FEP WORKING PAPERS

FEP WORKING PAPERS

Research Work
in Progress

n. 511 October 2013

ISSN: 0870-8541

A Mixed Micro-Macro Approach to Statistical Disclosure Control for Macrodata

Cristina Matias [1]
Pedro Campos [1,2]

[1] FEP-UP, School of Economics and Management, University of Porto
[2] LIAAD/INESC TEC

A MIXED MICRO-MACRO APPROACH TO STATISTICAL DISCLOSURE CONTROL FOR MACRODATA

Cristina Matias[1], Pedro Campos[2]

1. Faculty of Economics, University of Porto, Portugal,
cristinalobogrmatias@gmail.com
2. LIAAD INESC-TEC and Faculty of Economics, University of Porto, Portugal
pcampos@fep.up.pt

Abstract. National Statistics Offices, Central Banks, and any other organisms and agencies producing statistical information, disseminate data so that the individual information is sufficiently protected. At the same time, those entities aim at providing society with as much information as possible under this restriction. There is some contradiction between these two purposes, since high utility information is not always possible if one has to ensure data security against unauthorized accesses. Post-tabular techniques generate safe tables through non perturbative methods (such as cell suppression) or perturbative methods (such as rounding). Despite its effectiveness, these techniques prevent users from making a more detailed statistical analysis since the published data doesn't have the desired similarity with real values. For instance, cell suppression hides non-sensitive cells, leading to higher losses of information while perturbative methods may conceal the reality. In this paper we propose a new post-tabular perturbative method which applies mathematical restrictions directly on the respondents within each sensitive cell and computes safe values. Since this method focuses on respondents, it is possible to identify sensitive cells that don't represent disclosure risk. The comparative study between this technique and others commonly used, shows significant improvements in the data utility, keeping a low risk level.

Keywords: Confidentiality, Statistical Disclosure Control, Tabular data, Mixed Micro-Macro Approach, M3A.

1 INTRODUCTION

In the literature ([17], [5], [2]), three major reasons are essentially indicated to ensure data confidentiality: (i) Legislation that requires that statistical data are strictly confidential and used exclusively for statistical purposes, as is evidenced, for example, by Principle 6 of UNECE (United Nations Economic Commission for Europe) and law n .º 22/2008 of Portuguese Parliament; (ii) Trust of information providers (the respondents), a fundamental guarantee for the high quality and detail of information provided when collected by the Statistics authorities; (iii) Ethics in statistical profession, which is reflected in the European Statistics Code of Practice [1], including the principles of professional independence , impartiality and objectivity in the collection of statistical information .

One of the main concerns of Statistical Agencies is to disseminate tabular data with high utility ensuring data security against unauthorized accesses. Finding the best trade-off between utility and risk is one of the goals of Statistical Disclosure Control (SDC) techniques. Whether they are magnitude or frequency tables, Statistical Agencies have to follow the European Statistics Code of Practice [1] which highlights the need and importance of data dissemination with a high level of utility and a low risk level, i.e., the respondents' privacy should never be endangered.

A SDC technique is a three-step process which aims to generate a safe table to publish: first, sensitive cells of the original table are identified through a sensitive rule then, all the desired SDC techniques are applied to the original table, generating a set of possible tables to publish and finally, all the purposed tables are compared according to its risk and utility levels. The one with best results is chosen.

The two main approaches for SDC techniques are perturbative techniques, such as Rounding and Control Tabular Adjustment [2], which add noise to the table and modify some values, and non-perturbative techniques, such as Cell Suppression and Redesign [3], [4], which change the structure of the table or suppress information.

Cell suppression and rounding are very popular techniques within these two approaches. However, in some cases, they produce low-quality tables for a deeper data analysis. For instance, cell suppression has a huge impact in the released table since it hides both sensitive and non-sensitive cells, implying a great loss of information and jeopardizes the statistical analysis. The same happens with rounding techniques since they may generate very distinct data from the original, leading to misinterpretation of reality.

For these reasons we propose a technique focused in providing a table with a high level of utility, allowing users to perceive data, very close to reality, through statistical analysis. The Mixed Micro-Macro Approach (M3A) is a perturbative technique that modifies the original data through mathematical restrictions acting on the underlying respondents of each sensitive cell. The main idea is to directly protect the respondents (and not the cells), ensuring that an estimate of any intruder on the value of a contribution is, at least, at a minimum distance of d% from the original value. Compared to Cell Suppression and Rounding, M3A also provides tables with low risk level but with higher data quality/utility results.

The paper is structured as follows: Section 2. provides an overview of SDC processes and techniques for macrodata; Section 3 describes the M3A; Section 4 contains the evaluation of M3A and compares it with other macrodata techniques; Section 5 presents the main conclusions and topics for further work.

2 Statistical Disclosure Control for Macrodata

Many Statistical agencies operate or envision tools for ad hoc creation and visualization of aggregate tables [23]. The notion of safe data for macrodata relies on ensuring that all the published cells satisfy a safety requirement, given by the sensitive measure and applied method to the original table [5], [6] in order to prevent intruders' estimations.

Concerning macrodata protection, there are two types of SDC techniques: pre-tabular and post-tabular. The former is applied to microdata before aggregation and in those cases, microdata techniques are used. In the latter, data is protected after the table is created. Some research has been done recently, however, combining these two approaches. Giessing [23] proposes a new method based on an idea for post-tabular stochastic noise. The method proved to give encouraging results when tested on tabulations of German business tax statistics.

In this paper, we also apply and discuss post-tabular techniques, since M3A protects the data after knowing the table structure. To give a deeper understanding of this technique, we assume the most common table structure used by Statistical Agencies: a two-dimensional table (Definition 1), which has aggregated information about respondents.

Definition 1 – TWO-DIMENSIONAL TABLES

Let $T_{\ell,\sigma}$ be a table with two dimensions with $i = 1, \dots, \ell$ lines and $j = 1, \dots, \sigma$ columns, composed by $z = 1, \dots, \mathcal{C}$ cells represented by x_{ij}. The corresponding marginal totals of lines and columns is given by $x_{i.} = \sum_{j=1}^{\sigma} x_{ij}\ \forall\ i \in [1, \dots, \ell]$ and $x_{.j} = \sum_{i=1}^{\ell} x_{ij}, \forall\ j \in [1, \dots, \sigma]$, and $x_{..} = \sum_{j=1}^{\sigma} \sum_{i=1}^{\ell} x_{ij}$ defines the grand total which obeys to the consistency condition, $\sum_{i=1}^{\ell} x_{i.} = \sum_{j=1}^{\sigma} x_{.j} = \sum_{j=1}^{\sigma} \sum_{i=1}^{\ell} x_{ij}$.

A table is known as frequency table when its cells contain absolute or relative frequencies and as magnitude table when it provides a sum of quantitative variables of all respondents' contributions (Definition 2).

Definition 2 – RESPONDENTS CONTRIBUTIONS

Denote by $y_1, y_2, \dots, y_m$, the corresponding contribution of the respondents 1,2,..., m, for each cell x_z, where $y_1 \geq y_2 \geq \cdots \geq y_m$ and by $Y_z = \sum_{i=1}^{m} y_i$ the sum of contributions. Therefore, if Y_z reflects the total value of a cell, then:

$$Y_z = \sum_{i=1}^{m} y_i \overset{\text{def}}{=} x_z \qquad (1)$$

Each cell z contains $\mathcal{M} = 1, \dots, n, n+1, \dots, m$ respondents, where $\{1, \dots, n\}$ represents the n respondents with greater contribution.

Besides macrodata tables release aggregated information, some of their cells may represent risk to the respondents since an intruder may use the published values to achieve good estimations or derivate the respondents' contribution. Those cells are known as sensitive cells, denoted by $\mathscr{s}$, and are identified through sensitive rules.

The next sections present the three phases of a SDC process: risk evaluation through sensitive rules, table protection through SDC techniques and evaluation of the proposed tables to publish.

2.1 Sensitive rules

Sensitive rules are selected based on the table's data type and on the Statistical Agency intuition about variables and their assumptions about the public knowledge [3]. The commonly used rules are minimum frequency rule, (n, k) rule and the pq rule.

Minimum Frequency Rule.

The minimum frequency rule considers safe all the cells that have a minimum frequency of r respondents. Usually, r=3 [5], [6].

(n, k) rule.

Using this rule, a cell is identified as sensitive when the sum of the n higher contributions exceed $k\%$ of the cell's total, i.e., $\mathrm{y}_1 + \cdots + \mathrm{y}_n > k/100 \times \mathrm{Y}$, $0 < k < 100$, [6],[7]. Usually, the dominance rule parameters are $2 \leq n < 5$ and $k > 60$ [7].

(pq) rule.

The pq rule (priori-posteriori rule) uses two parameters p and q, with $p < q$, 0< p <100 and $0 < q$ <100, where it is assumed that before the table disclosure, any respondent's contribution may be estimated with $q\%$ of precision. A cell is considered sensitive if, after the disclosure, a respondent may estimate the contribution of another one with $p\%$ of precision [8].

2.2 SDC techniques for macrodata

SDC techniques usually safeguard the data confidentiality while providing data quality. Those techniques are classified as non-perturbative techniques, when they don't modify the data (Redesign and Cell Suppression), and as perturbative techniques, when they modify the data

(Rounding and Control Tabular Adjustment). In this paper, it will only be discussed Cell Suppression and Rounding techniques.

Rounding.

Rounding techniques change the original data by multiplying cell values by a rounding base b which is usually equal to 1, 3, 5 or 10 [9] or using better approach [5]:

- Chosen as being, at least, $J\%$ of the maximum value of a sensitive cell.
- Chosen as being, at least, $J\%$ of the higher contribution of a sensitive cell.
- Using the parameters of a dominance rule for sensitive cells, where the minimum value for b is given by $max_{\mathscr{s}}\{100/\mathscr{k} \sum_{t=1}^{n} y_t - \sum_{t=1}^{m} y_t\}$.

There are several types of rounding techniques. Here, we will present only the most comprehensive: conventional rounding, random rounding and controlled rounding.

Conventional rounding is a technique that rounds each internal cell and the marginal totals to the multiple nearest base b, implying that the table's additivity property is not guaranteed and the information released is poorly consistent.

Random rounding is similar to conventional rounding but here, cell values are rounded according to a probabilistic system defined as follow: Be Y_z the original value of cell z, which may be written by $\mathrm{Y}_z = (q_z + r_z) \times b$, where q_z is the quotient of dividing Y_z by b and r_z the rest of the division, so, for $0 \leq r < b$, q integer and b the base value, Y_z may be rounded up $\lceil \mathrm{Y}_z \rceil$ or down $\lfloor \mathrm{Y}_z \rfloor$, through the probability scheme (2) (deduced from [6]).

$$\begin{cases} \mathrm{P} \quad Y_z\rceil \equiv (q_z + 1)b) = (Y_z - q_z b)/b \\ \mathrm{I} \quad Y_z\rfloor \equiv q_z b) = 1 - (Y_z - q_z b)/b \end{cases} \qquad (2)$$

The controlled rounding [10], [11], [12] is a technique which keeps the consistency between the internal cell and the additivity relations of the table by using Linear Programming which identifies controlled rounding patterns of cells. The rounding pattern will minimize the information loss, defined by $\mathrm{Min}\left(\sum_{z=1}^{\mathcal{C}} |\mathrm{x}_z - [\mathrm{x}_z]|\right)$, where x_z is the original value of the cell and $[\mathrm{x}_z]$ its rounding value.

Cell Suppression

In cell suppression, some information is omitted and replaced by a symbol [13], [14]. With this technique, all the sensitive cells are suppressed (Primary Suppression, PS) and the safety of the table is supported by the additional suppression of non-sensitive cells (Secondary Suppression, SS) that will prevent the estimation of confidential information through the marginal totals published.

The main challenge of cell suppression technique is how to find the optimal suppression pattern (set of suppressed cells, SUP), which will depend on the SS. The suppression pattern will define the protection intervals to each sensitive cell, ensuring that the limits of the intervals are at a safe distance from the original value. To find those limits is necessary to solve Equation (3) [15].

$$\left.\begin{matrix} Min \\ Max \end{matrix}\right\} x_i, \text{such that}\, Ax = b, x \geq 0, \forall\, i \in SUP \qquad (3)$$

An interval is considered safe when, for a set of sensitive cells $\mathcal{s}$, verify that [15].

$$\begin{cases} \underline{x_{\mathcal{s}}} \quad x_{\mathcal{s}} - NPMin_{\mathcal{s}} \leq x_{\mathcal{s}} \leq x_{\mathcal{s}} + NPMax_{\mathcal{s}} \leq \overline{x_{\mathcal{s}}} \\ \qquad \overline{x_{\mathcal{s}}} - \underline{x_{\mathcal{s}}} \geq DNP_{\mathcal{s}} \end{cases} \qquad (4)$$

Where $\mathrm{NPMin}_{\mathcal{s}}$, $\mathrm{NPMax}_{\mathcal{s}}$ and $\mathrm{DNP}_{\mathcal{s}}$ are, respectively, the minimum and maximum protection levels and the protection level deviation that ensures that the interval isn't too short [16]. $\overline{\mathrm{x}_{\mathcal{s}}}$ and $\underline{\mathrm{x}_{\mathcal{s}}}$ are the maximum and minimum limits obtained solving (3).

To solve the Secondary Cell Suppression Problem is necessary to solve a Mixed Integer Programming Problem (MILP) which objective is to minimize the loss information subject to restriction (4). Since this is a NP-Hard problem [17], it should be used heuristics to find solutions near to the optimal.

The most common cell suppression heuristics are the Hypercube and HiTaS.

Hypercube [18] starts by subdividing a n-dimensional hierarchical table in sub-tables without substructure [6]. These sub-tables are protected successively by an iterative process that starts at the highest level. Then, for each primary suppression in the current sub-table, are built all the possible hypercubes with one of the cells of PS in the corner. For each hypercube, a lower limit is calculated according to the interval obtained with the suppression of the four corners of the hypercube and, for each interval, the loss of information is analyzed. The Hypercube with less loss of information is selected and the corners eliminated.

HiTaS uses a top-down approach, where a tree of two-dimensional sub-tables is built. In this methodology, the primary and secondary suppressions are computed, first, for the table-base (highest level) which will be reflected, subsequently, in the marginal totals of the sub-tables at a lower level. These marginal totals are fixed for the calculation of the secondary suppression, ensuring that the processing of sub-tables doesn't change what was already determined in the table at the previous level.

2.3 Evaluation of proposed tables

Once the Statistical Agency finishes the treatment of the tabular data, the proposed tables are represented in a R-U map that represents the trade-off between risk (R) and utility (U). The choice of the best table to disclose is defined by the partial order $\preccurlyeq_{\mathrm{RU}}$:

$$\mathbf{T_1} \preccurlyeq_{RU} \mathbf{T_2} \Leftrightarrow R(\mathbf{T_2}) < R(\mathbf{T_1}) \text{ and } U(\mathbf{T_2}) > U(\mathbf{T_1}) \qquad (5)$$

$\mathrm{T_1}$ is preferable to $\mathrm{T_2}$ whenever its risk is lower and its utility higher than $\mathrm{T_2}$.

In the next sub-sections we will describe the loss and risk measures. Resuming the table notion given in Definition 1, we defined the following notation: $\mathrm{T_{orig}}$ represents the original table and $\mathrm{T_{dis}}$ the disclosed table, each one with the same number of columns ($\mathscr{c}$) and lines (ℓ), then i = 1,…,ℓ and j = 1,…, $\mathscr{c}$, $\mathrm{T^{i}_{tab}}$ and $\mathrm{T^{j}_{tab}}$, $\mathrm{tab} = \{\mathrm{orig}, \mathrm{dis}\}$ represents all the lines i and all the columns j of $\mathrm{T_{tab}}$. Likewise, $\mathrm{T^{i}_{tab}(c)}$ and $\mathrm{T^{j}_{tab}(c)}$, $\mathrm{tab} = \{\mathrm{orig}, \mathrm{dis}\}$ denote the cell c of each line i and column j.

Some of the following metrics analyze the tables line by line but it's possible do it column by column through a simple adjustment in the formulas.

Measuring the distribution distortion.

The common used metrics, to measure the data distortion between the original table and the table proposed to publish, are the absolute difference between the grand total published and the original one (absolute distance), the sum of difference between each released cell and its respective original value (absolute distance per cell) and the Hellinger distance.

To measure the uncertainty, it may be also used an entropy analysis [19] between the original and the published table. The higher the value of entropy, the greater the level of uncertainty reported in the table. The measure of information loss is given by:

$$100 \times \frac{\sum_{\mathrm{k}=1}^{\ell} \mathrm{H}\left(\mathrm{T}_{\mathrm{div}}^{\mathrm{k}}\right) - \sum_{\mathrm{k}=1}^{\ell} \mathrm{H}\left(\mathrm{T}_{\mathrm{orig}}^{\mathrm{k}}\right)}{\sum_{\mathrm{k}=1}^{\ell} \mathrm{H}\left(\mathrm{T}_{\mathrm{orig}}^{\mathrm{k}}\right)} \qquad (6)$$

where, $\mathrm{H}\left(\mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}\right) = \frac{\mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}}{\sum_{\mathrm{c}\in\mathrm{k}} \mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}} \log\left(\frac{\mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}}{\sum_{\mathrm{c}\in\mathrm{k}} \mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}}\right), \mathrm{tab} = \{\mathrm{dis}, \mathrm{orig}\}$

Impact on variance

Similarly to distortion measures, it is also possible to compute the impact on variance for rows or columns of the tables.

Define $\mathcal{C}_{\mathrm{k}}$ as the number of cells contained in a line/column, and by $\mathrm{V}(\mathrm{T}_{\mathrm{tab}}) = \frac{1}{\mathcal{C}_{\mathrm{k}}-1} \sum_{\mathrm{c}\in\mathrm{k}} \left(\mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}(\mathrm{c}) - \left(\sum_{\mathrm{c}\in\mathrm{k}} \mathrm{T}_{\mathrm{tab}}^{\mathrm{k}}(\mathrm{c})/\mathcal{C}_{\mathrm{k}}\right)\right)^2$, $\mathrm{tab} = \{\mathrm{orig}, \mathrm{dis}\}$, the variance of the table. The loss information measure – relative variance - is given by (7).

$$100\times \frac{\sum_{k=1}^{\ell} V\left(\boldsymbol{T}_{div}^{k}\right) - \sum_{k=1}^{\ell} V\left(\boldsymbol{T}_{orig}^{k}\right)}{\sum_{k=1}^{\ell} V\left(\boldsymbol{T}_{orig}^{k}\right)} \qquad (7)$$

Impact in the association measures.

In a bivariate analysis of data it's common to apply statistical tests for categorical variables to test the relationship between them. Cramer's V, measures the correlation degree of categorical variables in a contingent table. This measures depends on the Pearson Chi-Square, $\chi^2 = \sum_{\mathrm{i}=1}^{\ell} \sum_{\mathrm{j}=1}^{\sigma} \left(\mathrm{o}_{\mathrm{ij}} - \mathrm{e}_{\mathrm{ij}}\right)^2 / \mathrm{e}_{\mathrm{ij}}$ with $(\ell - 1)(\sigma - 1)$degrees of freedom, where o_{ij} is the observed value of the cell, *m* the number of lines of the table, *n* the number of columns and e_{ij} its expected value, $\mathrm{e}_{\mathrm{ij}} = \frac{\sum_{\mathrm{k}=1}^{\ell} \mathrm{x}_{\mathrm{kj}} \times \sum_{\mathrm{k}=1}^{\sigma} \mathrm{x}_{\mathrm{ik}}}{\sum_{\mathrm{i}=1}^{\mathrm{m}} \sum_{\mathrm{j}=1}^{\sigma} \mathrm{x}_{\mathrm{ij}}}$. Cramer's V will be given by CV= $\sqrt{\frac{\chi^2 / \sum_{\mathrm{i}=1}^{\ell} \sum_{\mathrm{j}=1}^{\sigma} \mathrm{x}_{\mathrm{ij}}}{\min((\ell-1),(\sigma-1))}}$.

Relative Cramer's V is used as a relative measure of comparison between two tables, the published and the original:

$$100 \times \frac{CV(\mathbf{T}_{div}) - CV(\mathbf{T}_{orig})}{CV(\mathbf{T}_{orig})} \tag{8}$$

Risk analysis

Risk is a measure that quantifies how secure is the information proposed to be disclosed. Typically the risk analysis is made by counting the number or proportion of modified cells, calculating the percentage of cell that can be recalculated, and computing the inverse of the variance of the error (Eq. 9).

$$100 \times 1/\mathrm{V}(\mathrm{T}_{\mathrm{div}} - \mathrm{T}_{\mathrm{orig}}) \tag{9}$$

3 Mixed Micro-Macro Approach Technique

In this section we present the M3A – Mixed Micro-Macro Approach – a post-tabular technique based on the fact that macrodata tables are originated by aggregating granular information, coming from microdata.

Unlike the traditional SDC techniques, the sensitive cells in M3A are considered as potential risk cells and not as risk cells. This particularity will allow that a value of cell that was identified as being a sensitive cell can be released without incurring in disclosure risk.

The main idea behind M3A is to use microdata information within each cell to protect all the respondents (and not the cell as a whole), through mathematical restrictions. M3A is based on four assumptions (scenarios) that ensure the utility of macrodata and the data confidentiality since in these scenarios it's considered that any estimation made by an intruder will be at a safe distance of d% from the original value. The mathematical restriction that define these four scenarios computes a safe interval to all the sensitive cells and after computing all the safe intervals, M3A choses a safe value for those cells by minimizing the information loss of each internal cell and, at the same time, minimizing the global information loss of the table.

The main differences between M3A and the other SDC techniques for tabular data are the following:

- M3A has the assumption that a sensitive cell (identified through a sensitive rule) may not be a risk cell but a potential risk cell, i.e., due to the mathematical restrictions applied to the sensitive cells, there are some cases where it is possible to verify that a sensitive cell may be released since its original value is contained in the safe interval computed by M3A.

- Besides of being a post-tabular technique, because it acts under the table structure, the safety of the released table is obtained through a microdata analysis, i.e., the computation of the safe value of sensitive cell is made through the mathematical restrictions that use all the individual contributions within each sensitive.

Thus, in M3A are introduced the concepts of safe and unsafe cells. When a cell is not a sensitive cell it will always be a safe cell. When a cell is a sensitive cell, under M3A, two things may happen: If the original value of the cell is contained in the safe interval computed by M3A, then, the cell is safe. Otherwise, the cell is unsafe.

The following four scenarios describe the main situations where an intruder may try to estimate the contribution of a respondent:

Scenario I. A cell with one respondent will always be unsafe.

Scenario II. A cell with two respondents will always be unsafe since each respondent can compute the contribution of the other one.

Scenario III. A cell with three or more respondents will be unsafe when:

a) The respondent with the highest contribution can obtain an estimate close to the real value of the second largest contribution, and vice versa.

b) The respondent with the second higher contribution, knowing his position, tries to estimate the contribution of the remaining respondents, assuming that the higher contribution is, at least, equals to its.

c) The remaining respondents, knowing their positions in the contribution to the cell, proceed to the calculation of the average value of the respondents with a higher contribution, via subtraction of their contribution to the disclosed value.

Scenario IV. An intruder knowing the number of respondents contained on a cell, may calculate an average value as being the contribution of each respondent. We should guarantee that those estimations are not near to a true contribution value.

In order to assure that the four scenarios are considered in the presentation of the safe value to disclose we have to analyze each sensitive cell to find the value to disclose, defined by $\widetilde{\mathrm{T}}$. As we will see, there are a set of possible values, which are described by the interval $[0, \widetilde{\mathrm{T}}_{\min}] \cup [\widetilde{\mathrm{T}}_{\max}, +\infty[$. It will be chosen the value for which difference to the original value is minimum.

Therefore, to preserve the confidentiality for each sensitive cell it is necessary to solve the equations (10) and (11) for the four scenarios.

$$\boldsymbol{\widetilde{T}_{min}}=\begin{cases}\text{scenario I: } \tilde{T}_{min} \leq (1-d\%) \times y_1 \\ \text{scenario II, IIIa: } \tilde{T}_{min} \leq (1-d\%) \times y_2 + y_1 \\ \text{scenario II, IIIa: } \tilde{T}_{min} \leq (1-d\%) \times y_1 + \ y_2 \\ \text{scenario IIIb: } \tilde{T}_{min} \leq (1-d\%) \times \sum_{t=3}^{m} y_t + 2y_2 \\ \text{scenario IIIc: } \frac{\tilde{T}_{min}}{\varphi-1} \leq (1-d\%) \times y_\delta + \frac{y_\varphi}{\varphi-1} \\ \text{scenario IV: } \tilde{T}_{min} \leq (1-d\%) \times y_i \times m \end{cases} \tag{10}$$

$$\boldsymbol{\widetilde{T}_{max}}=\begin{cases}\text{scenario I: } \tilde{T}_{max} \geq (1+d\%) \times y_1 \\ \text{scenario II, IIIa: } \tilde{T}_{max} \geq (1+d\%) \times y_2 + y_1 \\ \text{scenario II, IIIa: } \tilde{T}_{max} \geq (1+d\%) \times y_1 + \ y_2 \\ \text{scenario IIIb: } \tilde{T}_{max} \geq (1+d\%) \times \sum_{t=3}^{m} y_t + 2y_2 \\ \text{scenario IIIc: } \frac{\tilde{T}_{max}}{\varphi-1} \geq (1+d\%) \times y_\delta + \frac{y_\varphi}{\varphi-1} \\ \text{scenario IV: } \tilde{T}_{max} \geq (1+d\%) \times y_i \times m \end{cases} \tag{11}$$

Where $0<d<100$ is the safety distance to the real value, $\mathrm{y_i}$ is the contribution of respondents $i = 1, \dots, m$, presented by a decreasing order of contribution and φ is a parameter defined by the user which corresponds to the φ^{th} respondent. Assuming the (n,k) rule, the parameter φ should be, at least n +1, and this would be sufficient to ensure the safety of the respondents within the cell. The value proposed to disclose ($\widetilde{\mathrm{T}}$), is contained in the interval $[0, \widetilde{\mathrm{T}}_{\min}] \cup [\widetilde{\mathrm{T}}_{\max}, +\infty[$.

Since M3A algorithm computes a safety interval of the values to be disclosed and ensures a high similarity to the original table, the best value to be disclosed will be one of the two interval limits: $\widetilde{\mathrm{T}}_{\min}$ or $\widetilde{\mathrm{T}}_{\max}$ which is closer to the original cell value. To do that, M3A considers a predefined sequence of steps for the treatment of the cells: First, an initial table is built as a copy of the original but leaving sensitive cells in blank; second, are calculated the optimal values to disclose, according to scenarios II, III and IV for cells with more than one respondent. The computed values are then introduced in the table to be published and the marginal totals computed. Finally, M3A solves the scenario I to cells with only one respondent and choses the value to be published using the computed marginal totals. When a cell with one respondent is solved, the disclosed value is introduced in the table and the marginal totals updated.

The treatment of information in phases allows a more efficient computation of the table to be published. Note that, when the real value of the cell is contained in the center of the unsafe interval $]\widetilde{\mathrm{T}}_{\mathrm{min}}, \widetilde{\mathrm{T}}_{\mathrm{max}}[$, as it happens with cell with a unique respondent, both limits have the same proximity to the original value, so, in normal conditions, choosing the optimal value to disclose implies the creation of $(2!)^{w}$ possible tables, where w is the number of cells with only one respondent. Since M3A computes and updates the marginal totals, the choice between $\widetilde{\mathrm{T}}_{\mathrm{min}}$ and $\widetilde{\mathrm{T}}_{\mathrm{max}}$ is made by minimizing the loss of global information, i.e., by choosing the value that minimizes the global information loss for the corresponding row and column of the cell.

As mentioned earlier, the fact that M3A uses the microdata information within each sensitive cell enables the identification of sensitive cell that don't put confidentiality at risk. The following example illustrates the difference between a sensitive cell and a risk cell, and how M3A deals with such situation:

Consider a cell containing 7 respondents whose contributions are: 10000, 8000, 1600, 1500, 1100, 900, 800, being the original total equal to 23900. Assuming a (n,k) rule with n=2 and k=75, $\varphi = n + 1 = 3$ and a safety level of d=10%. According to the sensitive rule, this is a sensitive cell (since the two higher contributions represent 75,31% of the cell value). However, by solving the inequalities of M3A we verified that the safe interval is given by the set $[0,9000] \cup [23600, +\infty[$. Since the original value of the cell is contained in the safe interval, this cell is sensitive but doesn't represent any risk, therefore, the cell is safe and the original value may be published.

It is important to note that this methodology is innovative in the sense that all the contributors for a specific cell and their corresponding risks are analyzed to determine if the cell, as whole, is at risk.

4 RESULTS

To compare M3A with rounding and cell suppression techniques, we collected from SABI database (Sistema de Análise de Balanços Ibéricos) [20], data from 2009 for 245,595 Portuguese companies: company's geographical location (22 districts of Portugal), number of plants (10 classes of values) and turnover, in millions Euros.

To apply the techniques we use the following parameters: sensitive rule (n, k) with n =2 and k =75, d=10 and $\varphi = n + 1 = 3$. M3A was implemented in R Language [21]; For the rounding technique, MS Excel was used and for Cell Suppression techniques we took the R library sdcTable [22]. Results are presented in Table 1.

Table 1. Comparing M3A, Rounding and Cell Suppression techniques

Measure	M3A	HiTaS*	Hypercube*	Conv. Round.
Absolute loss per cell (using internal cells)	30,147,956 (-7.38%)	78,002,116 (-19.09%)	78,074,683 (-19.11%)	64,969,303 (-15.90%)
Absolute loss total (using grand cells)	22,511,042 (-5.51%)	-	-	700,183 (-0.17%)
Entropy (Eq. 6)	-3.85649	19.33886	19.37221	Inf.
Relative Variance (Eq.7)	13.19138	16.09627	16.09622	0.9361081
Relative Cramer's V (Eq. 8)	6.444834	41.63242	41,61949	41.63242
Number of sensitive cells	44	44	44	44
Risk (Eq. 9)	5.46E-11	5.33E-11	5.33E-11	0.00181
Number of modified internal cells / suppressed cells	44 (40.4%)	53 (48,6%)	54 (49,5%)	108 (99.1%)

* In the calculation of risk and loss information measures for cell suppression an average value to each suppressed cell was assumed which restored the additivity property for table columns.

Table 1 confirms that the data treatment at a microdata level produces better results in data quality. Metrics such as Entropy, Relative Variance and Relative Cramer's V are intended to be near zero when there is higher similarity to the original table. Regarding the loss information by cell and by total, we obtained better results with M3A comparatively with the other methodologies.

A Mann-Whitney Wilcoxon test, used to compare the distribution of each row and column between the original table and the disclosed, proved that the table computed by M3A maintains the distribution of the original table.

Through the Risk-Utility map we compared the trade-off between risk and Absolute Loss per Cell, Entropy and Relative Variance and defined, as example, a thresh- old of 0.1%. It is visible that M3A technique has good results and a better trade-off.

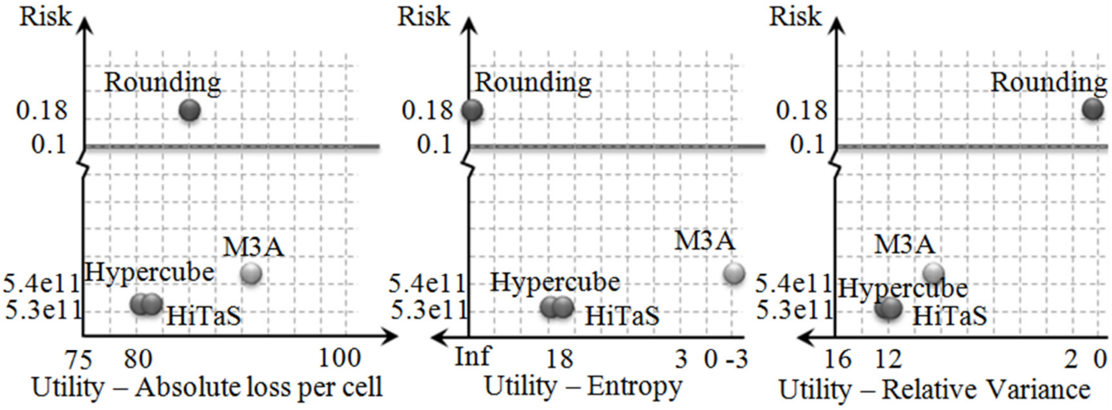

Fig. 1. Risk-Utility maps. For the Absolute Loss per Cell we used the percentage of information that was not modified. In the remaining graphs, the direction of x-axis goes from right to left, in order to maintain the original structure of the R-U map: tables positioned at the right of the map have high quality.

Besides the quantitative analysis we highlight the advantages (A) and disadvantages (D) of each technique by subject area.

Table 2. Advantages and Disadvantages of M3A, Rounding and Cell Suppression techniques.

Subject	M3A	Cell Suppression	Rounding
Non sensitive cells	**(A)** Remain unchanged	**(D)** Some are suppressed	**(D)** May be changed
Sensitive cells	**(A)** All analyzed. Some may not represent risk, and remain unchanged	**(D)** All suppressed	**(D)** All rounded
Disclosed totals	**(D)** Are not the originals, but are consistent with the released table	**(A)** Are the originals	**(A/D)** May be the originals if it is used the controlled rounding
Additivity property	**(A)** The table maintains this property	**(D)** Not applicable since information is suppressed	**(A/D)** This property is not assured
Deeper Statistical analysis	**(A)** It's possible to do since it maintains the characteristics of the original data	**(D)** It is not possible to do, since a lot of information is suppressed	**(D)** It is not possible to do, since some relevant detail is lost

As it can be seen, M3A is a method which aims to produce safe tables that allow users to obtain results very similar to reality when applying data analysis techniques to the published table. This is proved not only by the risk and utility measures but also by the Mann-Whitney Wilcoxon test whose p-values results indicate that the original data distribution is preserved. Although the users do not have access to the original marginal totals, this table provides a good statistical analysis and perception of reality that users cannot have through cell suppression, since this technique suppresses many cells, or through rounding, since those techniques lose relevant detail in the rounding process.

5 CONCLUSIONS

Mathematical restrictions considered in M3A technique are used to analyze all the microdata set, i.e., all the contributions contained in each sensitive cell of the table and propose safe values to disclose. Despite a microdata approach is implicit here, the treatment of the table is made in an aggregative perspective (cell by cell) and that is why this technique must be used only for tabular data.

By making a clear reading of the microdata, this technique achieved better results when compared with other SDC techniques, especially when compared to rounding and cell suppression. We concluded that, M3A:

- Obtains better results in terms of data quality, ensuring, at the same time, low levels of risk.
- The results provided by the measures absolute loss per cell and total loss, entropy, Relative Cramer's V, Relative Variance and number of modified internal cells proves that the information loss is significantly inferior when compared to the other SDC macrodata techniques. The Mann-Whitney Wilcoxon test also proved that the original distribution of the data is preserved.
- M3A doesn't remove data. Therefore, it is a good alternative to cell suppression.
- M3A allows gains in comparison to other techniques, and it is possible to find cases where sensitive cells are not risk cells.
- It is a technique with low complexity, it's easily understood and has an easy implementation and processing.

M3A is a new technique characterized by a different approach to SDC for tabular data which has potential to be improved. As current limitations we should enumerate the following: it does not preserve the original marginal totals and it's not applicable to tables with more than two dimensions, hierarchical table and tables with negative values. Therefore, in a future work we propose the inclusion of restrictions to ensure that the totals reported in the table are equal to the original, and the expansion of this method to tables with more dimensions or hierarchical tables and tables with negative values.

6 References

1. European Commission: European Statistics Code of Practice, for the National and Community Statistical Authorities. In: General and regional statistics, Methodologies and working papers. (2011)
2. Cox, L. H., Orelien, J.G, Shah B.V.: A Method for Preserving Statistical Distributions Subject to Controlled Tabular Adjustment. In Lecture Notes in Computer Science, J. Do-

mingo-Ferrer and L. Franconi (eds). Privacy in Statistical Databases- Vol. 4302, pp. 1-11 (2006)

3. Oganian, A., Domingo-Ferrer, J.: A posteriori disclosure risk measure for tabular data based on conditional entropy. In: Statistics and Operations Research Transactions, vol. 27, N.2, pp. 175-190 (2003)
4. Hundepool, A.: The ARGUS-Software. In: in Monographs of Official Statistics. Work session on Statistical Data Confidentiality, Eurostat – Office for Official Publications on the European Communities, Luxemburg, vol. 3, pp. 347-363 (2003)
5. Willenborg, L., de Wall, T.: Statistical Disclosure Control in practice. In: Lecture notes in Statistics, Vol. 111. Springer, Heidelberg (1996)
6. Hundepool. A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R, Naylor, J., Nordholt, E.S., Seri, G., Wolf, P.P.: Handbook on Statistical Disclosure Control. ESSNet – hand- book SDC, v1.2 (2010)
7. Domingo-Ferrer, J., Torra, V.: A Critique of the Sensitivity Rules Usually Employed for Statistical Table Protection. In: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no 5, pp. 545-556 (2002)
8. de Wall, T.: Processing of Erroneous and Unsafe Data. Phd thesis, Erasmus Research Institute of Management Doctoral Programme, Senaatszaal, Woudestein: ISBN-5892-045-3 (2003)
9. Salazar, J. J.: A New Approach to Round Tabular Data. In: Lecture Notes in Computer Science, Domingo-Ferrer, J., Franconi, L. (eds), Privacy in Statistical Databases, Vol. 4302, pp.25-34 (2006)
10. Cox, L. H., George, J. A.: Controlled Rounding for Tables with Subtotals. In: Annals of Operations Research, vol. 20, pp.141-157 (1989)
11. Doerr, B., Friedich, T., Klein, C., Osbild, R.: Unbiased Matrix Rounding. In: Lecture Notes in Computer Science, Arge.L, Freivalds, R. (eds.), Scandinavian Workshop on Algorithm Theory, vol. 4059, pp. 102-112 (2006)
12. Salazar, J.J.: Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. In: Mathematical Programming, K.M. Anstreicher and D. Ralph (eds). Mathematics and Statistics, Vol. 5, no 2-3, pp. 583-603 (2006)

13. Salazar, J. J.: Extending Cell Suppression to Protect Tabular Data Against Several Attackers. In. Lecture Notes in Computer Science, Domingo-Ferrer, J. (eds), Inference Control in Statistical Databases, vol. 2316, pp. 34-58 (2002)
14. Daalmans, J., de Waal, T.: A General Formulation of the Secondary Cell Suppression Problem. Discussion paper (10009), The Hague: Statistics Netherlands (2010).
15. Fischetti, M., Salazar, J. J.: Solving the Cell Suppression Problem on Tabular Data with Linear Constraints. In: Management Science, vol. 47, N.º 7, pp. 1008-1027 (2001)
16. Dobra, A., Fienberg, S.E.: Bounds for Cell Entries in Contingency Tables Induced by Fixed Marginal Total with Application to Disclosure Limitation. In: Statistical Journal of the United Nations ECE, vol. 18, pp. 363-371 (2001)
17. Cox, L. H., Zayatz, L. V.: An Agenda for Research in Statistical Disclosure Limitation. In: Journal of Official Statistics, vol. 11, Nº 2, pp.205-220 (1995)
18. Giessing, S., Repsilber, D.: Tools and Strategies to Protect Multiple Tables with the GHQUAR Cell Suppression Engine. In: Lecture Notes in Computer Science, Domingo- Ferrer, J. (eds), Inference Control in Statistical Databases, vol. 2316, pp. 181-192 (2002)
19. Gomatan, S., Karr,A.:Distortion Measures for Categorical Data Swapping. In: National Institute of Statistical Science, Technical Report, No 131 (2003)
20. Bureau van Dijk: Bureau van Dijk Electronic Publishing, Company information in an instant, Nortel Net-works, United Kingdom, Westacott Way (2003)
21. R Development Core Team: R: A Language and Environment for Statistical Computing. In: R Foundation for Statistical Computing, Vienna, Austria, Version 2.13.1, <http://www.R- project.org> (2011)
22. Meindl, B.: sdcTable: Statistical Disclosure Control for Tabular Data. R package, version 0.6.4, <http://CRAN.R-project.org/package=sdcTable> (2011)
23. Giessing, S., Post-tabular Stochastic Noise to Protect Skewed Business Data, Joint UNECE/Eurostat work session on statistical data confidentiality (Tarragona, Spain, 26-28 October 2011), United Nations Economic Commission for Europe (Unece) (2011).

Editorial Board (wps@fep.up.pt)
Download available at: `http://wps.fep.up.pt/wplist.php`
also in `http://ideas.repec.org/PaperSeries.html`

www.fep.up.pt