Information Sciences 180 (2010) 2834–2844

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Hybrid microdata using microaggregation ☆

Josep Domingo-Ferrer *, Úrsula González-Nicolás

*Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, UNESCO Chair in Data Privacy, Av. Päisos Catalans 26, E-43007 Tarragona, Catalonia, Spain*

ARTICLE INFO

*Article history:*
Received 20 April 2009
Received in revised form 25 February 2010
Accepted 10 April 2010

*Keywords:*
Statistical disclosure control
Microdata protection
Privacy-preserving data mining
Synthetic data
Hybrid data
Microaggregation

ABSTRACT

Statistical disclosure control (also known as privacy-preserving data mining) of microdata is about releasing data sets containing the answers of individual respondents protected in such a way that: (i) the respondents corresponding to the released records cannot be re-identified; (ii) the released data stay analytically useful. Usually, the protected data set is generated by either masking (i.e. perturbing) the original data or by generating 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 obtained by 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 hybrid data by combining microaggregation with any synthetic data generator. We show that numerical hybrid data exactly preserving means and covariances of original data and approximately preserving other statistics as well as some subdomain analyses can be obtained as a particular case with a very simple parameterization. The new method is competitive versus both the literature on hybrid data and plain multivariate microaggregation.



## 1. Introduction

One of the purposes of statistical disclosure control (SDC, [7,15,28]) is to protect static individual data, also called microdata. It is only recently that data collectors (statistical agencies and the like) have been persuaded to publish microdata. Before publishing data sets containing data on individual respondents (who can be people or organizations like enterprises), data should be protected in such a way that the respondents corresponding to the released records cannot be re-identified.

A microdata set **V** can be viewed as a file with $n$ records, where each record contains a number of attributes on an individual respondent. According to [23], attributes can be classified into the following non-disjoint categories: identifiers, quasi-identifiers/key attributes, confidential attributes and non-confidential attributes. *Identifiers* unambiguously identify the respondent; examples are the passport number, social security number, name-surname, etc. *Quasi-identifiers* or *key attributes* identify the respondent with some degree of ambiguity, but a combination of quasi-identifiers may provide unambiguous identification; examples are address, gender, age, telephone number, etc. *Confidential attributes* contain sensitive information

☆ 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.

* Corresponding author. Tel.: +34 977558109; fax: +34 977559710.
*E-mail addresses:* josep.domingo@urv.cat (J. Domingo-Ferrer), ursula.gonzaleznicolas@urv.cat (Ú. González-Nicolás).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.04.005

on the respondent; examples are salary, religion, political affiliation, health condition, etc. *Non-confidential attributes* do not contain sensitive information on the respondent. We assume in what follows that original microdata sets to be protected for subsequent release have been pre-processed to remove from them all identifiers.

Given an original microdata set $\mathbf{V}$, the goal of SDC is to compute a protected microdata set $\mathbf{V}'$ that can be released and such that: (i) the risk that $\mathbf{V}'$ is useful for a user or an intruder to determine confidential attribute values for a specific individual among those in $\mathbf{V}$ (disclosure risk) is low; (ii) user analyses (regressions, means, etc.) on $\mathbf{V}'$ and $\mathbf{V}$ yield the same or at least similar results.

Note that, unlike in conventional data protection [20], SDC-protected data must stay analytically useful. Microdata protection methods (see [7,28] 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 [29,21]). On the negative side, the utility of synthetic data is critically dependent on the validity of the models that are used to generate them [21]: there is no preservation guarantee for those statistics not included in the model.

### 1.1. Contribution and plan of this article

Hybrid data obtained by mixing the original data and synthetic data have been proposed in the literature [4,18], to combine the strengths of masked and synthetic data and neutralize their pitfalls. We show how to easily obtain hybrid data by combining microaggregation [9,13] with any synthetic data generator. We show that numerical hybrid data preserving means and covariances of original data as achieved by the Muralidhar-Sarathy procedure ([18], named for short MS in the sequel) can be obtained as a particular case 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 synthetic data generators and microaggregation. Section 3 describes a generic scheme for microaggregation-based hybrid data generation. In Section 4, a specific scheme for generating numerical microaggregation-based 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 5, including a comparison with MS and plain multivariate microaggregation supported by empirical results. Section 6 is a conclusion.

## 2. Background: synthetic data and microaggregation

### 2.1. Synthetic data generation

In the early 1990s, Rubin [22] suggested creating an entirely synthetic data set based on the original survey data and multiple imputation. The interesting point about multiple imputation is that it can be applied to *any type of data*, numerical or categorical. Furthermore, in theory the distribution of the generated synthetic data is the same as the joint distribution of the original data. As to disclosure risk, it is very low, since all generated values are synthetic: it is not zero, however, because if the imputation model is too good (overfit), the synthetic data might resemble too much the original data. In general, the choice of the imputation model is a non-trivial task critical both to utility (which distribution is preserved) and to disclosure risk; see [21] for a more detailed discussion.

Another synthetic data generator, called Information Preserving Statistical Obfuscation (IPSO), was proposed in [3] to synthesize *numerical data sets*. Informally, suppose that we have a data set with numerical attributes and $n$ records; 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 $\widehat{B}$ 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 $\widehat{B}$ and $\hat{\Sigma}$, sufficient for the multivariate normal case, are preserved. Thus, synthetic data produced by IPSO preserve the means and covariances of the original data.

More detailed and recent surveys on synthetic data generation can be found in [7,8].

### 2.2. Microaggregation

Microaggregation is a family of perturbative SDC methods originally defined for continuous data [6,9] and extended for categorical data in [27,13]. Computational improvements of microaggregation by combining it with other techniques can be found in [19,17]. Whatever the data type and the computational method, microaggregation can be defined operationally 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-generic algorithm [13] for the partition step in multivariate microaggregation.

**Algorithm 1** (*MDAV-generic*). (R: data set, k: integer)

(1) While $|R| \geqslant 3k$ do
  (a) Compute the average record $\bar{x}$ of all records in $R$. The average record is computed attribute-wise.
  (b) Consider the most distant record $x_r$ to the average record $\bar{x}$ using an appropriate 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 1(d).

end while

(2) If there are between $3k-1$ and $2k$ records in $R$:
  (a) Compute the average record $\bar{x}$ of the remaining records in $R$,
  (b) find the most distant record $x_r$ from $\bar{x}$,
  (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.

Implementation of MDAV-generic for a particular attribute type requires specifying how the average record is computed and what distance is used. For numerical attributes, computing averages using the arithmetic mean and using Euclidean distances are the natural choices. For ordinal and nominal data, see [13]. A measure of variance to evaluate the homogeneity of clusters obtained on hierarchical nominal data is given in [12].

**Note**. If attributes are all continuous, 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 [11], 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. A generic procedure for generating microaggregation-based hybrid data

Let $\mathbf{V}$ be an original data set consisting of $n$ records. On input an integer parameter $k \in \{1,\ldots,n\}$, 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 yields $\mathbf{V}' = \mathbf{V}$ (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:

- A generic synthetic data generator $\mathscr{S}(\mathbf{C}, \mathbf{C}', parms)$, that is, an algorithm which, given an original data (sub) set $\mathbf{C}$, generates a synthetic data (sub) set $\mathbf{C}'$ preserving the statistics or parameters or models of $\mathbf{C}$ specified in *parms*.
- A microaggregation heuristic, which, on input a set or $n$ 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** (***microhybrid***(**V**,**V**′,*parms*,*k*)).

(1) Call microaggregation (**V**,*k*). Let $C_1,\ldots,\ C_\kappa$ for some $\kappa$ be the resulting clusters of records.
(2) For $i = 1,\ldots,\kappa$ call $\mathscr{S}(C_i, C_i', parms)$.
(3) Output a hybrid data set **V**′ whose records are those in the clusters $C_1',\ldots,C_\kappa'$.

At Step 1 of procedure *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 *microhybrid* procedure bears some resemblance to the condensation approach proposed in [1]; however, *microhybrid* is more general because:

(i) It can be applied to any data type (condensation is designed for numerical data only);
(ii) Clusters do not need to be all of size $k$ (their sizes can vary between $k$ and $2k-1$);
(iii) Any synthetic data generator (chosen to preserve certain pre-selected statistics or models) can be used by *microhybrid*;
(iv) Instead of using an *ad hoc* clustering heuristic like condensation, *microhybrid* can use any of the best microaggregation heuristics cited above, which should yield higher within-cluster homogeneity and thus less information loss.

### 3.1. On the role of parameter k

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

- If $k = 1$, and *parms* include preserving the mean of each attribute in the original clusters, 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 the synthetic data generator $\mathscr{S}$ once for the entire data set.
- For intermediate values of $k$, several clusters are obtained at Step 1, whose parameters *parms* are preserved by the synthetic clusters generated at Step 2. As $k$ decreases, the number of clusters (whose parameters 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, depending on the synthetic generator used, there may be a lower bound for $k$ *higher than* 1. For example, if using IPSO with $|X|$ confidential attributes and $|Y|$ non-confidential attributes, it turns out that $k$ must be at least $|Y| + 1$; the reason is that IPSO fits to the cluster data a multivariate linear regression model with the $|Y|$ non-confidential attributes as independent variables, and it is well-known that, in a linear regression, the sample size must be greater than the number of independent variables.

### 3.2. On the type of attributes

Procedure *microhybrid* relies on a microaggregation procedure and on a generic synthetizer $\mathscr{S}$. It has been explained above how the MDAV-generic microaggregation procedure can be adapted to numerical, categorical ordinal and categorical nominal attributes. We have also reviewed two different synthesizers: while IPSO is only suitable for numerical attributes, multiple imputation is more general and can generate synthetic attributes of any type if an appropriate imputation model is taken.

If the original data set **V** contains numerical and categorical attributes, there are several possible options:

- Run *microhybrid* separately for each attribute type: once for numerical attributes, once for categorical ordinal attributes and once for categorical nominal attributes. For each attribute type, the appropriate average operator and the appropriate distance are used within MDAV-generic; also, an appropriate synthesizer is selected (e.g. multiple imputation with a suitable imputation model for each data type and/or IPSO for numeric attributes). In this way, three hybrid data sets are independently generated for each of the three attribute types. Assuming that only the confidential attributes are hybrid-generated, so that the non-confidential attributes stay unaltered, the latter can be used to link hybrid records in the three data sets in view of obtaining a single hybrid data set with all attributes.
- Run MDAV-generic only for one attribute type (e.g. numerical attributes) and impose the clustering obtained to the other attribute types. Then run the synthesizer within each cluster for all attributes (with the proper adaptations depending on the attribute type).

- Run MDAV-generic separately for each attribute type. This yields three different cluster partitions of the original data set, one for each attribute type. Then, try to reach a "consensus partition", which can be done in several ways [25]. Finally, run the synthesizer within each consensus cluster for all attributes.

## 4. Generating numerical microaggregation-based hybrid data

In the specific case where all attributes in the data set **V** are numerical, we can use the *microhybrid* procedure with the following choices:

- Take one of the following two microaggregation procedures:
  - Either MDAV-generic with the arithmetic mean as average operator and the Euclidean distance, that is, the MDAV algorithm described in [13] and implemented in the $\mu$-Argus [14] and SDCmicro [24] packages;
  - Or the $\mu$-Approx variable-size heuristic mentioned above.
- Take IPSO as a synthetic data generator, which implies that preserved *parms* are the attribute means and covariances.

Let us call the resulting procedure $\mathbb{R}$-*microhybrid*.

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

$$\overline{\mathbf{X}} = \overline{\mathbf{X}}', \quad \Sigma_{\mathbf{X}'\mathbf{X}'} = \Sigma_{\mathbf{X}\mathbf{X}}, \quad \Sigma_{\mathbf{X}'\mathbf{Y}} = \Sigma_{\mathbf{X}\mathbf{Y}} \tag{1}$$

**Proof.** Let $n$ be the number of records in **V** and $\mathbf{V}'$. Let $k$ be the parameter used to call $\mathbb{R}$-*microhybrid*, and let $C_1,\ldots,C_\kappa$ 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 **V**. Let $\mathbf{X}_i(=X(i)_1,\ldots,X(i)_L)$ be the confidential attributes restricted to cluster $C_i$ of the original data set. Let $\mathbf{X}'_i(=X'(i)_1,\ldots,X'(i)_L)$ 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

$$\overline{\mathbf{X}}'_i = \overline{\mathbf{X}}_i, \quad i = 1,\ldots,\kappa \tag{2}$$

$$\Sigma_{\mathbf{X}'_i\mathbf{X}'_i} = \Sigma_{\mathbf{X}_i\mathbf{X}_i}, \quad i = 1,\ldots,\kappa \tag{3}$$

$$\Sigma_{\mathbf{X}'_i\mathbf{Y}_i} = \Sigma_{\mathbf{X}_i\mathbf{Y}_i}, \quad i = 1,\ldots,\kappa \tag{4}$$

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

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

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

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

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

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

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

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

where $k_i$ is the actual size of cluster $C_i$, with $k \leqslant k_i \leqslant 2k-1$. By Eq. (3), we have that $s'(i)_{lm} = s(i)_{lm}$ for all $i,l,m$. Using Eq. (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 Eq. (6) holds with equality and we have $\Sigma_{\mathbf{X'X'}} = \Sigma_{\mathbf{XX}}$. A similar argument based on Eqs. (4) and (2) shows that $\Sigma_{\mathbf{X'Y}} = \Sigma_{\mathbf{XY}}$. □

### 4.2. 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 $C_1,\ldots,C_\kappa$ are formed. In other words, a very bad multivariate microaggregation procedure (yielding very poor intra-cluster homogeneity) 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.

We substantiate this confidence in two ways:

- In this section, we show that the lower the intra-cluster variance, the more approximately are third-order central moments preserved.
- In Section 5.3 below, 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 $\{C_i: i = 1,\ldots,\kappa\}$. 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 $j = 1,\ldots,k_i$, with $k \leqslant k_i \leqslant 2k-1$, denote the values of $X$, resp. $X'$, in $C_i$. If $B$ is such that $(x_{ij}-\bar{x}_i)^2 \leqslant B$ and $(x'_{ij}-\bar{x}_i)^2 \leqslant B$ for all $i,j$ then*

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

**Proof.** After some algebraic manipulation we can write

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

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

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

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 Eq. (7) from Eq. (8) we obtain:

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

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

## 5. Usability and performance assessment

No counterparts of the generic *microhybrid* procedure are currently identifiable in the literature. Therefore, we focus only on comparing the $\mathbb{R}$-*microhybrid* procedure above for numerical microdata with the alternative MS, which also preserves means and covariances, and with plain multivariate microaggregation [9,13]. Previous proposals for hybrid data generation not guaranteeing exact preservation of means and covariances (like [4]) are not considered.

### 5.1. Comparison with the Muralidhar–Sarathy hybrid generator

We recall the procedure MS [18] using the notation of Lemma 1. In that procedure, the hybrid values are generated as

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

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

$$\boldsymbol{\beta}^T = \Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}(\mathbf{I} - \boldsymbol{\alpha}^T)$$
$$\boldsymbol{\gamma} = (\mathbf{I} - \boldsymbol{\alpha})\overline{\mathbf{X}} - \boldsymbol{\beta}\overline{\mathbf{Y}}$$
$$\Sigma_{\mathbf{ee}} = (\Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}}) - \boldsymbol{\alpha}(\Sigma_{\mathbf{XX}} - \Sigma_{\mathbf{XY}}\Sigma_{\mathbf{YY}}^{-1}\Sigma_{\mathbf{YX}})\boldsymbol{\alpha}^T$$

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

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

- While $k$ is an integer between 1 and $n$, $\boldsymbol{\alpha}$ is a matrix with real-valued components.
- While the choice of the value of $k$ is very intuitive (see Section 3.1), the authors of MS admit that $\boldsymbol{\alpha}$ must be selected carefully to ensure that $\Sigma_{\mathbf{ee}}$ is positive semidefinite. They consider three options for specifying the $\boldsymbol{\alpha}$ matrix:
  (1) Take $\boldsymbol{\alpha}$ 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 $\boldsymbol{\alpha}$ 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 $\boldsymbol{\alpha}$ until positive semidefiniteness is achieved.
  (3) Taking $\boldsymbol{\alpha}$ 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 $(X_1,\ldots,X_L)$. 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 Section 5.3);
- There is also approximate preservation *over subdomains* (data subsets) of means, variances, covariances, and third and fourth-order central moments (see Section 5.3).

### 5.2. 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 *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 [13], microaggregation is a natural way to achieve $k$-anonymity [23]. 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.[1] 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, Tables 4 and 5 in Section 5.3 below show that $\mathbb{R}$-*microhybrid* yields a lower disclosure risk. Hence, it outperforms plain multivariate microaggregation in both information loss and disclosure risk.

### 5.3. Empirical performance assessment

We implemented and compared the performance of $\mathbb{R}$-*microhybrid* based on variable-size $\mu$-Approx microaggregation against the MS hybrid data generator, using two reference data sets [2] proposed in the European project CASC:

(1) The "Census" data set contains 1080 records with 13 numerical attributes. This data set was used in CASC and in [10,5,30,16,13,11]. Within this data set, as confidential attributes $X_1$ and $X_2$ we selected FICA (Social security retirement payroll deduction) and FEDTAX (Federal income tax liability), respectively; as non-confidential attributes $Y_1$ and $Y_2$, we selected INTVAL (Amount of interest income) and POTHVAL (Total other persons income).

[1] 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.

(2) The "EIA" data set contains 4092 records with 11 numerical attributes (plus two additional categorical attributes). This data set was used in CASC, in [5,11] and partially in [16] under the name "Creta". If records are viewed as points in a multidimensional space, points in "EIA" are less evenly distributed than points in "Census"; in fact, points in "EIA" tend to form natural clusters. As confidential attributes $X_1$ and $X_2$, we selected INDREVENUE (Revenue from sales to industrial consumers) and INDSALES (Sales to industrial consumers); as non-confidential attributes $Y_1$ and $Y_2$, we selected TOTREVENUE (Revenue from sales to all consumers) and TOTSALES (Sales to all consumers).

In order to conduct a fair comparison, we empirically determined values of $k$ for our method and $\boldsymbol{\alpha}$ for MS which yielded hybrid data with a similar disclosure risk. Disclosure risk was measured by using distance-based record linkage [26] between the hybrid and the original records. A hybrid record, actually a pair of values for the hybrid attributes $(X_1', X_2')$, was linked to the original record whose values for $(X_1,X_2)$ were at shortest Euclidean distance; if the pair of linked records shared the same values for the non-confidential attributes $(Y_1,Y_2)$ the match was considered correct. The percentage of correct matches was used as a measure of disclosure risk.

For MS, we took the same values of $\boldsymbol{\alpha}$ used in the empirical work presented in [18], that is

$$\boldsymbol{\alpha}_1 = \begin{pmatrix} 0.9 & 0.0 \\ 0.0 & 0.9 \end{pmatrix}; \quad \boldsymbol{\alpha}_2 = \begin{pmatrix} 0.8 & 0.0 \\ 0.0 & 0.3 \end{pmatrix}; \quad \boldsymbol{\alpha}_3 = \begin{pmatrix} 0.0 & 0.0 \\ 0.0 & 0.0 \end{pmatrix}.$$

For each value of $\alpha$ and each original data set, 10 hybrid data sets were generated and the average percentage of correct matches was computed. For our method, different values of $k$ were tried. For each value of $k$ and each original data set, 10 hybrid data sets were generated and the average percentage of correct matches was computed.

For the "Census" and the "EIA" data sets, Table 1 gives the average percentage of correct matches with MS using parameters $\alpha_1,\alpha_2$ and $\alpha_3$. For each data set, the table also lists those values of $k$ for which our method was found to yield the most similar percentage of correct matches to MS with $\alpha_1,\alpha_2,\alpha_3$, respectively. For "Census", the values of $k$ comparable to $\alpha_1,\alpha_2,\alpha_3$ turned out to be $k = 22,23,24$, respectively. For "EIA", they turned out to be $k = 80,90,120$, respectively. Note that as $\alpha$ changes from $\alpha_1$ to $\alpha_2$ and $\alpha_3$, the resulting hybrid data are less influenced by the original data (in fact for $\alpha_3$ they are purely synthetic); therefore, it is natural that the corresponding values of $k$ are also increasing (the higher $k$ with our method, the less influenced are the resulting hybrid data by the original data).

After generating hybrid data using MS and our method with, respectively, the values of $\alpha$ and $k$ in Table 1, sampling was performed to capture utility loss when the user restricts her analysis to a subset of the data. The following steps were performed:

- For each data set and each parameter value, samples of size 10% of the data set were drawn 100 times by simple random sampling.
- Let $\Theta'$ be a statistic over the hybrid data and $\Theta$ the corresponding statistic over the original data. Let $\theta_i'$ and $\theta_i$ be, respectively, the values taken by $\Theta'$ and $\Theta$ in the $i$th sample of the hybrid data and the corresponding $i$th sample of the original data. Define the mean variation for the pair of statistics $(\Theta', \Theta)$ over the 100 sample pairs as

$$\Delta(\Theta) = \frac{1}{100}\sum_{i=1}^{100}\frac{|\theta_i' - \theta_i|}{|\theta_i|}$$

Mean variations were computed for the following pairs of statistics: mean of $X_1'$ vs $X_1$ (named $\Delta(m_1)$), mean of $X_2'$ vs $X_2$ ($\Delta(m_2)$), variance of $X_1'$ vs $X_1$ ($\Delta(\sigma_1^2)$), variance of $X_2'$ vs $X_2$ ($\Delta(\sigma_2^2)$), covariance $Y_1X_1$ vs $Y_1X_1'$($\Delta(\sigma_{11})$), covariance $Y_1X_2$ vs $Y_1X_2'$($\Delta(\sigma_{12})$), covariance $Y_2X_1$ vs $Y_2X_1'$($\Delta(\sigma_{21})$), covariance $Y_2X_2$ vs $Y_2X_2'$($\Delta(\sigma_{22})$), third-order central moment of $X_1'$ vs $X_1$($\Delta(m_1^3)$), third-order central moment of $X_2'$ vs $X_2$($\Delta(m_2^3)$), fourth-order central moment of $X_1'$ vs $X_1$($\Delta(m_1^4)$) and fourth-order central moment of $X_2'$ vs $X_2$($\Delta(m_2^4)$).

Tables 2 and 3 report, respectively for the "Census" and the "EIA" data sets, the above mean variations. The following can be seen:

**Table 1**
Comparable values of $\alpha$ and k, i.e. yielding similar percentages of correct matches for MS and $\mathbb{R}$-*microhybrid* with $\mu$-Approx microaggregation, respectively. Data sets considered: "Census" and "EIA".

| Data set | $\alpha$ | % Correct matches | $k$ | % Correct matches |
|---|---|---|---|---|
| "Census" | $\alpha_1$ | 0.2 | 22 | 0.2 |
| | $\alpha_2$ | 0.1 | 23 | 0.1 |
| | $\alpha_3$ | 0.0 | 24 | 0.0 |
| "EIA" | $\alpha_1$ | 0.6 | 80 | 0.6 |
| | $\alpha_2$ | 0.4 | 90 | 0.4 |
| | $\alpha_3$ | 0.0 | 120 | 0.0 |

**Table 2**
Average mean variation for several statistics on corresponding 10% samples of the hybrid data set and the original data set. Hybrid methods considered: MS and $\mathbb{R}$-*microhybrid* with $\mu$-Approx microaggregation, respectively. Correct matches computed for overall data set (no sampling). Data set: "Census".

| Statistic<br>% Correct matches | $\alpha_1$<br>0.60 | $k = 22$<br>0.60 | $\alpha_2$<br>0.40 | $k = 23$<br>0.40 | $\alpha_3$<br>0.00 | $k = 24$<br>0.00 |
|---|---|---|---|---|---|---|
| $\Delta(m_1)$ | 0.0168 | 0.0029 | 0.0175 | 0.0036 | 0.0584 | 0.0048 |
| $\Delta(m_2)$ | 0.0203 | 0.0048 | 0.0215 | 0.0056 | 0.0699 | 0.0066 |
| $\Delta(\sigma_1^2)$ | 0.0694 | 0.0131 | 0.0747 | 0.0155 | 0.1253 | 0.0199 |
| $\Delta(\sigma_2^2)$ | 0.0625 | 0.0200 | 0.0770 | 0.0246 | 0.1188 | 0.0347 |
| $\Delta(\sigma_{11})$ | 0.2097 | 0.0293 | 0.2584 | 0.0333 | 0.5267 | 0.0443 |
| $\Delta(\sigma_{12})$ | 0.1722 | 0.0346 | 0.2522 | 0.0524 | 0.4723 | 0.0488 |
| $\Delta(\sigma_{21})$ | 0.3600 | 0.0595 | 0.4364 | 0.0870 | 1.2375 | 0.1619 |
| $\Delta(\sigma_{22})$ | 0.2437 | 0.0736 | 0.3223 | 0.2937 | 1.2618 | 0.1637 |
| $\Delta(m_1^3)$ | 3.4807 | 0.5501 | 5.1872 | 0.8345 | 62.6002 | 0.5509 |
| $\Delta(m_2^3)$ | 0.1990 | 0.0309 | 0.2303 | 0.0345 | 0.5332 | 0.0427 |
| $\Delta(m_1^4)$ | 0.5597 | 0.1192 | 0.6987 | 0.2022 | 1.9927 | 0.2412 |
| $\Delta(m_2^4)$ | 0.1830 | 0.0600 | 0.2458 | 0.0901 | 0.4939 | 0.1165 |

**Table 3**
Average mean variation for several statistics on corresponding 10% samples of the hybrid data set and the original data set. Hybrid methods considered: MS and $\mathbb{R}$-*microhybrid* with $\mu$-Approx microaggregation, respectively. Correct matches computed for the overall data set (no sampling). Data set: "EIA".

| Statistic<br>% Correct matches | $\alpha_1$<br>0.20 | $k = 80$<br>0.20 | $\alpha_2$<br>0.10 | $k = 90$<br>0.10 | $\alpha_3$<br>0.00 | $k = 100$<br>0.00 |
|---|---|---|---|---|---|---|
| $\Delta(m_1)$ | 0.0129 | 0.0147 | 0.0150 | 0.0150 | 0.0444 | 0.0165 |
| $\Delta(m_2)$ | 0.0118 | 0.0113 | 0.0231 | 0.0123 | 0.0378 | 0.0155 |
| $\Delta(\sigma_1^2)$ | 0.0195 | 0.0705 | 0.0276 | 0.0732 | 0.0919 | 0.0849 |
| $\Delta(\sigma_2^2)$ | 0.0172 | 0.0435 | 0.0164 | 0.0559 | 0.0812 | 0.0665 |
| $\Delta(\sigma_{11})$ | 0.0128 | 0.0363 | 0.0204 | 0.0424 | 0.0670 | 0.0509 |
| $\Delta(\sigma_{12})$ | 0.0132 | 0.0344 | 0.0204 | 0.0387 | 0.0641 | 0.0443 |
| $\Delta(\sigma_{21})$ | 0.0116 | 0.0303 | 0.0300 | 0.0378 | 0.0608 | 0.0442 |
| $\Delta(\sigma_{22})$ | 0.0112 | 0.0297 | 0.0289 | 0.0330 | 0.0599 | 0.0391 |
| $\Delta(m_1^3)$ | 0.0691 | 0.1838 | 0.0860 | 0.1964 | 0.3225 | 0.2170 |
| $\Delta(m_2^3)$ | 0.0807 | 0.2854 | 0.0949 | 0.3379 | 0.3493 | 0.3671 |
| $\Delta(m_1^4)$ | 0.0532 | 0.1259 | 0.0695 | 0.1381 | 0.2482 | 0.1520 |
| $\Delta(m_2^4)$ | 0.0552 | 0.1978 | 0.0695 | 0.2272 | 0.2668 | 0.2390 |

**Table 4**
Average mean variation for several statistics on corresponding 10% samples of the hybrid data set and the original data set. Hybrid method: $\mathbb{R}$-*microhybrid* with $\mu$-Approx microaggregation and several values of $k$. Correct matches computed for the overall data set (no sampling); between parenthesis, percentage of correct matches if plain multivariate microaggregation was used. Data set: "Census".

| Statistic<br>% Correct matches | $k = 7$<br>3.30 (11.11) | $k = 10$<br>2.00 (7.69) | $k = 15$<br>1.00 (5.00) | $k = 20$<br>0.40 (3.70) | $k = 22$<br>0.20 (3.52) | $k = 23$<br>0.10 (3.43) | $k = 24$<br>0.00 (2.87) |
|---|---|---|---|---|---|---|---|
| $\Delta(m_1)$ | .0012 | .0016 | .0023 | .0035 | .0029 | .0036 | .0048 |
| $\Delta(m_2)$ | .0017 | .0024 | .0033 | .0052 | .0048 | .0056 | .0066 |
| $\Delta(\sigma_1^2)$ | .0042 | .0074 | .0101 | .0151 | .0131 | .0155 | .0199 |
| $\Delta(\sigma_2^2)$ | .0063 | .0110 | .0151 | .0230 | .0200 | .0246 | .0347 |
| $\Delta(\sigma_{11})$ | .0084 | .0126 | .0206 | .0464 | .0293 | .0333 | .0443 |
| $\Delta(\sigma_{12})$ | .0070 | .0137 | .0219 | .0507 | .0346 | .0524 | .0488 |
| $\Delta(\sigma_{21})$ | .0144 | .0247 | .0488 | .0584 | .0595 | .0870 | .1619 |
| $\Delta(\sigma_{22})$ | .0220 | .0261 | .0407 | .0513 | .0736 | .2937 | .1637 |
| $\Delta(m_1^3)$ | .1940 | .2377 | .3057 | .6841 | .5501 | .8345 | .5509 |
| $\Delta(m_2^3)$ | .0108 | .0200 | .0236 | .0376 | .0309 | .0345 | .0427 |
| $\Delta(m_1^4)$ | .0404 | .0853 | .1059 | .1066 | .1192 | .2022 | .0901 |
| $\Delta(m_2^4)$ | .0202 | .0412 | .0473 | .0517 | .0600 | .0901 | .1165 |

- For "Census", $\mathbb{R}$-*microhybrid clearly* outperforms MS for all parameter values and for all considered statistics. "Clear outperformance" means that the mean variation using MS is between 5 and 10 times greater than the mean variation using $\mathbb{R}$-*microhybrid* for all statistics, except for $\Delta(m_1^3)$ when $\alpha_3$ and $k = 24$ are taken, in which case it is more than 100 times greater.

**Table 5**
Average mean variation for several statistics on corresponding 10% samples of the hybrid data set and the original data set. Hybrid method: $\mathbb{R}$-*microhybrid* with $\mu$-Approx microaggregation and several values of $k$. Correct matches computed for the overall data set (no sampling); between parenthesis, percentage of correct matches if plain multivariate microaggregation was used. Data set: "EIA".

| Statistic<br>% Correct matches | $k = 10$<br>7.80 (7.50) | $k = 15$<br>4.50 (5.06) | $k = 20$<br>3.30 (3.76) | $k = 80$<br>0.60 (0.93) | $k = 90$<br>0.40 (0.83) | $k = 120$<br>0.00 (0.64) |
|---|---|---|---|---|---|---|
| $\Delta(m_1)$ | .0047 | .0079 | .0080 | .0147 | .0150 | .0165 |
| $\Delta(m_2)$ | .0035 | .0060 | .0060 | .0113 | .0123 | .0155 |
| $\Delta(\sigma_1^2)$ | .0311 | .0408 | .0426 | .0705 | .0732 | .0849 |
| $\Delta(\sigma_2^2)$ | .0154 | .0262 | .0297 | .0435 | .0559 | .0665 |
| $\Delta(\sigma_{11})$ | .0174 | .0232 | .0238 | .0363 | .0424 | .0509 |
| $\Delta(\sigma_{12})$ | .0116 | .0159 | .0177 | .0344 | .0387 | .0443 |
| $\Delta(\sigma_{21})$ | .0131 | .0190 | .0201 | .0303 | .0378 | .0442 |
| $\Delta(\sigma_{22})$ | .0090 | .0135 | .0159 | .0297 | .0330 | .0391 |
| $\Delta(m_1^3)$ | .0929 | .0975 | .1200 | .1838 | .1964 | .2170 |
| $\Delta(m_2^3)$ | .1683 | .1604 | .2126 | .2854 | .3379 | .3671 |
| $\Delta(m_1^4)$ | .0363 | .0627 | .0728 | .1259 | .1381 | .1520 |
| $\Delta(m_2^4)$ | .0570 | .1005 | .1145 | .1978 | .2272 | .2390 |

- For "EIA", $\mathbb{R}$-*microhybrid* with $k = 100$ *slightly* outperforms MS for all statistics with $\alpha_3$, but is *slightly* outperformed by MS for all statistics with the other two parameter choices ($\alpha_1$,$k = 80$ and $\alpha_2$,$k = 90$, respectively). "Slight outperformance" means that the ratio between mean variations is between 1 and 3 in either sense.

An interesting feature of $\mathbb{R}$-*microhybrid* method is that the trade-off between disclosure risk and information loss depends only on parameter $k$. Tables 4 and 5 show, respectively for "Census" and "EIA", how taking smaller values of $k$ increases the disclosure risk (percentage of correct matches) and reduces information loss (mean variation for the considered statistics). Furthermore, the tables show that the disclosure risk incurred by $\mathbb{R}$-*microhybrid* is lower than the one incurred by plain multivariate microaggregation for the same $k$.

## 6. Conclusions

We have presented a new hybrid data generation method whose goal is to produce hybrid microdata sets that can be released with low disclosure risk and acceptable data utility. The method combines microaggregation and any synthetic data generator. Depending on a single integer parameter $k$, it can yield data which are very close to the original data (or even the original data themselves if $k = 1$) 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 in the hybrid data generation alternatives proposed so far.

For the specific case of numerical microdata, 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 has been empirically shown;
- For subdomains (i.e. data subsets), approximate preservation of means, variances, covariances, and third-order and fourth-order central moments has been empirically shown; this feature was not offered by the current hybrid or synthetic data generation methods in the literature.

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 lower disclosure risk.

## Acknowledgements

We are indebted to Dr. Josep M. Mateo-Sanz for his insightful comments. This work was partly supported by the Government of Catalonia under grant 2009 SGR 1135, by the Spanish Government through projects TSI2007–65406-C03–01 "E-AEGIS" and CONSOLIDER INGENIO 2010 CSD2007–00004 "ARES", and by the European Comission 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] C.C. Aggarwal, P.S. Yu, A condensation approach to privacy-preserving data mining, in: E. Bertino, S. Christodoulakis, D. Plexousakis, V. Christophides, M. Koubarakis, K. Böhm, E. Ferrari (Eds.), Advances in Database Technology – EDBT 2004, Lecture Notes in Computer Science, vol. 2992, Berlin Heidelberg, 2004, pp. 183–199.

[2] R. Brand, J. Domingo-Ferrer, J.M. Mateo-Sanz, Reference data sets to test and compare SDC methods for protection of numerical microdata, European Project IST-2000-25069 CASC, 2002. <http://neon.vb.cbs.nl/casc>.
[3] J. Burridge, Information preserving statistical obfuscation, Statistics and Computing 13 (2003) 321–327.
[4] R. Dandekar, M. Cohen, N. Kirkendall, Sensitive micro data protection using latin hypercube sampling technique, in: J. Domingo-Ferrer (Ed.), Inference Control in Statistical Databases, Lecture Notes in Computer Science, vol. 2316, Springer, Berlin Heidelberg, 2002, pp. 245–253.
[5] R. Dandekar, J. Domingo-Ferrer, F. Sebé, LHS-based hybrid microdata vs rank swapping and microaggregation for numeric microdata protection, in: J. Domingo-Ferrer (Ed.), Inference Control in Statistical Databases, Lecture Notes in Computer Science, vol. 2316, Springer, Berlin Heidelberg, 2002, pp. 153–162.
[6] D. Defays, P. Nanopoulos, Panels of enterprises and confidentiality: the small aggregates method, in: Proceedings of 92 Symposium on Design and Analysis of Longitudinal Surveys, Statistics Canada, Ottawa, 1993, pp. 195–204.
[7] J. Domingo-Ferrer, A survey of inference control methods for privacy-preserving data mining, in: C.C. Aggarwal, P. Yu (Eds.), Privacy-Preserving Data Mining: Models and Algorithms, Advances in Database Systems, vol. 34, Springer, New York, 2008, pp. 53–80.
[8] J. Domingo-Ferrer, J. Drechsler, S. Polettini, Report on synthetic data files, Technical report, ESSNET-SDC Project, 2009. <http://neon.vb.cbs.nl/casc>.
[9] J. Domingo-Ferrer, J.M. Mateo-Sanz, Practical data-oriented microaggregation for statistical disclosure control, IEEE Transactions on Knowledge and Data Engineering 14 (1) (2002) 189–201.
[10] J. Domingo-Ferrer, J.M. Mateo-Sanz, V. Torra, Comparing SDC methods for microdata on the basis of information loss and disclosure risk, in: Pre-proceedings of ETK – NTTS'2001, vol. 2, Eurostat, Luxemburg, 2001, pp. 807–826.
[11] J. Domingo-Ferrer, F. Sebé, A. Solanas, A polynomial-time approximation to optimal multivariate microaggregation, Computers & Mathematics with Applications 55 (4) (2008) 714–732.
[12] J. Domingo-Ferrer, A. Solanas, A measure of nominal variance for hierarchical nominal attributes, Information Sciences 178 (24) (2008) 4644–4655; J. Domingo-Ferrer, A. Solanas, Erratum in Information Sciences 179 (20) (2009) 3732.
[13] J. Domingo-Ferrer, V. Torra, Ordinal, continuous and heterogenerous *k*-anonymity through microaggregation, Data Mining and Knowledge Discovery 11 (2) (2005) 195–212.
[14] A. Hundepool, A. Van de Wetering, R. Ramaswamy, L. Franconi, A. Capobianchi, P.-P. DeWolf, J. Domingo-Ferrer, V. Torra, R. Brand, S. Giessing, $\mu$-ARGUS version 4.0 Software and User's Manual, Statistics Netherlands, Voorburg NL, May, 2005. <http://neon.vb.cbs.nl/casc>.
[15] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, R. Lenz, J. Longhurst, E. Schulte-Nordholt, G. Seri, P.-P. DeWolf, Handbook on Statistical Disclosure Control (version 1.0), Eurostat (CENEX SDC Project Deliverable), 2006. <http://neon.vb.cbs.nl/CENEX/>.
[16] M. Laszlo, S. Mukherjee, Minimum spanning tree partitioning algorithm for microaggregation, IEEE Transactions on Knowledge and Data Engineering 17 (7) (2005) 902–911.
[17] A. Martínez-Ballesté, A. Solanas, J. Domingo-Ferrer, J.M. Mateo-Sanz, A genetic approach to multivariate microaggregation for database privacy, in: 23rd IEEE International Conference on Data Engineering Workshops – ICDE 2007, 2007, pp. 180–185.
[18] K. Muralidhar, R. Sarathy, Generating sufficiency-based non-synthetic perturbed data, Transactions on Data Privacy 1 (1) (2008) 17–33. <http://www.tdp.cat/issues/tdp.a005a08.pdf>.
[19] A. Oganian, A.F. Karr, Combinations of SDC methods for microdata protection, in: J. Domingo-Ferrer, L. Franconi (Eds.), Privacy in Statistical Databases – PSD 2006, Lecture Notes in Computer Science, vol. 4302, Springer, Berlin Heidelberg, 2006, pp. 102–113.
[20] A. Parakh, S. Kak, Online data storage using implicit security, Information Sciences 179 (19) (2009) 3323–3331.
[21] J.P. Reiter, Releasing multiply-imputed, synthetic public-use microdata: an illustration and empirical study, Journal of the Royal Statistical Society A 168 (1) (2005) 185–205.
[22] D.B. Rubin, Discussion of statistical disclosure limitation, Journal of Official Statistics 9 (2) (1993) 461–468.
[23] P. Samarati, Protecting respondents' identities in microdata release, IEEE Transactions on Knowledge and Data Engineering 13 (6) (2001) 1010–1027.
[24] M. Templ, Statistical disclosure control for microdata using the R-package sdcMicro, Transactions on Data Privacy 1 (2) (2008) 67–85.
[25] A.P. Topchy, M.H.C. Law, A.K. Jain, A.L. Fred, Analysis of consensus partition in cluster ensemble, in: Fourth IEEE International Conference on Data Mining – ICDM'04, 2004, pp. 225–232.
[26] V. Torra, J. Domingo-Ferrer, Record linkage methods for multidatabase data mining, in: V. Torra (Ed.), Information Fusion in Data Mining, Springer, Berlin Heidelberg, 2003, pp. 101–132.
[27] V. Torra, Microaggregation for categorical variables: a median based approach, in: J. Domingo-Ferrer, V. Torra (Eds.), Privacy in Statistical Databases-PSD 2004, Lecture Notes in Computer Science, vol. 3050, Springer, Berlin Heidelberg, 2004, pp. 162–174.
[28] L. Willenborg, T. DeWaal, Elements of Statistical Disclosure Control, Springer-Verlag, New York, 2001.
[29] W.E. Winkler, Re-identification methods for masked microdata, in: J. Domingo-Ferrer, V. Torra (Eds.), Privacy in Statistical Databases, Lecture Notes in Computer Science, vol. 3050, Springer, Berlin Heidelberg, 2004, pp. 216–230.
[30] W.E. Yancey, W.E. Winkler, R.H. Creecy, Disclosure risk assessment in perturbative microdata protection, in: J. Domingo-Ferrer (Ed.), Inference Control in Statistical Databases, Lecture Notes in Computer Science, vol. 2316, Springer, Berlin Heidelberg, 2002, pp. 135–152.