

**WP. 11**  
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE  
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION  
STATISTICAL OFFICE OF THE  
EUROPEAN COMMUNITIES (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**  
(Bilbao, Spain, 2-4 December 2009)

**Topic (ii): Synthetic and hybrid data**

## **MICROAGGREGATION-BASED NUMERICAL HYBRID DATA**

**Invited Paper**

Prepared by Josep Domingo-Ferrer, Universitat Rovira i Virgili, Catalonia, Spain

# Microaggregation-Based Numerical Hybrid Data

Josep Domingo-Ferrer

Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics,  
UNESCO Chair in Data Privacy, Av. Països Catalans 26, E-43007 Tarragona,  
Catalonia.  
(josep.domingo@urv.cat)

**Abstract.** In statistical disclosure control of microdata, the usual approach is to either mask (*i.e.* perturb) the original data or generate synthetic (*i.e.* simulated) data preserving some pre-selected statistics of the original data. Masked data may approximately preserve a broad range of distributional characteristics, although very few of them (if any) are exactly preserved; on the other hand, synthetic data exactly preserve the pre-selected statistics and may seem less disclosive than masked data, but they do not preserve at all any statistics other than those pre-selected. Hybrid data mixing the original data and synthetic data have been proposed in the literature to combine the strengths of masked and synthetic data. We show how to easily obtain numerical hybrid data exactly preserving means and covariances by combining microaggregation with the IPSO synthetic generator. Furthermore, our method approximately preserves other statistics as well as some subdomain analyses. For those reasons and also due to its very simple parameterization, the new method is competitive versus both the literature on hybrid data and plain multivariate microaggregation.

## 1 Introduction

Given an original microdata set  $\mathbf{V}$ , the goal of statistical disclosure control (SDC) of microdata is to release a protected microdata set  $\mathbf{V}'$  in such a way that:

1. Disclosure risk (*i.e.* the risk that a user or an intruder can use  $\mathbf{V}'$  to determine confidential attribute values for a specific individual among those in  $\mathbf{V}$ ) is low.
2. User analyses (regressions, means, etc.) on  $\mathbf{V}'$  and on  $\mathbf{V}$  yield the same or at least similar results.

Microdata protection methods (see [5, 12] for further details) can generate the protected microdata set  $\mathbf{V}'$

- either by *masking original data*, *i.e.* generating a modified version  $\mathbf{V}'$  of the original microdata set  $\mathbf{V}$  using some kind of perturbation, sampling or granularity reduction;
- or by *generating synthetic data*  $\mathbf{V}'$  that preserve some statistical properties of the original data  $\mathbf{V}$ .

Masked data may approximately preserve a broad range of distributional characteristics, although very few of them (if any) are exactly preserved. So they seem a good option for the case in which *both* circumstances below concur:

- The data protector has no precise idea about what types of analyses will be carried out by the users of  $\mathbf{V}'$  (this is the most usual situation);
- The users can tolerate some accuracy loss in the results of their analyses on  $\mathbf{V}'$ , with respect to the results they would obtain on the original data set  $\mathbf{V}$ . This assumption is realistic if the alternative to getting protected data  $\mathbf{V}'$  is for users to get no data at all (no research possible) or be forced to declare their exact planned computations to the data protector for the latter to run them on the original data  $\mathbf{V}$  (cumbersome research).

On the other hand, the strong points of synthetic data are that they exactly preserve the pre-selected statistics and may seem less disclosive than masked data, because published records are simulated and not derived from modification of any particular original record (even if overfitted synthetic data might lead to disclosure [13]). On the negative side, the utility of synthetic data is zero for those statistics not pre-selected for preservation by the data protector.

## 1.1 Our contribution

Hybrid data mixing the original data and synthetic data have been proposed in the literature [3, 10], to combine the strengths of masked and synthetic data and neutralize their pitfalls. We show how to easily obtain numerical hybrid data by combining microaggregation [7, 9] with the IPSO synthetic data generator [2]. The resulting numerical hybrid data preserve means and covariances of original data as achieved by the Muralidhar-Sarathy procedure ([10], named for short MS in the sequel), but with a simpler and more intuitive parameterization. Furthermore, unlike MS, the new proposal approximately preserves subdomain analyses (*i.e.* analyses restricted to subsets of the data) and it outperforms plain multivariate microaggregation regarding both information loss and disclosure risk.

Section 2 gives background on IPSO and microaggregation. In Section 3, a scheme for generating microaggregation-based numerical hybrid data is described, and it is proven that the resulting hybrid data preserve the means and the covariances of the original data. An assessment of usability and performance is given in Section 4. Section 5 is a conclusion.

## 2 Background: the IPSO generator and microaggregation

### 2.1 The IPSO synthetic generator

A procedure called Information Preserving Statistical Obfuscation (IPSO) is proposed in [2] to synthesize numerical data sets. The basic form of IPSO will be called

here IPSO-A. Informally, suppose that we have a data set with  $n$  records and numerical attributes; attributes can be split in two sets  $X$  and  $Y$ , where the former are the confidential outcome attributes and the latter are non-confidential attributes (*e.g.* quasi-identifiers). Then  $X$  are taken as dependent and  $Y$  as independent attributes.

In the above setting, conditional on the specific quasi-identifier attributes  $y_i$ , the confidential attributes  $X_i$  are assumed to follow a multivariate normal distribution with covariance matrix  $\Sigma = \{\sigma_{jk}\}$  and a mean vector  $y_i B$ , where  $B$  is the matrix of regression coefficients.

Let  $\hat{B}$  and  $\hat{\Sigma}$  be the maximum likelihood estimates of  $B$  and  $\Sigma$  derived from the complete data set  $(y, x)$ . IPSO outputs a data matrix  $x'$  such that, when a multivariate multiple regression model is fitted to  $(y, x')$ , both statistics  $\hat{B}$  and  $\hat{\Sigma}$ , sufficient for the multivariate normal case, are preserved. Thus, synthetic data output by IPSO preserve the means and covariances of the original data.

We briefly describe the operation of IPSO for the univariate case, with a single confidential attribute  $X$  and a single non-confidential attribute  $Y$  (see [2] for details on the multivariate case). In the first step, a linear regression model is constructed to predict the values of  $X$  using  $Y$ . From the model, the intercept  $\hat{\beta}_0$  and the slope  $\hat{\beta}_1$  are estimated, and the predicted values of  $x_i$  are computed as

$$\hat{x}_i = \hat{\beta}_0 + \hat{\beta}_1 \times y_i$$

Next, a vector of noise terms  $A$  of size  $n$  is generated from a standard normal distribution. Attribute  $A$  is then regressed on  $Y$  and  $X$  and the residuals from this regression are computed. Let  $B$  represent the residuals.  $B$  has mean 0 and is orthogonal to both  $Y$  and  $X$ . Let  $\sigma_{BB}^2$  be the variance of  $B$ . Then  $B$  is transformed to a new attribute  $C$  as follows:

$$c_i = \frac{b_i}{\sigma_{BB}} \sigma_{ee}, \quad i = 1, 2, \dots, n$$

where  $\sigma_{ee}^2 = \sigma_{XX}^2 - \frac{(\sigma_{XY})^2}{\sigma_{YY}^2}$ , and  $\sigma_{XX}^2$ ,  $\sigma_{YY}^2$ , and  $\sigma_{XY}$  are the variance of  $X$ ,  $Y$  and the covariance between  $X$  and  $Y$ , respectively. Finally, the synthetic values are generated as

$$x'_i = \hat{x}_i + c_i, \quad i = 1, 2, \dots, n$$

The resulting synthetic confidential attribute  $X'$  has *exactly* the same mean and variance as  $X$ .

## 2.2 Microaggregation

Microaggregation is a family of perturbative SDC methods originally defined for continuous data [4, 7] and extended for categorical data in [9]. Whatever the data type, microaggregation can be operationally defined in terms of the following two steps:

**Partition:** The set of original records is partitioned into several clusters in such a way that records in the same cluster are *similar* to each other and so that the number of records in each cluster is at least  $k$ .

**Aggregation:** An aggregation operator (for example, the mean for continuous data or the median for categorical data) is computed for each cluster and is used to replace the original records. In other words, each record in a cluster is replaced by the cluster’s prototype.

In the remainder of this paper, we will be interested *only in the partition step* of microaggregation. We recall the MDAV algorithm [9] for the partition step in multivariate microaggregation of numerical data.

**Algorithm 1 (MDAV( $R$ : data set,  $k$ : integer))**

1. *While*  $|R| \geq 3k$  *do*
  - (a) *Compute the average (mean) record*  $\tilde{x}$  *of all records in*  $R$ . *The average record is computed attribute-wise.*
  - (b) *Consider the most distant record*  $x_r$  *to the average record*  $\tilde{x}$  *using the Euclidean distance.*
  - (c) *Find the most distant record*  $x_s$  *from the record*  $x_r$  *considered in the previous step.*
  - (d) *Form two clusters around*  $x_r$  *and*  $x_s$ , *respectively. One cluster contains*  $x_r$  *and the*  $k - 1$  *records closest to*  $x_r$ . *The other cluster contains*  $x_s$  *and the*  $k - 1$  *records closest to*  $x_s$ .
  - (e) *Take as a new data set*  $R$  *the previous data set*  $R$  *minus the clusters formed around*  $x_r$  *and*  $x_s$  *in the last instance of Step 1d.*

*end while*
2. *If there are between*  $3k - 1$  *and*  $2k$  *records in*  $R$ :
  - (a) *compute the average record*  $\tilde{x}$  *of the remaining records in*  $R$
  - (b) *find the most distant record*  $x_r$  *from*  $\tilde{x}$
  - (c) *form a cluster containing*  $x_r$  *and the*  $k - 1$  *records closest to*  $x_r$
  - (d) *form another cluster containing the rest of records;*

*else (less than*  $2k$  *records in*  $R$ ) *form a new cluster with the remaining records.*

For numerical attributes, variable-size microaggregation is an alternative to MDAV. In this type of microaggregation, the partition step yields clusters of size varying between  $k$  and  $2k - 1$  records depending on the distribution of the data. Variable-size heuristics, like  $\mu$ -Approx [8], usually yield higher intra-cluster similarity (more homogeneous records within each cluster) than fixed-size heuristics such as MDAV, especially if the original data form natural clusters.

### 3 Microaggregation-based numerical hybrid data

Let  $\mathbf{V}$  be an original data set consisting of  $n$  records. On input an integer parameter  $k \in \{1, \dots, n\}$ , the procedure described in this section generates a hybrid data set  $\mathbf{V}'$ . The greater  $k$ , the more synthetic is  $\mathbf{V}'$ . Extreme cases are: i)  $k = 1$ , which would yield  $\mathbf{V}' = \mathbf{V}$  (the output data are exactly the original input data); and ii)  $k = n$ , which yields a completely synthetic output data set  $\mathbf{V}'$ .

The procedure calls two algorithms:

- The IPSO synthetic data generator  $IPSO(\mathbf{C}, \mathbf{C}')$ , which, given an original data (sub)set  $\mathbf{C}$ , generates a synthetic data (sub)set  $\mathbf{C}'$  preserving means and covariances of  $\mathbf{C}$ .
- A microaggregation heuristic, *e.g.* MDAV or  $\mu$ -Approx, which, on input a set of  $n$  records and parameter  $k$ , partitions the set of records into clusters containing between  $k$  and  $2k - 1$  records. Cluster creation attempts to maximize intra-cluster homogeneity.

**Procedure 1** ( $\mathbb{R}$ -*microhybrid*( $\mathbf{V}, \mathbf{V}', \text{parms}, k$ ))

1. Call *microaggregation*( $\mathbf{V}, k$ ). Let  $C_1, \dots, C_\kappa$  for some  $\kappa$  be the resulting clusters of records.
2. For  $i = 1, \dots, \kappa$  call  $IPSO(C_i, C'_i)$ .
3. Output a hybrid data set  $\mathbf{V}'$  whose records are those in the clusters  $C'_1, \dots, C'_\kappa$ .

At Step 1 of procedure  $\mathbb{R}$ -*microhybrid* above, clusters containing between  $k$  and  $2k - 1$  records are created. Then at Step 2, a synthetic version of each cluster is generated. At Step 3, the original records in each cluster are replaced by the records in the corresponding synthetic cluster (instead of replacing them with the average record of the cluster, as done in conventional microaggregation). The  $\mathbb{R}$ -*microhybrid* procedure bears some resemblance to the condensation approach proposed in [1]; however,  $\mathbb{R}$ -*microhybrid* is more general because:

- i) Clusters do not need to be all of size  $k$  (their sizes can vary between  $k$  and  $2k - 1$ );
- ii) Instead of using an *ad hoc* clustering heuristic like condensation,  $\mathbb{R}$ -*microhybrid* can use any of the best microaggregation heuristics cited above, which should yield higher within-cluster homogeneity and thus less information loss.
- iii)  $\mathbb{R}$ -*microhybrid* can be generalized to non-numerical data by using non-numerical microaggregation and synthetic data generators other than IPSO, as proposed in the journal version of this paper [6].

### 3.1 On the role of parameter $k$

We justify here the role of parameter  $k$  in *microhybrid*:

- If  $k = 1$ , the output is the same original data set, because the procedure creates  $n$  clusters (as many as the number of original records). With  $k = 1$ , even variable-size heuristics will yield all clusters of size 1, because the maximum intra-cluster similarity is obtained when clusters consist all of a single record.
- If  $k = n$ , the output is a single synthetic cluster: the procedure is equivalent to calling IPSO once for the entire data set.
- For intermediate values of  $k$ , several clusters are obtained at Step 1, and means and covariances are preserved within each synthetic cluster generated at Step 2. As  $k$  decreases, the number of clusters (whose means and covariances are preserved in the data output at Step 3) increases, which causes the output data to look more and more like the original data. Each cluster can be regarded as a constraint on the synthetic data generation: the more constraints, the less freedom there is for generating synthetic data, and the output resembles more the original data. This is why the output data can be called hybrid.

It must be noted here that, when using IPSO with  $|X|$  confidential attributes and  $|Y|$  non-confidential attributes, it turns out that  $k$  must be at least  $|X| + 2|Y| + 1$ ; otherwise there are not enough degrees of freedom for the generator to work.

### 3.2 Mean vector and covariance matrix exact preservation

We show here that  $\mathbb{R}$ -*microhybrid* exactly preserves the mean vector and the covariance matrix of the original data set, which are sufficient statistics when the underlying distribution of data is multivariate normal.

**Lemma 1 (Preservation of means and covariances)** *Let  $\mathbf{V}$  be an original data set whose attributes are numerical and fall into confidential attributes  $\mathbf{X}(= X_1, \dots, X_L)$  and non-confidential attributes  $\mathbf{Y}(= Y_1, \dots, Y_M)$ . Let  $\mathbf{V}'$  be a hybrid data set obtained from  $\mathbf{V}$  using  $\mathbb{R}$ -microhybrid, whose attributes are  $\mathbf{X}'(= X'_1, \dots, X'_L)$  (hybrid versions of  $\mathbf{X}$ ) and  $\mathbf{Y}$ . Then the means and covariances of the confidential attributes in  $\mathbf{V}$  and  $\mathbf{V}'$  are exactly the same, that is, it holds that*

$$\bar{\mathbf{X}} = \bar{\mathbf{X}'}, \quad \Sigma_{\mathbf{X}'\mathbf{X}'} = \Sigma_{\mathbf{X}\mathbf{X}}, \quad \Sigma_{\mathbf{X}'\mathbf{Y}} = \Sigma_{\mathbf{X}\mathbf{Y}} \quad (1)$$

*Proof:* Let  $n$  be the number of records in  $\mathbf{V}$  and  $\mathbf{V}'$ . Let  $k$  be the parameter used to call  $\mathbb{R}$ -*microhybrid*, and let  $C_1, \dots, C_\kappa$  be the clusters obtained in Step 1. Let  $\mathbf{X}_i, \mathbf{Y}_i$  be the confidential and the non-confidential attributes restricted to cluster  $C_i$  of the original data set  $\mathbf{V}$ . Let  $\mathbf{X}_i(= X(i)_1, \dots, X(i)_L)$  be the confidential attributes restricted to cluster  $C_i$  of the original data set. Let  $\mathbf{X}'_i(= X'(i)_1, \dots, X'(i)_L)$  be the confidential attributes restricted to cluster  $C_i$  of the output data set. Since  $\mathbf{X}'_i$  is the output obtained by applying IPSO to  $\mathbf{X}_i$ , it holds that

$$\bar{\mathbf{X}}'_i = \bar{\mathbf{X}}_i, \quad i = 1, \dots, \kappa \quad (2)$$

$$\Sigma_{\mathbf{X}'_i \mathbf{X}'_i} = \Sigma_{\mathbf{X}_i \mathbf{X}_i}, \quad i = 1, \dots, \kappa \quad (3)$$

$$\Sigma_{\mathbf{X}'_i \mathbf{Y}_i} = \Sigma_{\mathbf{X}_i \mathbf{Y}_i}, \quad i = 1, \dots, \kappa \quad (4)$$

From Equations (2), it follows that  $\bar{\mathbf{X}} = \bar{\mathbf{X}}'$ .

We now pick any two indices  $l, m \in \{1, \dots, L\}$  and consider the components at row  $l$  and column  $m$  of  $\Sigma_{\mathbf{X}'\mathbf{X}'}$  and  $\Sigma_{\mathbf{X}\mathbf{X}}$ :

$$s'_{lm} = \frac{\sum_{j=1}^n x'_{l,j} x'_{m,j}}{n} - \bar{X}'_l \bar{X}'_m$$

$$s_{lm} = \frac{\sum_{j=1}^n x_{l,j} x_{m,j}}{n} - \bar{X}_l \bar{X}_m$$

Since  $\bar{\mathbf{X}} = \bar{\mathbf{X}}'$ , proving that  $s'_{lm} = s_{lm}$  amounts to checking whether

$$\sum_{j=1}^n x'_{l,j} x'_{m,j} \stackrel{?}{=} \sum_{j=1}^n x_{l,j} x_{m,j} \quad (5)$$

The check in Equation (5) can be rewritten by taking clusters into account as

$$\sum_{i=1}^{\kappa} \sum_{j=1}^{k_i} x'(i)_{l,j} x'(i)_{m,j} \stackrel{?}{=} \sum_{i=1}^{\kappa} \sum_{j=1}^{k_i} x(i)_{l,j} x(i)_{m,j} \quad (6)$$

where  $k_i$  is the actual size of cluster  $C_i$ , with  $k \leq k_i \leq 2k - 1$ . By Equations (3), we have that  $s'(i)_{lm} = s(i)_{lm}$  for all  $i, l, m$ . Using Equations (2) this implies for all clusters  $C_i$

$$\sum_{j=1}^{k_i} x'(i)_{l,j} x'(i)_{m,j} = \sum_{j=1}^{k_i} x(i)_{l,j} x(i)_{m,j}$$

Therefore the check in Equation (6) holds with equality and we have  $\Sigma_{\mathbf{X}'\mathbf{X}'} = \Sigma_{\mathbf{X}\mathbf{X}}$ . A similar argument based on Equations (4) and (2) shows that  $\Sigma_{\mathbf{X}'\mathbf{Y}} = \Sigma_{\mathbf{X}\mathbf{Y}}$ .  $\square$

### 3.3 Microaggregation and approximate preservation of other analyses

It can be seen from its proof that Lemma 1 holds (that is, means and covariances are exactly preserved) no matter how the clusters  $C_1, \dots, C_\kappa$  are formed. In other words, a very bad microaggregation procedure could be used. The added value of using good microaggregation algorithms like MDAV or  $\mu$ -Approx, which try to maximize intra-cluster homogeneity, is that small intra-cluster variances are obtained. In this way, at least for small  $k$ , the hybrid records are generated in a very constrained way: indeed, for small  $k$ , there are a lot of clusters for which the mean and a small intra-cluster variance should be exactly preserved by the hybrid data. This gives reasonable confidence that additional statistics or subdomain analyses may be approximately preserved.

This confidence is substantiated as follows. In this section, we show that the lower the intra-cluster variance, the more approximately are third-order central moments preserved. In [6], we explore how well means and covariances are preserved for random subsets of records and several choices of  $k$ ; we also examine the influence of  $k$  on the approximate preservation of higher-order central moments.

**Lemma 2 (Approximate preservation of third-order central moments)** *Let  $X$  be an attribute in the original data set  $\mathbf{V}$  consisting of  $n$  records and  $X'$  be the corresponding attribute in the hybrid data set  $\mathbf{V}'$  obtained with  $\mathbb{R}$ -microhybrid using parameter  $k$  and a clustering  $\{C_i : i = 1, \dots, \kappa\}$ . Let  $\bar{x}$ ,  $\bar{x}_i$  be the averages of  $X$  over  $\mathbf{V}$  and  $C_i$ , respectively. Let  $x_{ij}$ , resp.  $x'_{ij}$ , for  $j = 1, \dots, k_i$ , with  $k \leq k_i \leq 2k - 1$ , denote the values of  $X$ , resp.  $X'$ , in  $C_i$ . If  $B$  is such that  $(x_{ij} - \bar{x}_i)^2 \leq B$  and  $(x'_{ij} - \bar{x}_i)^2 \leq B$  for all  $i, j$  then*

$$\left| \sum_{i,j} (x'_{ij} - \bar{x})^3 - \sum_{i,j} (x_{ij} - \bar{x})^3 \right| \leq 2n \max(1, B^{3/2})$$

*Proof:* After some algebraic manipulation we can write

$$\sum_{i,j} (x_{ij} - \bar{x})^3 = \sum_{i,j} (x_{ij} - \bar{x}_i)^3 + 3 \sum_i [(\bar{x}_i - \bar{x}) \sum_j (x_{ij} - \bar{x}_i)^2] + \sum_i k_i (\bar{x}_i - \bar{x})^3 \quad (7)$$

Also, we can use that  $\mathbb{R}$ -microhybrid yields an  $X'$  with the same intra-cluster means  $\bar{x}_i$  as  $X$  to write

$$\sum_{i,j} (x'_{ij} - \bar{x})^3 = \sum_{i,j} (x'_{ij} - \bar{x}_i)^3 + 3 \sum_i [(\bar{x}_i - \bar{x}) \sum_j (x_{ij} - \bar{x}_i)^2] + \sum_i k_i (\bar{x}_i - \bar{x})^3 \quad (8)$$

where, in the last equality, we have used that the intra-cluster variances of  $X'$  are also the same as those of  $X$ . Now if we subtract Equation (7) from Equation (8) we obtain:

$$\begin{aligned} \left| \sum_{i,j} (x'_{ij} - \bar{x})^3 - \sum_{i,j} (x_{ij} - \bar{x})^3 \right| &= \left| \sum_{i,j} (x'_{ij} - \bar{x}_i)^3 - \sum_{i,j} (x_{ij} - \bar{x}_i)^3 \right| \\ &\leq \sum_{i,j} |x'_{ij} - \bar{x}_i|^3 + \sum_{i,j} |x_{ij} - \bar{x}_i|^3 \leq 2n \max(1, B^{3/2}) \end{aligned}$$

where, in the last inequality, we have used that if  $z^2 \leq B$  then either  $z^2 \geq 1$ , which implies  $|z|^3 \leq (\sqrt{B})^3$ , or  $z^2 < 1$ , which implies  $|z|^3 < 1$ ; hence,  $|z|^3 \leq \max(1, B^{3/2})$ .  $\square$

## 4 Usability and performance assessment

We compare the  $\mathbb{R}$ -microhybrid procedure with the alternative MS, which also preserves means and covariances, and with plain multivariate microaggregation [7, 9]. Previous proposals for hybrid data generation not guaranteeing exact preservation of means and covariances (like [3]) are not considered.

### 4.1 Comparison with the Muralidhar-Sarathy hybrid generator

We recall the procedure MS using the notation of Lemma 1. In that procedure, the hybrid values are generated as

$$\mathbf{x}'_j = \gamma + \mathbf{x}_j \alpha^T + \mathbf{y}_j \beta^T + \mathbf{e}_i, \quad j = 1, \dots, n$$

In order to enforce the preservation of means and covariances specified by Equations (1), the following equalities are necessary:

$$\beta^T = \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}\mathbf{X}} (\mathbf{I} - \alpha^T)$$

$$\gamma = (\mathbf{I} - \alpha) \bar{\mathbf{X}} - \beta \bar{\mathbf{Y}}$$

$$\Sigma_{\mathbf{ee}} = (\Sigma_{\mathbf{X}\mathbf{X}} - \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}\mathbf{X}}) - \alpha (\Sigma_{\mathbf{X}\mathbf{X}} - \Sigma_{\mathbf{X}\mathbf{Y}} \Sigma_{\mathbf{Y}\mathbf{Y}}^{-1} \Sigma_{\mathbf{Y}\mathbf{X}}) \alpha^T$$

where  $\mathbf{I}$  is the identity matrix and  $\Sigma_{\mathbf{ee}}$  is the covariance matrix of the noise terms  $\mathbf{e}$ .

Thus,  $\alpha$  completely specifies the procedure, similarly to  $k$  in our  $\mathbb{R}$ -*microhybrid* procedure. However, there are differences:

- While  $k$  is an integer between 1 and  $n$ ,  $\alpha$  is a matrix with real-valued components.
- While the choice of the value of  $k$  is very intuitive (see Section 3.1 above), the authors of MS admit that  $\alpha$  must be selected carefully to ensure that  $\Sigma_{\mathbf{ee}}$  is positive semidefinite. They consider three options for specifying the  $\alpha$  matrix:
  1. Take  $\alpha$  as a diagonal matrix with all values in the diagonal being equal. In this case,  $\Sigma_{\mathbf{ee}}$  is positive semidefinite and the value of the hybrid attribute  $X'_i$  depends only on  $X_i$ , but not on  $X_j$  for  $j \neq i$ . All confidential attributes  $X_i$  are perturbed at the same level.
  2. Take  $\alpha$  as a diagonal matrix, with values in the diagonal being not all equal. In this case,  $X'_i$  still depends only on  $X_i$ , but not on  $X_j$  for  $j \neq i$ . The differences are that the confidential attributes are perturbed at different levels and there is no guarantee that  $\Sigma_{\mathbf{ee}}$  is positive semidefinite, so it may be necessary to try several values of  $\alpha$  until positive semidefiniteness is achieved.
  3. Taking  $\alpha$  as a non-diagonal matrix does not guarantee positive semidefiniteness either and the authors of MS do not see any advantage in it, although it would be the only way to have  $X'_i$  depend on several attributes among  $(X_1, \dots, X_L)$ . With  $\mathbb{R}$ -*microhybrid*, the dependence of  $X'_i$  on the original confidential attributes is the one provided by the underlying IPSO method.

Beyond preserving means and covariances like MS,  $\mathbb{R}$ -*microhybrid* goes a step further by offering the following properties when a good microaggregation heuristic yielding a small intra-cluster variance is used:

- There is approximate preservation of third-order central moments (see Lemma 2 above) and fourth-order central moments (see [6]);
- There is also approximate preservation *over subdomains* (data subsets) of means, variances, covariances, and third and fourth-order central moments (see [6]).

## 4.2 Comparison with plain multivariate microaggregation

As recalled in Section 2.2 above, plain multivariate microaggregation consists of a partition step and an aggregation step. The latter step (not used in  $\mathbb{R}$ -*microhybrid*) consists of replacing the records within each cluster with the average record: for each attribute, the value in the average record is the average of the attribute values in the cluster; in the case of numerical data, the average is the arithmetic mean.

If applied to quasi-identifiers as proposed in [9], microaggregation is a natural way to achieve  $k$ -anonymity [11]. However, if applied to confidential attributes, microaggregation causes an undesirable variance reduction in the released data set with respect to the original data set: after the aggregation step, the variance within each cluster becomes zero<sup>1</sup>. In this respect,  $\mathbb{R}$ -*microhybrid* is clearly superior to plain microaggregation, because it *preserves the intra-cluster variances and covariances*.

Thus, *when applied to confidential attributes* and for the same clustering of the original data set,  $\mathbb{R}$ -*microhybrid* causes less information loss than plain multivariate microaggregation. Regarding disclosure risk,  $\mathbb{R}$ -*microhybrid* is no worse than plain multivariate microaggregation, as it uses the same clustering; in fact, our empirical work in [6] shows that  $\mathbb{R}$ -*microhybrid* yields a lower disclosure risk, so that it outperforms plain multivariate microaggregation in both information loss and disclosure risk.

## 5 Conclusions

We have presented a new method whose goal is to produce hybrid numerical microdata sets that can be released with low disclosure risk and acceptable data utility. The method combines microaggregation and the IPSO synthetic generator. Depending on a single integer parameter  $k$ , it can yield data which are very close to the original data (for small  $k$ ) or entirely synthetic data (when  $k$  is equal to the number of records in the data set). Thus, the parameterization of the method is simpler and more intuitive for users than the alternatives proposed so far.

We have shown that the hybrid data set obtained *preserves the mean vector and the covariance matrix of the original data set*. Furthermore, if a good microaggregation heuristic yielding a small intra-cluster variance is used:

- Approximate preservation of third-order central moments has been proven;
- Approximate preservation of fourth-order central moments is empirically shown in [6];
- For subdomains (*i.e.* data subsets), approximate preservation of means, variances, covariances, and third-order and fourth-order central moments is empirically shown in [6]; this feature was not offered by the current hybrid or synthetic data generation methods in the literature.

---

<sup>1</sup>The better the microaggregation heuristic used in the partition step, the smaller is the intra-cluster variance before aggregation, and the smaller is the information loss caused by enforcing a zero variance in the partition step.

Last but not least, compared to plain multivariate microaggregation, the new method offers better data utility for confidential attributes (due to variance and covariance preservation) and at the same time it achieves a slightly lower disclosure risk.

## Disclaimer and acknowledgments

The authors are with the UNESCO Chair in Data Privacy, but they are solely responsible for the views expressed in this paper, which do not necessarily reflect the position of UNESCO nor commit that organization. We are indebted to Dr. Josep M. Mateo-Sanz for his insightful comments. This work was partly supported by the Spanish Government through projects TSI2007-65406-C03-01 “E-AEGIS” and CONSOLIDER INGENIO 2010 CSD2007-00004 “ARES”, by the Government of Catalonia under grant 2009 SGR 1135 and by Eurostat-European Commission under grant no. 25200.2005.003-2007.670 “ESSnet on Statistical Disclosure Control”. The first author is partially supported as an ICREA Acadèmia researcher by the Government of Catalonia.

## References

- [1] Aggarwal, C. C. and Yu, P. S. (2004) “A condensation approach to privacy preserving data mining”. In E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, and E. Ferrari, editors, *Advances in Database Technology - EDBT 2004*, volume 2992 of *Lecture Notes in Computer Science*, pages 183–199, Berlin Heidelberg: Springer.
- [2] Burrige, J. (2003) “Information preserving statistical obfuscation”. *Statistics and Computing*, 13:321–327.
- [3] Dandekar, R., Cohen, M. and Kirkendall, N. (2002) “Sensitive micro data protection using Latin hypercube sampling technique”. In J. Domingo-Ferrer, editor, *Inference Control in Statistical Databases*, volume 2316 of *Lecture Notes in Computer Science*, pages 245–253, Berlin Heidelberg: Springer.
- [4] Defays, D. and Nanopoulos, P. (1993) “Panels of enterprises and confidentiality: the small aggregates method”. In *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys*, pages 195–204, Statistics Canada.
- [5] Domingo-Ferrer, J. (2008) “A survey of inference control methods for privacy-preserving data mining”. In C. C. Aggarwal and P. Yu, editors, *Privacy-Preserving Data Mining: Models and Algorithms*, volume 34 of *Advances in Database Systems*, pages 53–80, New York: Springer.
- [6] Domingo-Ferrer, J. and González-Nicolás, Ú. (2009) “Hybrid microdata using microaggregation” (manuscript).

- [7] Domingo-Ferrer, J. and Mateo-Sanz, J. M. (2002) “Practical data-oriented microaggregation for statistical disclosure control”. *IEEE Transactions on Knowledge and Data Engineering*, 14(1):189–201.
- [8] Domingo-Ferrer, J., Sebé, F. and Solanas, A. (2008) “A polynomial-time approximation to optimal multivariate microaggregation”. *Computers & Mathematics with Applications*, 55(4):714–732.
- [9] Domingo-Ferrer, J. and Torra, V. (2005). “Ordinal, continuous and heterogeneous  $k$ -anonymity through microaggregation”. *Data Mining and Knowledge Discovery*, 11(2):195–212.
- [10] Muralidhar, K. and Sarathy, R. (2008) “Generating sufficiency-based non-synthetic perturbed data”. *Transactions on Data Privacy*, 1(1):17–33. <http://www.tdp.cat/issues/tdp.a005a08.pdf>.
- [11] Samarati, P. (2001) “Protecting respondents’ identities in microdata release”. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027.
- [12] Willenborg, L. and DeWaal, T. (2001) *Elements of Statistical Disclosure Control*. New York: Springer-Verlag.
- [13] Winkler, W. E. (2004) “Re-identification methods for masked microdata”. In J. Domingo-Ferrer and V. Torra, editors, *Privacy in Statistical Databases*, volume 3050 of *Lecture Notes in Computer Science*, pages 216–230, Berlin Heidelberg: Springer.

# Dealing with edit constraints in microdata protection: microaggregation

Vicenç Torra, Isaac Cano, Guillermo Navarro-Arribas

IIIA - Artificial Intelligence Research Institute,  
CSIC - Spanish Council for Scientific Research,  
Campus UAB s/n, 08193 Bellaterra (Catalonia, Spain)  
({vtorra, cano, guille}@iiia.csic.es)

## Abstract.

In this paper we discuss how most edit constraints can be taken into account in an effective way through microaggregation. We discuss different edit constraints and some variations of microaggregation that permits to deal with such constraints. We will also present our software to formalize and deal with such constraints in an automatic way.

## 1 Introduction

When perturbation methods are used to protect statistical data they can introduce undesirable errors in the data. For instance, data editing [9, 12, 4] is a field of statistical disclosure control that is devoted to the analysis and correction of raw data for their improvement. The basic idea is that data should satisfy a set of requirements (or constraints) before their release. E.g. non negative values are not permitted for people's age. Data editing is typically applied to the original data and, in any case, before any perturbation takes place. Thus, the perturbation can introduce inconsistency in the data.

The study of perturbation methods in the presence of data edits has not been considered until very recently [16, 14]. We provide a discussions about some data edits and how can they be preserved in microaggregation. We also describe a framework to automate the microaggregation of constrained data, which uses XML as the base format both for the microdata to be processed and to express the edit constraints. The framework identifies the required edit constraints and applies modifications of the microaggregation method in order to perturb the data while satisfying the edit constraints.

The structure of the paper is as follows. Section 2 provides an overview of microaggregation, and Section 3 discusses the data edits in microaggregation. Section 4, and 5 presents our implementation and results, and Section 6 concludes the paper.

## 2 Overview of microaggregation

In this paper we show how microaggregation [3] can cope with edit constraints.

From the operational point of view, microaggregation is defined in terms of partition and aggregation:

- **Partition.** Records are partitioned into several clusters, each of them consisting of at least  $k$  records.
- **Aggregation.** For each of the clusters a representative (the centroid) is computed, and then original records are replaced by the representative of the cluster to which they belong to.

In most cases microaggregation is applied to numeric data, even so, it can also be applied to categorical data [15], either nominal or ordinal [8]. Moreover microaggregation normally considers a crisp partition of the records (as the k-means clustering), but there is also some works that do consider the use of fuzzy  $c$ -means to partition the dataset [19], and then aggregate the records accordingly. Although our work is mainly focused on crisp clustering with numeric data, we will also consider other possibilities if appropriate.

In the rest of the paper we will use the following notation. We consider a microdata file with  $n$  records  $x_1, \dots, x_n$  that take values over a set of variables  $V_1, \dots, V_m$ . We express the value for record  $x_i$  in variable  $V_j$  by  $x_{i,j}$ .

The function  $\mathbb{C}$  is the cluster representative or centroid, which we assume to be a function of the data in the cluster. More specifically, we presume that the representative of the variable  $V$  is a function of the values of the records for  $V$ , that is,  $\mathbb{C}(x_1, \dots, x_N)$ . Similarly, the representative for variable  $V_i$  is  $\mathbb{C}(x_{1,i}, \dots, x_{N,i})$ .

Note that in most cases edit constraints are preserved by providing a specific function  $\mathbb{C}$ , which preserves the constraint.

## 3 Data editing and microaggregation

Data editing can broadly be defined as the process of detecting errors in statistical data [2]. In general the whole data editing process can be very costly, even requiring human supervision in some stages [4]. For this reason it is very desirable that the statistical disclosure control methods used in edited data do not introduce new errors, so data does not need to be edited again.

The editing process is usually formalized as a set of edit constraints, that the data should satisfy. We present the generic classification of edit constraints from [16], and show their applicability in a slightly modified version of the Census data set [1]. The modification of the dataset is minimal and restricted to the addition of three variables in order to be able to show the applicability of the types of data edits.

We depart from an XML representation of the microdata. Our microdata file has a simple generic format, where data are stored by rows, and each value is an

element labeled with the variable name. It resembles most of current XML standards for data spreadsheets such as Office OpenXML or OpenDocument. The simplicity of this format and the availability of a great number of tools for XML processing makes it easy to obtain it from not only other similar XML files but also directly from database tables, or more generic microdata files.

In order to provide a standardized and already in-use language to represent the edit constraints we have used Schematron [10], which is a rule-based validation language for XML. Unlike common schema languages for XML such as W3C Schema, RELAX NG, or DTDs, which can express rules about the structure of the document, Schematron also provides semantic rules, which makes it very suitable to express edit constraints.

Schematron expresses pattern rules both as *asserts* or *reports*. An *assert* tags positive assertions about the document, while the *report* tags negative assertions. The assertions themselves are declared as the attribute *test* with an XPath [20] expression. We have chosen to express edit constraints as Schematron *asserts*, because they do more clearly express the semantic of the constraint, but *reports* could also be used to achieve the same effect. These Schematron rules can be easily checked against the XML microdata file directly and automatically by an XSLT engine.

What follows is a description of the different data edit constraints, how are they encoded as Schematron rules<sup>1</sup>, and how microaggregation can be applied to preserve the constraints.

### 3.1 Linear constraints (EC-LC)

A variable can be expressed as a linear combination of a set of other variables. For example, the following relation between *family income*, *person income*, and *other persons income* should hold (cf. Fig. 1):

$$\begin{aligned} \text{EC-LC: } & \textit{other person income} + \textit{person income} \\ & = \textit{family income} \\ & \Rightarrow V9 + V10 = V14 \end{aligned}$$

A linear constraint can be expressed, if we assume that  $V$  is the dependent variable, as  $V = \sum_{i=1}^K \alpha_i V_i$ , for some values  $\alpha_i$  and variables  $V_i$ .

Assuming that the original data (already edited) satisfies the linear constraint, so  $x_j = \sum_{i=1}^K \alpha_i x_{j,i}$ , we need to consider which function is suitable for computing the cluster representative. The most general solution for  $\mathbb{C}$  in this case is,

$$\mathbb{C}(x_1, \dots, x_N) = \frac{1}{N} \sum_{i=1}^N x_i \quad (1)$$

---

<sup>1</sup>Note that in these examples, the XPath expression in the test attribute, uses the entities '&lt;'; and '&gt;'; as the symbols '<', and '>', because the expression is contained in a string.

```

<pattern name="EC-LC: V9+V10 = V14" >
  <rule context="Record" >
    <assert test="1.0 * number(V9)
      + 1.0 * number(V10)
      = number(V14)" >
      Linear Constraint
      total_others <value-of select="V9" />
      + total_personal <value-of select="V10" />
      != total_family <value-of select="V14" />
    </assert>
  </rule>
</pattern>

```

Figure 1: Schematron rule for EC-LC.

```

<pattern name="EC-NLC: V5*V15 = V16" >
  <rule context="Record" >
    <assert test="number(V5)
      * number(V15)
      = number(V16)" >
      Non-linear Constraint
      fed. income tax <value-of select="V5" />
      * inv. state income tax
      <value-of select="V15" />
      != fed./state ratio
      <value-of select="V16" />
    </assert>
  </rule>
</pattern>

```

Figure 2: Schematron rule for EC-NLC.

Note that it coincides with the arithmetic mean.

Preservation of linear constraints in fuzzy microaggregation (microaggregation based on fuzzy clustering algorithms), can also be achieved. In [17], the authors provide a fuzzy c-means algorithm, which preserves linear constraints.

### 3.2 Non-linear constraints (EC-NLC)

Some numerical variables satisfy a non-linear relation. For example (cf. Fig. 2):

$$\begin{aligned}
 \text{EC-NLC: } & \text{fed. inc. tax} * \text{inv. state inc. tax} \\
 & = \text{ratio fed.-state inc. tax} \\
 & \Rightarrow V5 * V15 = V16
 \end{aligned}$$

In this case we can follow the same approach considering multiplicative variables. Formally, we consider variables  $V, V_1, \dots, V_K$  satisfying  $V = \prod_{i=1}^K V_i^{\alpha_i}$ . In this case the most general solution for  $\mathbb{C}$  is,

$$\mathbb{C}(x_1, \dots, x_N) = \prod_{i=1}^N x_i^{1/N} \tag{2}$$

Note that it coincides with the geometric mean.

### 3.3 Constraints on the possible values (EC-PV)

The values of a given variable are restricted to a predefined set. For example, stating that the value of variable *employer contribution for health care* should be in the interval  $[0, 7500]$  (cf. Fig. 3).

$$\text{EC-PV: } \text{employer contrib. health} \in [0, 7500] \Rightarrow V3 \in [0, 7500]$$

Or for example, consider an attribute *age* where a value of 18.5 may not make sense, and only integer positive values are permitted. Other similar constraint could involve subsets of variables, which could be reformulated in similar terms.

```

<pattern name="EC-PV: V3 in [0, 7500]">
  <rule context="Record">
    <assert test="0 &lt;= number(V3) and
      number(V3) &lt;= 7500">
      Constraints on possible values
      Employer contribution for health care
      <value-of select="V3"/>
      is not in the interval [0, 7500]
    </assert>
  </rule>
</pattern>

```

Figure 3: Schematron rule for EC-PV

```

<pattern name="EC-LC: if V8 &lt; 1115 then
  V13 &lt;= V12">
  <rule context="Record">
    <assert test="not(number(V8) &lt; 1115) or
      (number(V13) &lt;= number(V12))">
      IF total person earnings
      <value-of select="V8"/>
      &lt; 1115 THEN
      total wage and salary
      <value-of select="V13"/>
      &lt;= taxable income
      <value-of select="V12"/>
      does not hold
    </assert>
  </rule>
</pattern>

```

Figure 4: Schematron rule for EC-GV.

In order to enforce constraints on the possible values, we can require the cluster representative to be in the interval defined between the minimum and the maximum of the elements in the cluster, that is, it has to satisfy *internality*. Formally,

$$\min_i x_i \leq \mathbb{C}(x_1, \dots, x_N) \leq \max_i x_i$$

Note that if the constraint is that  $x_i \in [a, b]$  for some  $a$  and  $b$ , it is clear that for edited data, we have  $x_i \in [a, b]$ , and thus, this constraint implies that  $\mathbb{C}(x_1, \dots, x_N) \in [a, b]$ . It can be proved that both Eq. (1) and Eq. (2) do satisfy internality [16].

This constraint is commonly found in categorical nominal data. E.g. *vehicle*  $\in \{\text{car}, \text{motorcycle}, \text{truck}, \dots\}$ . By using the plurality rule (or mode) as the function  $\mathbb{C}$ , this constraint is preserved. This aggregator (which can be generalised as the weighted plurality rule) selects the most frequent element from the cluster.

The microaggregation of categorical ordinal data preserving this constraint can also be achieved, by using the median (also the weighted median, or convex median) as the function  $\mathbb{C}$ .

$$\mathbb{C}(x_1, \dots, x_N) = \begin{cases} x_{\sigma(\lfloor (N+1)/2 \rfloor)} & \text{if } N \text{ is even} \\ x_{\sigma((N+1)/2)} & \text{if } N \text{ is odd} \end{cases}$$

where  $\{\sigma(1), \dots, \sigma(N)\}$  is a permutation of  $\{1, \dots, N\}$  such that  $x_{\sigma(i-1)} \leq x_{\sigma(i)}$  for all  $i = \{2, \dots, N\}$ .

Both the plurality and median operators, which satisfy internality, are commonly used in the microaggregation of categorical data [7].

### 3.4 One variable governs the possible values of another one (EC-GV)

The values of a variable are constrained by the values of another one. E.g., considering the relations between three variables *total person earnings*, *taxable income* and *amount* as (cf. Fig. 4):

EC-GV: IF *total person earnings* < 1115  
 THEN *taxable income* ≤ *amount*  
 ⇒ IF  $V_8 < 1115$  THEN  $V_{13} \leq V_{12}$   
 ⇒  $\text{not}(V_8 < 1115)$  or  $(V_{13} \leq V_{12})$

In general any monotonic function  $\mathbb{C}$ , permits us to generate a protected file with  $V_1 < V_2$  for variables  $V_1$  and  $V_2$  if in the original file it also holds  $V_1 < V_2$ . In fact,  $x_{i,j} \leq x_{i,k}$  for all  $i$  and  $j \neq k$  implies  $\mathbb{C}(x_{1,j}, \dots, x_{N,j}) \leq \mathbb{C}(x_{1,k}, \dots, x_{N,k})$ , corresponds to the monotonicity of  $\mathbb{C}$ . Note also that Eq. (1) and Eq. (2) are monotonic. Simple EC-GV constraints, such as: EC-GV1 :  $V_3 \leq V_7$  are satisfied by using a monotonic function  $\mathbb{C}$ .

In the case of categorical data, this constraint only makes sense in ordinal data, and the median is a monotonic function.

Other EC-GV constraints such as the one presented in Section 3, which can be summarized as:

IF  $V_8 < 1115$  THEN  $V_{13} \leq V_{12}$

can be satisfied by partitioning the dataset in subsets according to the antecedent in the rule, and then applying microaggregation separately to each subset using a monotonic function  $\mathbb{C}$ . In this case the data is partitioned in two sets, one with records satisfying  $V_8 < 1115$ , and the other with records with  $V_8 \geq 1115$ . Note that the same strategy works for categorical ordinal data.

### 3.5 Other types

Other classes of constraints might be considered. For example, constraints on non-numerical variables (ordinal or categorical), ...

If we consider data editing in the context of perturbative statistical disclosure control an additional constraint is normally assumed.

- *Values are restricted to exist in the domain.* Not only the values should belong to a predefined set (as in EC-PV constraints), but the values should really exist in the domain. For example due to the edit constraint EC-PV previously presented, the variable *employer contrib. health* has to be in the interval  $[0, 7500]$ , but if in the original data all values are under 500 the perturbation

introduced in the masked data cannot cause a record to have a value of 800. Another example can be an attribute with the *town o village of residence* of the individual. The protected microdata cannot introduce for example a town that was not in the original microdata. (cf. Fig. 5)

```

<pattern name="Value-domain restriction">
  <rule context="Record">
    <let name="original"
        value="document('microdata.xml')"/>
    <assert test="exists(index-of
                        ($ original //Record/V1, V1))">
      Value <value-of select="V1"/>
      is not in domain of original values for V1
    </assert>
  </rule>
</pattern>

```

Figure 5: Schematron rule for 'values are restricted to exist in the domain'.

An appropriate operator for  $\mathbb{C}$  that satisfies this constraint is the *median*, which has already been used for microaggregation in [13]. Other operators such as order statistics and boolean max-min functions [19] could also be used.

The median (as well as the order statistics) are monotonic functions. Due to this, they could also be applied in the case of constraints where one variable governs the possible values of another one (EC-GV). This monotonicity makes the median also suitable for constraints on the possible values (EC-PV). Regarding the other constraints, linear and non-linear, it is important to note that the functions introduced in Eq. (1) and Eq. (2) cannot deal with this constraint.

Note that in this case the operators discussed for categorical data (mode, and median) satisfy this constraint.

## 4 Implementation details

The constrained microaggregation has been implemented with the aim of providing an automated process to microaggregate edited data.

The original microdata from the Census dataset is processed together with the specification of the data edits as Schematron rules, and then the data is microaggregated according to the edit constraints. Finally, edit constraints can be checked in the masked file to verify the edit constraints. Note that we have only considered numerical data.

As usual in microaggregation, variables are microaggregated by groups. In our case, all variables involved in an edit constraint are grouped together. Note that for constraints EC-PV, and EC-GV, both the arithmetic mean and the geometric mean

can be used as the cluster representative function. We use the arithmetic mean in these cases because it yields better results regarding the information loss of the protected data.

## 5 Results

We have considered two scenarios. In scenario *S1* we have microaggregated the file considering all edit constraints and the remaining variables, the ones that are not involved in any edit constraint, are microaggregated together grouping them in groups of size 3. In the scenario *S2* we have microaggregated the whole dataset without taking into account the edit constraints using the arithmetic mean to compute  $\mathbb{C}$  and again making groups of 3 variables.

For each scenario we have measured its utility and protection. To compute the information loss the Probabilistic Information Loss [11] (*PIL*) measure has been used. On the other hand, to compute the disclosure risk (*DR*), it was taken into account two broadly used measures, the Distance Based Record Linkage (*DBRL*) and the Interval Disclosure (*ID*) [5, 6]. Hence the average disclosure risk is the arithmetic mean of *DBRL* and *ID*. Finally, the *SCORE* is computed as a mean of the *PIL* and *DR*.

The experiments have shown that microaggregating while considering the edit constraints slightly affects the information loss and disclosure risk. Usually, the desired value of  $k$  is taken between 4 and 10. As it can be seen, in our case, for such values of  $k$  the *SCORE* values are very similar. Although in *S1* the minimum *SCORE* is 37.223 for  $k = 10$ , all other *SCORE* values for the range  $k \in [4, 10]$  are closely similar. The same applies to scenario *S2*, where the minimum *SCORE* is 34.531 for  $k = 6$ . Note that the lower score, the better, and that only scores under 50 are of interest (this is the score of unprotected data). The difference between both minimum values of the *score* is compensated with a preservation of the edit constraints in the original and masked dataset in case of scenario *S1*. Moreover, a score of 37.223 (from scenario *S1*) is considered a good one, providing a proper trade off between information loss and disclosure risk.

To get a graphical representation of *PIL*, *DR*, and *SCORE* we have plotted in the Figs. 6 and Fig. 7 their relationship for all  $k$  values from 1 to 99. In these figures it is shown that the *score* remains almost constant because of the greater the information loss the lower the disclosure risk in almost the same proportion.

## 6 Conclusions

In this paper we have proposed a new framework for the automatic perturbation of data through microaggregation taking into account the requirements or the constraints that the data elements have to satisfy. We have assessed the information loss and disclosure risk when considering or not the edit constraints in the microag-

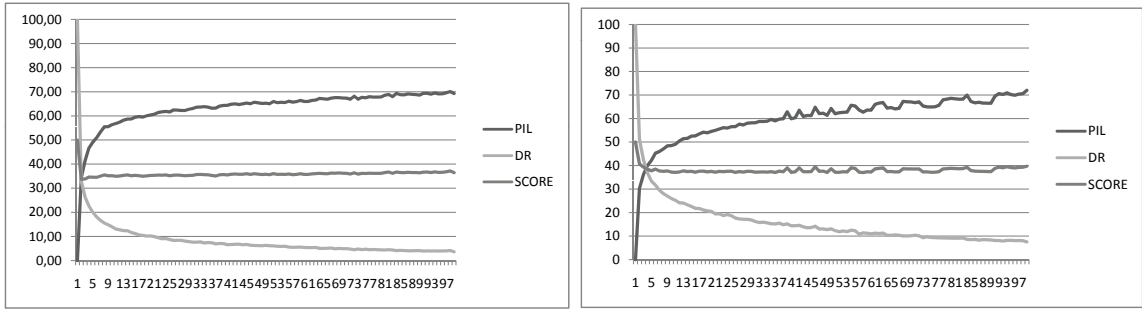


Figure 6: Scatter plot showing PIL and Figure 7: Scatter plot showing the relationship between PIL and DR with respect to group's size  $k$  for scenario  $S1$ .  
 relationship between PIL and DR with respect to group's size  $k$  for scenario  $S2$ .

gregation process. We have presented the results when microaggregating taking into account 4 different types of edit constraints,  $EC - PV$ ,  $EC - GV$ ,  $EC - LC$  and  $EC - NLC$ . As future work we consider the extension of the approach to deal with categorical attributes and to support more edit constraints.

## Acknowledgments

Partial support by the Spanish MICINN (projects eAEGIS TSI2007-65406-C03-02, ARES - CONSOLIDER INGENIO 2010 CSD2007-00004, and TSI2006-03481), and the "Institut d'Estadística de Catalunya (IDESCAT)" is acknowledged.

## References

- [1] U.S. Census Bureau. Data Extraction System. <http://www.census.gov/>.
- [2] Chambers, R. Evaluation criteria for statistical editing and imputation. *National Statistics Methodological* series No.28, Jan 2001.
- [3] Deejay's, D., Nanopoulos, P. Panels of enterprises and confidentiality: the small aggregates method, in *Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada*, 1993, pp. 195–204.
- [4] De Waal, T. An overview of statistical data editing. Statistics Netherlands. 2008.
- [5] Domingo-Ferrer, J., Torra, V., (2001) "A quantitative comparison of disclosure control methods for microdata, Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies,". Elsevier, pp. 111 – 133.
- [6] Domingo-Ferrer, J., Torra, V. (2001) "Disclosure control methods and information loss for microdata, Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies", Elsevier, pp. 91 – 110.

- [7] Domingo-Ferrer, J., Torra, V., (2005), “Ordinal, continuous and heterogeneous k-anonymity through microaggregation,” *Data Mining and Knowledge Discovery*, vol. 11, no. 2, pp. 195 – 212.
- [8] Domingo-Ferrer, J., Torra, V. Ordinal, continuous and heterogeneous k-anonymity through microaggregation, *Data Mining and Knowledge Discovery*, pp. 195–212, Jan. 2005.
- [9] Granquist, L. The new view on editing, *Int. Statistical Review* 65:3 381-387. 1997.
- [10] ISO/IEC. Information technology – Document Schema Definition Language (DSDL) – Part 3: Rule-based validation – Schematron. ISO/IEC 19757-3:2006 Standard JTC1/SC34, 2006.
- [11] Mateo-Sanz, J.M., Domingo-Ferrer, J., Sebé, F. “Probabilistic information loss measures in confidentiality protection of continuous microdata,” *Data Mining and Knowledge Discovery*, vol. 11, pp. 181 – 193. Sep 2005. ISSN: 1384-5810
- [12] Pierzchala, M. A review of the state of the art in automated data editing and imputation, in *Statistical Data Editing*, Vol. 1, Conference of European Statisticians Statistical Standards and Studies N. 44, UN Statistical Commission and Economic Commission for Europe, 10-40. 1995
- [13] Sande, G. Exact and approximate methods for data directed microaggregation in one or more dimensions, *Int. J. of Unc., Fuzz. and Knowledge Based Systems* 10:5 459–476. 2002.
- [14] Shlomo, N., De Waal, T. Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. *Journal of Official statistics*. Vol. 24, No. 2, pp. 229–253. 2008.
- [15] Torra, V. Microaggregation for categorical variables: A median based approach, in *Proc. Privacy in Statistical Databases (PSD 2004)*, ser. LNCS, vol. 3050, Jan. 2004, pp. 162–174.
- [16] Torra, V. Constrained microaggregation: Adding constraints for data editing, *Transactions on Data Privacy*, vol. 1, no. 2, pp. 175–186, 2008.
- [17] V. Torra, On the Definition of Linear Constrained Fuzzy c-Means, *Proc. of the EUROFUSE 2009* Sep. 2009, pp. 61-66.
- [18] Torra, V., Miyamoto, S. Evaluating fuzzy clustering algorithms for microdata protection. in *Proc. Privacy in Statistical Databases (PSD 2004)*, ser. LNCS, vol. 3050, Jan. 2004, pp. 162–174.
- [19] Torra, V., Narukawa, Y. *Modeling decisions: information fusion and aggregation operators*, Springer. 2007.
- [20] W3C. XML Path Language (XPath) 2.0. W3C Recommendation, Jan. 2007.