongly indicates that the assumed imputation model, which in our application is assumed to follow a multivariate normal distribution, is incorrect. In order to improve the statistical results of the two imputation approaches presented in this article, the underlying statistical model should be improved. Further research is required to develop such better statistical models as well as computationally tractable methods to handle such models.

When imputing a missing value in a record in our implementation of the FM approach, we use the previously imputed values in this record as auxiliary information. In this way we try to preserve the correlation structure between the imputed values as much as possible. Using previously imputed values in order to impute a missing value has an obvious drawback: if the stochastic imputation process leads to a bad imputed value, this affects all subsequently imputed values in this record. It remains to be examined if the results of the FM approach improve, or deteriorate, if we do not use the previously imputed values as auxiliary information but instead use only the observed data as auxiliary information.

The imputation approaches we have developed in this article can be applied to general linear edit restrictions. If only non-negativity edits are specified, one could possibly also use tobit and logit models instead of our approaches. Such models automatically ensure that each variable to be imputed attains a non-negative value. The use of tobit or logit models for imputation subject to non-negativity edits remains to be examined.

Acknowledgments

The authors would like to thank the reviewers and the editor of the SORT for their useful comments on earlier versions of this article.

References

Banff Support Team (2008). *Functional Description of the Banff System for Edit and Imputation*. Technical Report, Statistics Canada.

Chambers, R. (2003). Evaluation Criteria for Statistical Editing and Imputation. In: *Methods and Experimental Results from the EUREDIT Project* (ed. J.R.H. Charlton) (available on http://www.cs.york.uk/euredit/).

Chvátal, V. (1983). *Linear Programming*. W.H. Freeman and Company, New York.

De Waal, T. (2001). SLICE: Generalised software for statistical data editing. In: *Proceedings in Computational Statistics* (ed. J.G. Bethlehem and P.G.M. Van der Heijden), Physica-Verlag, StateNew York, 277–282.

De Waal, T. (2005). Automatic error localisation for categorical, continuous and integer data, *Statistics and Operations Research Transactions*, 29, 57–99.

De Waal, T. and W. Coutinho (2005). Automatic editing for business surveys: an assessment of selected algorithms. *International Statistical Review*, 73, 73–102.

Duffin, R.J. (1974). On Fourier's analysis of linear inequality systems. *Mathematical Programming Studies*, 1, 71–95.

Fylstra, D. L. Lasdon, J. Watson and A. Warren (1998). Design and use of the Microsoft Excel Solver, *Interface*, 28, 29–55.

Geweke, J. (1991). *Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities*. Report, University of Minnesota.

Kalton, G. en D. Kasprzyk (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1–16.

Kovar, J. and P. Whitridge (1990). Generalized edit and imputation system; overview and applications. *Revista Brasileira de Estadistica,* 51, 85–100.

Kovar, J. en P. Whitridge (1995). Imputation of business survey data. In: *Business Survey Methods* (ed. B.G. Cox, D. A. Binder, B.N. Chinnappa, A. Christianson, M. J. Colledge & P.S. Kott), John Wiley & Sons, New York, 403–423.

Lasdon, L.S. and S. Smith (1992). Solving large sparse non-linear programs using GRG, *ORSA Journal on Computing*, 4, 2–15.

Little, R.J.A. and D.B. Rubin (2002). *Statistical Analysis with Missing Data (second edition).* John Wiley & Sons, New York.

Longford, N.T. (2005). *Missing Data and Small-Area Estimation*. Springer, New York.

Nemhauser, G.L. and L.A. Wolsey (1988). *Integer and Combinatorial Optimisation.* Wiley, New York.

Pannekoek, J. and T. De Waal (2005). Automatic edit and imputation for business surveys: the Dutch contribution to the EUREDIT project. *Journal of Official Statistics*, 21, 257–286.

Pannekoek, J., N. Shlomo and T. De Waal (2008). *Calibrated Imputation of Numerical Data under Linear Edit Restriction*. UN/ECE Work Session on Statistical Data Editing, Vienna.

Raghunathan, T.E., J.M. Lepkowski, J. Van Hoewyk and P. Solenberger (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

Rubin, D.B. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica*, 57, 3–18.

Särndal, C.-E. and S. Lundström (2005). *Estimation in Surveys with Nonresponse*, John Wiley & Sons, Chichester.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Tempelman, C. (2007). *Imputation of Restricted Data*. Doctoral thesis, University of Groningen.

Van Buuren, S. and C.G.M., Oudshoorn (1999). *Flexible Multivariate Imputation by MICE*. TNO Preventie en Gezondheid, TNO/PG 99.054, Leiden.

Van Buuren, S. and C.G.M. Oudshoorn C.G.M. (2000). *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*. Report PG/VGZ/00.038, TNO Preventie en Gezondheid, Leiden.

Winkler, W.E. and L.A. Draper (1997). The SPEER edit system. *Statistical Data Editing (Volume 2); Methods and Techniques*, United Nations, Geneva.

Selected article from

XII Conferencia Española de Biometría 2009

SORT 35 (1) January-June 2011, 65-80

Poverty comparisons when TIP curves intersect

Miguel A. Sordo[1,*] and Carmen D. Ramos[2]

Universidad de Cádiz

Abstract

Non-intersection of TIP curves is recognized as a criterion to compare two income distributions in terms of poverty. The purpose of this paper it to obtain comparable poverty results for income distributions whose TIP curves intersect (possibly more than once). To deal with such situations, a sequence of higher-degree dominance criteria between TIP curves is introduced. The normative significance of these criteria is provided in terms of a sequence C_n of nested classes of linear poverty measures with the property that, as the order n of the class increases, the measures become more and more sensitive to the distribution of income among the poorest.

MSC: 91B82, 60E15

Keywords: Poverty measure, poverty ordering, CPG curve, TIP curve, poverty.

1. Introduction

Since the seminal paper of Sen (1976) on poverty measurement, a large body of literature dealing with this topic has been published. Because an important reason for measuring poverty is to make comparisons, part of the literature has developed by focusing on partial poverty orderings, which require unanimity in poverty rankings for a class of measures that obey some normative principles, with a fixed poverty line (see Zheng (2000) for a review of this topic).

[*] The research of Miguel A. Sordo was supported by Ministerio de Ciencia e Innovación (grant MTM2009-08326) and Consejería de Economía Innovación y Ciencia (grant P09-SEJ-4739). The research of Carmen D. Ramos was supported by Consejería de Economía Innovación y Ciencia (grant P09-SEJ-4739)

[1] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Económicas y Empresariales, Universidad de Cádiz, Duque de Nájera, 8, 11002 Cádiz, Spain. mangel.sordo@uca.es

[2] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Sociales y de la Comunicación, Universidad de Cádiz, Avda. de la Universidad s/n, 11405 Jerez de la Frontera, Spain. carmen.ramos@uca.es

Received: November 2009
Accepted: November 2010

Poverty orderings are sometimes based on comparisons of TIP curves. The TIP (Three I's of Poverty) curve (Jenkins and Lambert, 1997) cumulates the poverty gaps of the bottom p proportion of the population. In order to introduce this dominance device, consider an income random variable X with distribution F and let F^{-1} be the corresponding right continuous quantile function defined by

$$F^{-1}(t) = \sup\{x : F(x) \leq t\}, \; t \in [0,1].$$

Let $z > 0$ the poverty line. The proportion of poor people, $r_z(X)$, is given by

$$r_z(X) = \sup\{F(x) : x < z\},$$

and the censored quantile function F_z^{-1} is defined, for all $t \in [0,1]$, as

$$F_z^{-1}(t) = \begin{cases} F^{-1}(t) & \text{if} \quad t < r_z(F) \\ z & \text{if} \quad t \geq r_z(F) \end{cases}.$$

Censored quantiles are, therefore, just the incomes $F^{-1}(t)$ for those in poverty (below z) and z for those whose income exceeds the poverty line. The poverty gap associated with income $F^{-1}(t)$ is defined as $z - F_z^{-1}(t)$. The TIP curve (also sometimes referred to as the Cumulative Poverty Gap curve or the Poverty Profile curve; see Spencer and Fisher (1992), Shorrocks (1995, 1998) and Jenkins and Lambert (1998a, 1998b) associated to X is given by

$$G_X(p,z) = \int_0^p \left(z - F_z^{-1}(t)\right) dt, \; p \in [0,1]. \tag{1}$$

In this paper, motivated by the second-order TIP dominance criterion introduced by Sordo *et al.* (2007), a family of higher-degree poverty orderings based on comparisons of TIP curves is considered. Although one finds in the literature several results concerning different notions of high-degree poverty orderings (including those by Shorrocks and Foster (1987), Foster and Shorrocks (1988) or Zheng (1999)) a higher-degree dominance criterion based on TIP curves has not been considered before. The normative significance of this family of orderings is provided in terms of a class C of poverty measures which has attracted a growing interest in recent years (see Davidson and Duclos (2000), Duclos and Grégoire (2002), Duclos and Araar (2006) and Sordo *et al.* (2007)). Members of this class have the following functional form:

$$I_X(\Phi,z) = \int_0^1 \left(z - F_z^{-1}(t)\right) d\Phi(t), \tag{2}$$

where the poverty gaps are weighted with a continuous probability distribution, Φ, with support $\Delta(\Phi) \subseteq [0,1]$. The class C, is analogous to the class of linear inequality measures proposed by Mehran (1976) for inequality indices and Yaari (1988), for social welfare indices. Following Duclos and Araar (2006), members of C satisfy the following axioms: Pareto (the measure does not increase whenever someone's income increases), focus (the measure depends only on the income of the poor), symmetry (permuting the incomes has no influence on the value of the measure) and replication invariance (the measure is not affected by the pooling of several identical populations). In addition, members of

$$C_1 = \{I(\Phi, z) \in C \text{ such that } \Phi \text{ is concave}\} \tag{3}$$

satisfy the Pigou-Dalton Principle of Transfer (any mean-preserving transfer from a poor person to a poorer person that leaves unchanged their relative rank in the distribution, must decrease poverty). Duclos and Araar (2006, Section 10.1) shows that the class C_1 can be characterized in terms of the first-degree TIP dominance criterion as follows:

$$G_X(p,z) \leq G_Y(p,z) \text{ for all } p \in [0,1] \Leftrightarrow I_X(\Phi, z) \leq I_Y(\Phi, z)\,, \text{for all } I(\Phi, z) \in C_1.$$

As shown by Sordo *et al.* (2007), when TIP curves intersect, comparable results are possible by restricting attention to a class C_2 whose members satisfy the Diminishing Transfer Principle, which strengthens the Pigou-Dalton Principle of Transfer by requiring that the reduction of poverty resulting from a transfer from a poor person to a poorer person is higher the poorer the recipient. Namely

$$C_2 = \{I(\Phi, z) \in C_1 \text{ such that } \phi \text{ is convex, where } \Phi'(t) = \phi(t) \text{ almost everywhere}\} \tag{4}$$

Specifically, Sordo *et al.* (2007) show that

$$I_X(\Phi, z) \leq I_Y(\Phi, z)\,, \text{for all } I(\Phi, z) \in C_2$$

if and only if

$$\int_0^p G_X(t,z)\,dt \leq \int_0^p G_Y(t,z)\,dt, \text{ for all } p \in [0,1] \text{ and } G_X(1,z) \leq G_Y(1,z). \tag{5}$$

However, even (5) can be a strong requirement for many pair of distributions, which can fail to satisfy it. This justifies the convenience of employing a weaker criterion to compare income distributions in terms of poverty. In this paper, to deal with such situations, a sequence of higher-degree poverty orderings, which generalizes (5), is considered and its normative significance is provided in terms of a family C_n of classes of poverty measures which generalizes (4).

In Section 2, we introduce the family C_n and the *n*degree TIP curve orderings. The main characterization is stated in Section 3. An example is given in Section 4 and Section 5 contains final remarks and conclusions.

2. Poverty measures and high-degree poverty orderings

The family C given by functionals of the form (2) contains some important measures. A subclass $S \subset C$ of particular interest emerges from considering the weight function

$$\Phi_n(p) = \{1-(1-p)^n\},\ n \geq 1. \tag{6}$$

As noted by Duclos (2000) and Duclos and Grégoire (2002), $I_X(\Phi_n, z)$, $n > 1$, depends upon an ethical parameter n, which captures the sensitivity of poverty measurement to "exclusion" or "relative deprivation" aversion: the greater the value of n, the more weight is given to the relative deprivation of the poor. They refer to $I_X(\Phi_n, z) = S_X(n,z)$ as the equally distributed equivalent (EDE) poverty gap that is socially equivalent to the actual distribution of poverty gaps and compare its properties with those of additive poverty indices. $S_X(n,z)$ also can be interpreted as the higher poverty gap in a sample of n randomly selected poor individuals. The class S contains some poverty measures that are well known from the literature. It includes the so-called "per-capita income gap" or $FGT(1)$ proposed by Foster *et al.* (1984) and is obtained when $n = 1$ (that is, $\Phi_n(t)$ is the uniform distribution on $(0,1)$). The Thon (1979), Chakravarty (1983) and Shorrocks (1995) poverty indices are obtained when $n = 2$.

Two more general subclasses of C, which turn out to be crucial in the course of this work, are defined below. Let $\Phi^{(i)}$ denote the ith derivative of Φ, $i = 1,2,...$

Definition 1 *C_n is the class of indices $I_X(\Phi, z) \in C$ such that Φ is at least n times differentiable, $(-1)^i \Phi^{(i+1)} \geq 0$ for $i = 0,1,..,n-1$ and $(-1)^{n-1}\Phi^{(n)}$ is non-increasing[1]. C_n^* is the class of indices $I_X(\Phi,z) \in C_n$ such that $\Phi^{(i)}(1) = 0$, $i = 1,\ldots,n$.*

For $n = 1$ and $n = 2$, C_n reduces to (3) and (4) respectively. Note that $C_{k+1} \subset C_k$ for $k = 1,2,\ldots$ Also, note that $S_X(k,z) \in C_n^*$ for $k \geq n+1$. On the other hand, that not every measure of interest of C_n belongs to S, is shown by the measure proposed by Thon (1983), obtained from (2) by choosing

$$\Phi(t) = \frac{c^2}{4(c-1)} - \frac{1}{c-1}\left(\frac{c}{2} - t\right)^2,\ c > 2.$$

It can be shown, by an argument similar to that used by Duclos (2000), that a social decision-marker who employs $I_X(\Phi, z) \in C_k$, with $k \geq 1$, is more sensitive to transfers occurring within the lower part of the distribution and, as k increases, the weight assigned to the effect of these transfers also increases.

As we follow in the next section, comparisons of income distributions according to the indices of the classes C_n and C_n^*, for all integer $n \geq 1$, can be characterized by means

[1] We also include in C_n indices $I_F(\Phi, z)$ where $\Phi^{(n)}$ exists except possibly at a countable number of points. Thus, $I_F(\Phi_{n,x}, z)$, where $\Phi_{n,x}$ is the $n$iterated integration of a "wedge" function of the form $\Phi_x(t) = (t-x)^+ = \max\{t-x, 0\}$, $x \in (0,1)$, is included in C_n.

of a family of stochastic orderings based on comparing TIP areas and equally distributed equivalent (EDE) poverty gaps. Given a poverty line z, denote $G_X^{[1]}(p,z) = G_X(p,z)$, $0 \leq p \leq 1$, and define

$$G_X^{[n]}(p,z) = \int_0^p G_X^{[n-1]}(t,z)dt, \text{ for } n = 2,3,\ldots \text{ and } 0 \leq p \leq 1. \tag{7}$$

Definition 2 *Given two income random variables X and Y and a common poverty line z, we say that X dominates Y in the nth degree TIP curve ordering (denoted by $X \geq_{TIP(n,z)} Y$) if $S_X(k,z) \geq S_Y(k,z)$ for $k = 1,2,\ldots,n$ and $G_X^{[n]}(p,z) \geq G_Y^{[n]}(p,z)$ for all $p \in [0,1]$.*

Before obtaining further results, we need the following easy-to-prove auxiliary lemma.

Lemma 3 *For any real value x, denote $x^+ = \max\{x,0\}$.*
(i) For a fixed $p \in [0,1]$, the functions $\Psi_{p,1}$, defined by $\Psi_{p,1}(t) = (t-p)^+$ and $\Psi_{p,n}(t)$, defined by

$$\Psi_{p,n}(t) = \int_0^t \Psi_{p,n-1}(x)dx, \; n = 2,3,\ldots, \tag{8}$$

satisfy $\Psi_{p,n}^{(k)}(t) \geq 0$ for $k = 1,2,\ldots,n-1$ and $\Psi_{p,n}^{(n)}(t) \geq 0$ except at $t = p$.

(ii) $I_X(\Phi_{p,n},z)$ belongs to C_n, where

$$\Phi_{p,n}(t) = 1 - \Psi_{p,n}(1-t), \; n = 1,2,\ldots \tag{9}$$

We also need the following useful result.

Lemma 4 *For each $n \geq 2$, we have*

$$G_X^{[n]}(p,z) = \int_0^1 G_X^{[n-k]}(t,z)d\Phi_{1-p,k}(t), \;\; k = 1,2,\ldots,n-1. \tag{10}$$

Proof. Let $n \geq 2$ fixed. We use induction on k to prove the lemma. For $k = 1$, we have from (9) that

$$\Phi_{1-p,1}(t) = 1-(p-t)^+ = \begin{cases} 1-p+t & \text{if} \quad t \leq p \\ 1 & \text{if} \quad t > p \end{cases}. \tag{11}$$

The right-hand side of (10) equals

$$\int_0^1 G_X^{[n-1]}(t,z)d\Phi_{1-p,1}(t) = \int_0^p G_X^{[n-1]}(t,z)dt,$$

which is $G_X^{[n]}(p,z)$, the left-hand side. For $k=2$, we have

$$\Phi_{1-p,2}(t) = 1 - \int_0^{1-t} (u-1+p)^+ \, du. \tag{12}$$

By using the properties of the Riemann–Stieltjes integral, the right-hand side of (10) equals

$$\int_0^1 G_X^{[n-2]}(t,z) d\Phi_{1-p,2}(t) = \int_0^p (p-t) \, dG_X^{[n-1]}(t,z). \tag{13}$$

Taking account that $G_X^{[n-1]}(0,z) = 0$, integration by parts in (13) yields

$$\int_0^p G_X^{[n-1]}(t,z) \, dt,$$

the left-side hand. Let $k \geq 3$ and assume that the result holds for $k-1$. It follows from (8) and (9) that

$$\Phi_{1-p,k}(t) = 1 - \int_0^{1-t} \Psi_{1-p,k-1}(u) du$$

and, therefore, the right-hand side of (10) equals

$$\int_0^1 \Psi_{1-p,k-1}(1-t) dG_X^{[n-k+1]}(t,z). \tag{14}$$

Taking into account (9), $\Psi_{1-p,k-1}(0) = 0$ if $k \geq 3$ and $G_X^{[n-k+1]}(0,z) = 0$, integration by parts in (14) yields

$$\int_0^1 G_X^{[n-(k-1)]}(t,z) d\Phi_{1-p,k-1}(t),$$

which is $G_X^{[n]}(p,z)$ by applying the induction hypothesis. □

In the following result, we prove that, for each fixed $p \in [0,1]$, $G_X^{[n]}(p,z)$ belongs to C_n.

Theorem 5 *For each fixed $p \in [0,1]$ and $n \geq 1$, $G_F^{[n]}(p,z) \in C_n$.*

Proof. The proof consists in proving that

$$I_X(\Phi_{1-p,n}, z) = G_X^{[n]}(p,z) \tag{15}$$

holds for all $n = 1, 2, \ldots$ For $n = 1$, using (11), the left-hand side of (15) equals

$$\int_0^1 \left(z - F_z^{-1}(t)\right) d\Phi_{1-p,1}(t) = \int_0^p \left(z - F_z^{-1}(t)\right) dt,$$

which is $G_F^{[1]}(p,z)$, the right-hand side. For $n = 2$, using (12), the left-hand side of (15) equals

$$\int_0^1 \left(z - F_z^{-1}(t)\right) d\Phi_{1-p,2}(t) = \int_0^p (p-t)\, dG_X^{[1]}(t,z)$$

which is, using integration by parts,

$$\int_0^p G_X^{[1]}(t,z)\, dt,$$

that is, $G_X^{[2]}(p,z)$, the right-hand side. Let $n \geq 3$. Taking into account (9), the left-hand side of (15) equals

$$\int_0^1 \left(z - F_z^{-1}(t)\right) d\Phi_{1-p,n}(t) = \int_0^1 \Psi_{1-p,n-1}(1-t) dG_X^{[1]}(p,z).$$

Integration by parts and the facts that

$$\Psi_{1-p,n-1}(0) = 0 \text{ for } n \geq 3$$

and

$$G_X^{[1]}(0,z) = 0$$

yield

$$-\int_0^1 G_X^{[1]}(p,z) d\Psi_{1-p,n-1}(1-t) \tag{16}$$

or, equivalently, using again (9),

$$\int_0^1 G_X^{[1]}(p,z) d\Phi_{1-p,n-1}(t).$$

which is $G_X^{[n]}(p,z)$ by Lemma 4. □

It is well-known that $S_X(1,z)$ (the per-capita income gap) is $G_X^{[1]}(1,z)$ and $S_X(2,z)$ (the Thon-Chakravarty-Shorrocks indice) is two times the area underneath the curve $G_X^{[1]}(p,z)$. Now, we generalize these results by showing that $S_X(n,z)$ is $n!$ times the area underneath the curve $G_X^{[n-1]}(p,z)$ for all $n \geq 1$.

Theorem 6 *For all $n \geq 1$, we have $S_X(n,z) = n! G_X^{[n]}(1,z)$.*

Proof. From (10) we have

$$G_X^{[n]}(1,z) = \int_0^1 G_X^{[1]}(t,z) d\Phi_{0,n-1}(t). \tag{17}$$

It can be easily shown from (9) that

$$\Phi_{0,n-1}(t) = 1 - \frac{(1-t)^{n-1}}{n-1!}$$

Therefore, integration by parts in (17) yields

$$\frac{1}{n-1!} \int_0^1 (1-t)^{n-1} dG_X^{[1]}(t,z)$$

and this is the same as

$$\frac{1}{n!} \int_0^1 \left(z - F_z^{-1}(t)\right) d\Phi_n(t)$$

where $\Phi_n(t)$ is given by (6), which is $\frac{1}{n!} S_X(n,z)$. □

3. Characterizations

Now, we characterize the *n*th degree TIP curve dominance in terms of the class C_n defined in Section 2.

Theorem 7 *Let X and Y be two income random variables. For integers $n \geq 1$, we have*

$$X \geq_{TIP(n,z)} Y \text{ if and only if } I_X(\Phi,z) \geq I_Y(\Phi,z) \text{ for all } I \in C_n.$$

Proof. Necessary condition is immediate since $S_X(k,z) \in C_n$ for $k = 1,2,\ldots,n$ and, from Theorem 5, $G_X^{[n]}(p,z)$ also belongs to C_n for all $p \in [0,1]$. Therefore, by hypothesis, $S_X(k,z) \geq S_Y(k,z)$ for $k = 1,2,\ldots,n$ and $G_X^{[n]}(p,z) \geq G_Y^{[n]}(p,z)$, which means $X \geq_{TIP(n,z)} Y$.

In order to prove the sufficient condition suppose, firstly, that $X \geq_{TIP(1,z)} Y$ and take $I_X(\Phi,z) \in C_1$. Then, Φ is a concave distribution function on $[0,1]$ and there exists some non-negative, non-increasing and integrable function φ such that

$$\Phi(t) = \int_0^t \varphi(x) dx.$$

Therefore,

$$I_X(\Phi,z)=\int_0^1\left(z-F_z^{-1}(t)\right)d\Phi(t)=\int_0^1\varphi(t)\,dG_X^{[1]}(t,z).$$

Via integration by parts, we have

$$I_X(\Phi,z)=\varphi(1)\,S_X(1,z)-\int_0^1 G_X^{[1]}(t,z)d\varphi(t)\,. \tag{18}$$

Since

$$G_X^{[1]}(t,z)\geq G_Y^{[1]}(t,z) \text{ for all } t\in[0,1]$$

(in particular, $S_X(1,z)\geq S_Y(1,z)$),

$$\varphi(1)\geq 0 \text{ and } d\varphi(t)\leq 0,$$

it follows from (18) that $I_X(\Phi,z)\geq I_Y(\Phi,z)$.

Now, suppose $X\geq_{TIP(n,z)}Y$ and take $I(\Phi,z)\in C_n$, with $n\geq 2$. The first step consists in proving, by induction on n, that

$$I_X(\Phi,z)=\sum_{k=1}^{n-2}(-1)^{k+1}\Phi^k(1)\,S_X(k,z)-\int_0^1(-1)^{n-1}\Phi^{n-1}(t)\,dG_X^{[n-1]}(t,z). \tag{19}$$

For $n=2$, (19) is confirmed by using again the properties of the Riemann–Stieltjes integral:

$$I_X(\Phi,z)=\int_0^1\left(z-F_z^{-1}(t)\right)d\Phi(t)=\int_0^1\Phi'(t)\,dG_X^{[1]}(t,z),$$

which is the right-hand side of (19). Now suppose inductively that (19) holds for n and show the result holds for $n+1$. Let $I_X(\Phi,z)\in C_{n+1}$. Note, via integration by parts, that

$$\int_0^1\Phi^{n-1}(t)\,dG_X^{[n-1]}(t,z)=\Phi^{n-1}(1)\,S_X(n-1,z)-\int_0^1 G_X^{[n-1]}(t,z)d\Phi^{n-1}(t)$$

which is the same as

$$\Phi^{n-1}(1)\,S_X(n-1,z)-\int_0^1\Phi^n(t)\,dG_X^{[n]}(t,z). \tag{20}$$

Since $C_{n+1}\subset C_n$, by the induction hypothesis, $I_F(\Phi,z)$ satisfies (19) and by replacing (20) in (19) we obtain

$$I_X(\Phi,z)=\sum_{k=1}^{n-1}(-1)^{k+1}\Phi^k(1)S_X(k,z)-\int_0^1(-1)^n\Phi^n(t)\,dG_X^{[n]}(t,z)$$

as required. This proves that (19) holds for all $I_X(\Phi,z)\in C_n$, for all $n\geq 2$. Next, observe that, for $I_X(\Phi,z)\in C_n$, the function

$$\alpha(t)=(-1)^{n-1}\Phi^{n-1}(t) \tag{21}$$

is increasing and concave on $(0,1)$ and we can write

$$\alpha(t)=\alpha(1)-\int_t^1\mu(x)dx \tag{22}$$

where $\mu=\alpha'$ (almost everywhere) is non-negative and non-increasing. It is easy to see, by integration by parts, that (22) is the same as writing

$$\alpha(t)=\alpha(1)-\mu(1)(1-t)+\int_0^1(p-t)^+\,d\mu(p)\,. \tag{23}$$

Substitution of (21) into (23) yields

$$\begin{gathered}(-1)^{n-1}\Phi^{n-1}(t)=(-1)^{n-1}\Phi^{n-1}(1)+(-1)^n\Phi^n(1)(1-t)+\\(-1)^{n+1}\int_0^p(p-t)\,d\Phi^n(p)\,.\end{gathered} \tag{24}$$

By substituting (24) into (19) and rearranging terms we have that

$$\begin{gathered}I_X(\Phi,z)=\sum_{k=1}^{n-1}(-1)^{k+1}\Phi^k(1)S_X(k,z)+\\(-1)^{n+1}\Phi^n(1)\int_0^1(1-t)\,dG_X^{[n-1]}(t,z)+\\(-1)^n\int_0^1\int_0^p(p-t)\,dG_X^{[n-1]}(t,z)d\Phi^n(p)\,.\end{gathered} \tag{25}$$

(Fubini's Theorem has been applied in the last term). Since

$$\int_0^p(p-t)\,dG_X^{[n-1]}(t,z)=G_X^{[n]}(p,z)$$

and, consequently,

$$\int_0^1(1-t)\,dG_X^{[n-1]}(t,z)=S_X(n,z),$$

(25) can be rewritten as follows:

$$I_X(\Phi,z)=\sum_{k=1}^{n}(-1)^{k+1}\Phi^k(1)S_X(k,z)+(-1)^n\int_0^1 G_X^{[n]}(p,z)d\Phi^n(p). \tag{26}$$

We complete the proof by noting that

$$(-1)^{k+1}\Phi^k(1)\geq 0 \text{ for } k=1,\ldots,n,$$

$$S_X(k,z)\geq S_Y(k,z), \text{ for } k=1,\ldots,n,$$

$$G_X^{[n]}(p,z)\geq G_Y^{[n]}(p,z) \text{ for all } p\in[0,1]$$

and

$$(-1)^n d\Phi^n(p)\geq 0. \qquad \square$$

If we restrict attention to the class C_n^*, then a comparison of nTIP curves is enough to obtain a characterization. The proof of the next result follows easily from (26).

Corollary 8 *Let X and Y be two income random variables. For integers $n\geq 1$, we have*

$$G_X^{[n]}(p,z)\geq G_Y^{[n]}(p,z) \textit{ for all } p\in[0,1]$$

if and only if

$$I_X(\Phi,z)\geq I_Y(\Phi,z) \textit{ for all } I\in C_n^*.$$

4. An example

It is well-known that empirical income distribution data fit well to lognormal form (see, for example, Harrison (1981) and Cowell (1999)). Moreover, the use of the lognormal model is “probably the most standard approximation of empirical data distributions in the applied literature” (Bourguignon, 2003, page 11). See Lambert (2009) and references therein for applications of this model in poverty analysis. Recall that a lognormal random variable X has a density function of the form

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}, \quad x>0, \quad \sigma>0,\ \mu\in\mathbb{R}$$

and the mean and the standard deviation are given, respectively, by $E\left[X\right] = \exp\left\{\mu + \frac{\sigma^2}{2}\right\}$ and $SD(X) = \sqrt{\left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}}$.

In order to illustrate the applicability of the comparison method proposed in this paper, we have simulated two samples with sizes $n = m = 100$ from two underlying lognormal distributions X and Y, with respective means $E\left[X\right] = 9030$ and $E\left[Y\right] = 9010$ and standard deviations $SD(X) = SD(Y) = 3100$. The reference poverty line is set at $z = 6500$ (since $F_X(z) = 0.21$ and $F_Y(z) = 0.19$, this choice appears to be a reasonable poverty line for poverty comparisons between these models[2]).

In order to compare the poverty associated to these income distributions, we start by comparing the corresponding "per-capita income gaps" $S_X\left(1,z\right)$ and $S_Y\left(1,z\right)$, which represent the sum of the poverty gaps of the poor. The evaluation of these indices with DAD 4.5 (a programme freely distributed by Duclos *et al.* (2006), designed to facilitate the analysis of social welfare, inequality and poverty), gives $S_X\left(1,z\right) = 236.67$ and $S_Y\left(1,z\right) = 234.57$ and we can say that poverty, as measured by this index, is greater in X than in Y. However, "any choice of a single measure is apt to be arbitrary" (Foster, 1984, page 242), and different choices may produce different conclusions. We reduce this arbitrariness by considering a broader class of poverty measures than $S\left(1,z\right)$, given by

$$C_1 = \left\{I\left(\Phi,z\right) \text{ of the form (2) such that } \Phi \text{ is concave}\right\}.$$

Each member of C_1 is interpreted as a weighted sum of the poverty gaps of the poor. Obviously, it is impossible to check poverty orderings for all measures in C_1 and we prefer to plot the corresponding TIP curves $G_X^{[1]}\left(p,z\right)$ and $G_X^{[1]}\left(p,z\right)$. Following Duclos and Araar (2006, Section 10.1), non-intersection of these curves is equivalent to the unanimous ordering generated by the class C_1. Unfortunately, Figure 1 shows that the TIP curves cross twice (the first is at around $p = 0,11$ and the second is at around $p = 0,19$), therefore the inequality $I_X\left(\Phi,z\right) \leq I_Y\left(\Phi,z\right)$ fails to be satisfied for some member of C_1. In other words, the comparison between X and Y in terms of poverty measures in C_1 is ambiguous.

Fortunately, as we have shown in Section 3, an unambiguous ordering between X and Y is still possible by focussing on a subclass of C_1 and moving from the first degree TIP ordering to the second degree TIP ordering (and, more generally, to the n-degree TIP ordering, $n \geq 2$). The second degree TIP ordering requires the evaluation of $S_X\left(k,z\right)$ and $S_Y\left(k,z\right)$ for $k = 1,2$, and the comparisons of the curves $G_X^{[2]}\left(p,z\right)$ and $G_Y^{[2]}\left(p,z\right)$. The evaluation of $S_X\left(2,z\right)$ and $S_Y\left(2,z\right)$ with DAD 4.5 gives $S_X\left(2,z\right) = 444.75$ and $S_Y\left(2,z\right) = 439.71$. Therefore, we have

$$S_X\left(1,z\right) > S_Y\left(1,z\right) \text{ and } S_X\left(2,z\right) > S_Y\left(2,z\right).$$

[2] In Spain, for example, the percentage of persons below the poverty line is 19.6% (Quality of Life Survey, 2008, I.N.E.)

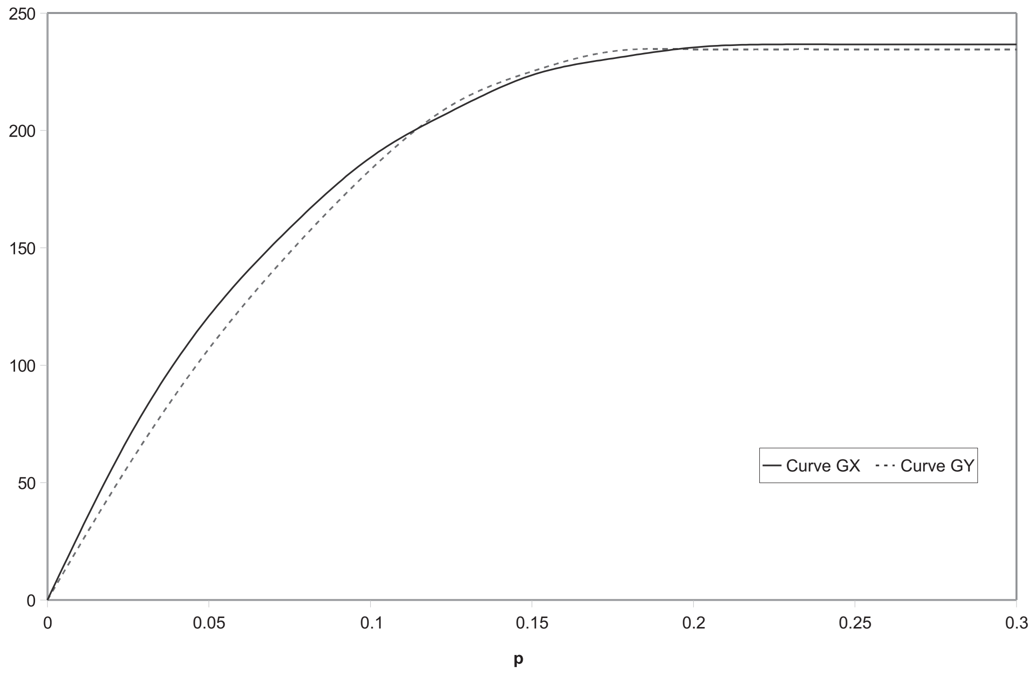


Figure 1: $(p, G^{[1]}(p,z))$

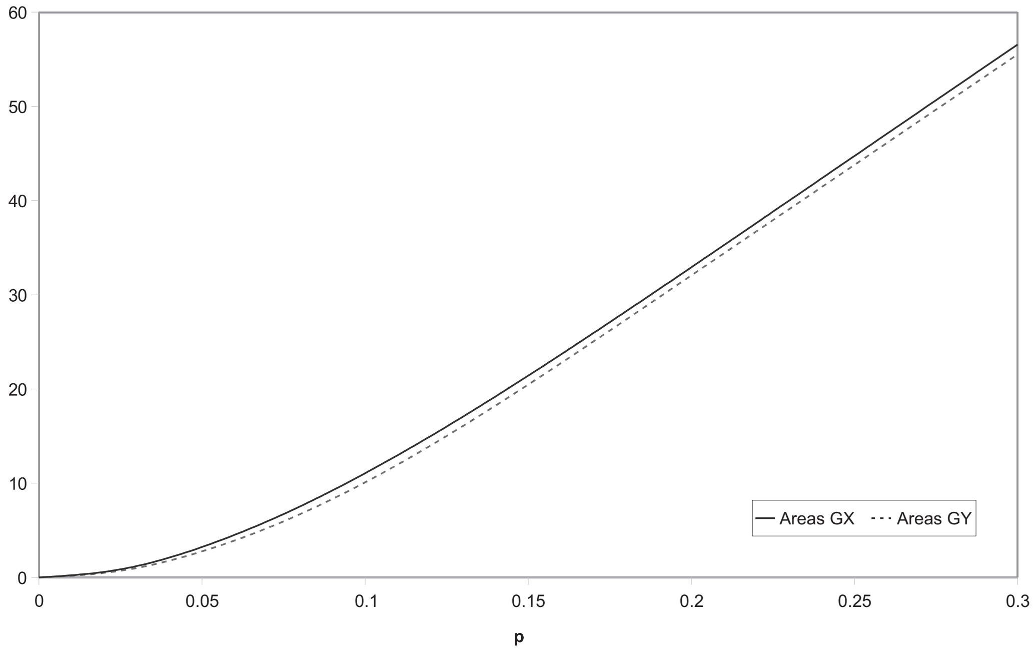


Figure 2: $(p, G^{[2]}(p,z))$

Moreover, Figure 2 shows that $G_X^{[2]}(p,z)$ is above $G_Y^{[2]}(p,z)$ for all p in $(0,1)$ (the curves are plotted up to $p = 0.3$; this is due to the fact that the second cross between the TIP curves is at around $p = 0.19$; therefore, from this p on, $G_X^{[2]}(p,z)$ is everywhere above $G_Y^{[2]}(p,z)$). Thus, $X \geq_{TIP(2,z)} Y$ holds and from Theorem 7 it follows that $I_X(\Phi,z) \geq I_Y(\Phi,z)$ for all measures in

$$C_2 = \left\{ I(\Phi,z) \in C_1 \text{ such that } \Phi' \text{ is convex} \right\}.$$

We conclude this illustration by noting that increasing the degree of dominance (moving from the first degree TIP ordering to the second degree TIP ordering) makes poverty in X unambiguously larger than in Y. Since any index of C_2 not belonging to C_1 is more sensitive to the distribution of income among the poorest, this is equivalent to saying that poverty in X is unambiguously larger than in Y when sufficient weight is given to the effect of income changes among the bottom of the distribution.

5. Final remarks

In this paper, we have tried to advance in obtaining comparable poverty results when TIP curves intersect, by considering a sequence of dominance criteria (the *n*degree TIP curve dominance) based on TIP areas and *S*-indices. The normative meaning of these criteria has been provided in terms of a class C_n of linear rank-based poverty measures with the property that, the larger the value of n, the greater the weight assigned to the effect of income changes among the bottom of the distribution.

Duclos and Grégoire (2002) have shown that the properties of the *S*-indices compare rather well with those of the FGT (Foster *et al.*, 1984) additive indices. Some results in this work confirm this conclusion. Given an income distribution F, a poverty line z and a non-negative integer α, the $FGT(\alpha)$ index is defined by

$$FGT_\alpha(F,z) = \int_0^1 (z-x)^\alpha \, dF(x).$$

Foster and Shorrocks (1998a) note that

$$FGT_\alpha(F,z) = \alpha! F_{\alpha+1}(z) \text{ for all } z, \tag{27}$$

where $F_1(x) = F(x)$, $F_k(x) = \int_0^x F_{k-1}(t)\,dt$, $k = 1,2,..$ and provide the following link between the poverty order induced by$FGT(\alpha)$ for all $z \in (0,\infty)$ and the $(\alpha+1)$th degree stochastic dominance:

$$FGT_\alpha(F,z) \geq FGT_\alpha(H,z) \ \forall\, z \in (0,\infty) \Longleftrightarrow F_{\alpha+1}(z) \geq H_{\alpha+1}(z) \forall\, z \in (0,\infty)$$

The relation

$$S_X(n,z) = n!G_X^{[n]}(1,z) \tag{28}$$

stated in Theorem 6 is somewhat similar to (27) and suggests that the role played by $S_X(n,z)$ in the dual approach (in the sense of Duclos and Araar, 2006) is as important as the role of FGT indices in the primal one. The characterization

$$S_X(n,z) \geq S_Y(n,z) \forall\, z \in (0,\infty) \Longleftrightarrow G_X^{[n]}(1,z) \geq G_Y^{[n]}(1,z) \forall\, z \in (0,\infty)$$

(which follows from Theorem 6) shows that X has unambiguously more poverty that Y with respect to the poverty measure $S(n,z)$ for all $z \in (0,\infty)$ if, and only if, the area underneath the curve $G_X^{[n-1]}(p,z)$ is bigger than the area underneath $G_Y^{[n-1]}(p,z)$ for all $z \in (0,1)$. (28) also reveals an important interrelationship among poverty-line orderings by different members of S. Since

$$G_X^{[n]}(1,z) \geq G_Y^{[n]}(1,z) \Longrightarrow G_X^{[k]}(1,z) \geq G_Y^{[k]}(1,z) \text{ for } k \geq n$$

it follows from (28) that

$$S_X(n,z) \geq S_Y(n,z) \forall z \Longrightarrow S_X(k,z) \geq S_Y(k,z) \forall z, \forall k \geq n.$$

Aknowledgement

We thank the referee for providing constructive suggestions, which helped to improve the final version.

References

Bourguignon, F. (2003). The growth elasticity of poverty reduction: explaining heterogeneity across countries and time periods. Chapter 1, in Teicher, T. S. and S. J. Turnovsky (eds). *Inequality and Growth: Theory and Policy implications*. Cambridge MA: MIT Press.

Chakravarty, S. (1983). A new index of poverty. *Mathematical Social Sciences*, 6, 307–313.

Cowell, F.A. (1999). *Measurement of Inequality*. In Atkinson, A.B. and Bourguignon F. (eds). Handbook of income distributions. North-Holland, Amsterdam.

Davidson, R., Duclos, J. (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica*, 68, 1435–1464.

Duclos, J. (2000). Gini indices and the redistribution of income. *International Tax and Public Finance*, 7, 141–162.

Duclos, J., Grégoire, P. (2002). Absolute and relative deprivation and the measurement of poverty. *Review of Income and Wealth*, 48, 471–492.

Duclos J., Araar, A. (2006). *Poverty and Equity: Measurement, Policy and Estimation with DAD*. Springer. New York.

Duclos, J., Araar, A. and Fortin, C. (2006). DAD: a software for Distributive Analysis / Analyse Distributive, MIMAP programme, International Development Research Centre, Government of Canada, and CIRPÉE, Université Laval.

Foster, J.E. (1984). On economic poverty: a survey of aggregate measures. In R.L. Basmann and G. F. Rhodes (eds). *Advances in Econometrics*, 3. Connecticut: JAI Press, 215–251.

Foster, J.E., Greer, J., Thorbecke, E. (1984). Notes and comments. A class of decomposable poverty measures. *Econometrica*, 52, 761–766.

Foster, J.E., Shorrocks, A.F. (1988). Notes and comments. Poverty orderings. *Econometrica*, 56, 173–177.

Foster, J.E., Shorrocks, A.F. (1988). Poverty orderings and welfare dominance. *Social Choice and Welfare*, 5, 179–198.

Harrison, A.J. (1981). Earning by sizes: a tale of two distributions. *Review of Economic Studies*, 48, 621–631.

Jenkins, S.P., Lambert, P.J. (1993). Ranking income distributions when needs differ. *Review of Income and Wealth*, 39, 337–356.

Jenkins, S.P., Lambert, P.J. (1997). Three I's of poverty curves, with an analysis of UK poverty trends. *Oxford Economics Papers*, 49, 317–327.

Jenkins, S.P., Lambert, P.J. (1998). Ranking poverty gap distributions: further tips for poverty analysis. *Research on Economic Inequality*, 8, 31–38.

Jenkins, S.P., Lambert, P.J. (1998). Three I's of poverty curves and poverty dominance: tips for poverty analysis. *Research on Economic Inequality*, 8, 39–56.

Lambert, P.J. (2009). Pro-poor growth and the lognormal income distribution. Working Paper No. 2009-130, ECINEQ, Milan.

Mehran, F. (1976). Linear measures of income inequality. *Econometrica*, 44, 805–809.

Sen, A. (1976). Poverty: an ordinal approach to measurement. *Econometrica*, 44, 219–231.

Shorrocks, A.F. (1995). Notes and comments. Revisiting the Sen poverty index. *Econometrica*, 63, 1225–1230.

Shorrocks, A.F. (1998). Deprivation profiles and deprivation indices. Ch. 11 in *The Distribution of Household Welfare and Household Production*, ed. S. Jenkins *et al.*, Cambridge University Press.

Shorrocks, A., Foster, J. (1987). Transfer sensitive inequality measures. *Review of Economic Studies*, 54, 485–497.

Sordo, M.A., Ramos, C.D., Ramos, H.M. (2007). Poverty measures and poverty orderings. *SORT*, 31, 169–180.

Spencer, B.D., Fisher, S. (1992). On comparing distributions of poverty gaps. *Sankhya: The Indian Journal of Statistics*, Series B 54, 114–126.

Thon, D. (1979). On measuring poverty. *Review of Income and Wealth*, 25, 429–440.

Thon, D. (1983). A poverty measure. *Indian Economic Journal*, 30, 55–70.

Yaari, M.E. (1988). A controversial proposal concerning inequality measurement. *Journal of Economic Theory*, 44, 381–397.

Zheng, B. (1999). On the power of poverty orderings. *Social Choice and Welfare*, 16, 349–371.

Zheng, B. (2000). Poverty orderings. *Journal of Economics Surveys*, 14, 427–466.

Information for authors and subscribers

Information for authors and subscribers

Submitting articles to SORT

Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.cat) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a $\LaTeX\,2_{\varepsilon}$.

In any case, upon request the journal secretary will provide authors with $\LaTeX\,2_{\varepsilon}$ templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

Publishing rights and authors' opinions

The publishing rights for SORT belong to the Institut d'Estadística de Catalunya since 1992 through express concession by the Technical University of Catalonia. The journal's current legal registry can be found under the reference number DL B-46.085-1977 and its ISSN is 1696-2281. Partial or total reproduction of this publication is expressly forbidden, as is any manual, mechanical, electronic or other type of storage or transmission without prior written authorisation from the Institut d'Estadística de Catalunya.

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following examples:

Citations
Mahalanobis (1936), Rao (1982b)

Journal articles
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9, 73-84.

Books
Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Parts of books
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Web files or "pages"
Nielsen, S. F. (2001). *Proper and improper multiple imputation* http://www.stat.ku.dk/˜feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

SORT *(Statistics and Operations Research Transactions)*
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.cat

How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

Subscription form
SORT *(Statistics and Operations Research Transactions)*

Name ____

____

Organisation ____

____

Street Address ____

____

Zip/Postal code ____________________ City ____________________

State/Country ____________________ Tel. ____________________

Fax ______________ NIF/VAT Registration Number ______________

E-mail ____

Date ____________________

Signature

I wish to subscribe to **SORT** ***(Statistics and Operations Research Transactions)*** for the year 2011 (volume 35)

Annual subscription rates:

– Spain: €22 (4% VAT included)
– Other countries: €25 (4% VAT included)

Price for individual issues (current and back issues):

– Spain: €15/issue (4% VAT included)
– Other countries: €17/issue (4% VAT included)

Method of payment:

□ Bank transfer to account number 2013-0100-53-0200698577

□ Automatic bank withdrawal from the following account number
□□□□ □□□□ □□ □□□□□□□□□□

□ Check made payable to the Institut d´Estadı́stica de Catalunya

Please send this subscription form (or a photocopy) to:

SORT ***(Statistics and Operations Research Transactions)***
Institut d´Estadı́stica de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

Bank copy

Authorisation for automatic bank withdrawal in payment for

SORT ***(Statistics and Operations Research Transactions)***

The undersigned ____

authorises Bank/Financial institution ______________________________

located at (Street Address) ________________________________

Zip/postal code ____________________ City ____________________

Country ____

to draft the subscription to **SORT** ***(Statistics and Operations Research Transactions)*** from my account

number □□□□ □□□□ □□ □□□□□□□□□□

Date ______________________

Signature

SORT ***(Statistics and Operations Research Transactions)***
Institut d´Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45