Hybrid Microdata via Model-Based Clustering

Anna Oganian[1] and Josep Domingo-Ferrer[2]

[1] Georgia Southern University
Department of Mathematical Sciences
P.O. Box 8093, Statesboro, GA 30460-8093, U.S.A.
aoganyan@georgiasouthern.edu
[2] Universitat Rovira i Virgili
UNESCO Chair in Data Privacy
Department of Computer Engineering and Maths
Av. Països Catalans 26, E-43007 Tarragona, Catalonia
josep.domingo@urv.cat

Abstract. In this paper we propose a new scheme for statistical disclosure limitation which can be classified as a hybrid method of protection, that is, a method that combines properties of perturbative and synthetic methods. This approach is based on model-based clustering with the subsequent synthesis of the records within each cluster. The novelty is that the clustering and synthesis methods have been carefully chosen to fit each other in view of reducing information loss. The model-based clustering tries to obtain clusters such that the within-cluster data distribution is approximately normal; then we can use a multivariate normal synthesizer for the local synthesis of data. In this way, some of the non-normal characteristics of the data are captured by the clustering, so that a simple synthesizer for normal data can be used within each cluster. Our method is shown to be effective when compared to other disclosure limitation strategies.

Keywords and Phrases: Statistical disclosure limitation (SDL), hybrid SDL methods, mixture models, model-based clustering, expectation-maximization (EM) algorithm.

1 Introduction

To protect microdata which contain sensitive variables, agencies apply disclosure limitation methods to the data prior to their release. These methods can be divided in two groups: masking methods, that release a modified version of the original microdata, and synthetic methods, that release synthetic records as random samples drawn from the distribution representing original data.

Examples of masking methods include: data swapping, in which data values are swapped for selected records; noise addition, in which noise is added to numerical data values to reduce the likelihood of exact matching on key variables or to distort the values of sensitive variables; and microaggregation, which is briefly reviewed next. See [14] for more details.

J. Domingo-Ferrer and I. Tinnirello (Eds.): PSD 2012, LNCS 7556, pp. 103–115, 2012.

Microaggregation can be viewed as cardinality-constrained clustering which can be applied to numerical and categorical variables [33,7]. It consists of a partition step and an aggregation step:

- In the partition step, the set of the original records is partitioned into a number of clusters each containing at least k records for some preset integer k and with the aim that the records within each cluster be as homogeneous as possible. For example, for continuous variables, the sum of squares criterion is a common measure of homogeneity in clustering [34,9,13]. An important feature of microaggregation is that the number of records per cluster should be at least k, which is a parameter of the method.
- In the aggregation step, an aggregation operator (for example, the mean for continuous data or the median for categorical ordinal data) is computed and used to replace the original records. So, the released masked data set consists of the cluster means/medians and the parameter k is responsible for the utility/risk trade-off.

Regarding synthetic methods, the crux is to obtain a good data generation model. Often synthetic data are generated using sequential modeling strategies, similar to those for imputation of missing data in [23]. Re-identification disclosure risk is considered to be very low for synthetic data, because original records are not released. However, if the model is overfitted, synthetic records may resemble very much the original ones. Another shortcoming of synthetic data may be their lack of flexibility: if the data use does not coincide with the model used to synthesize the data, results obtained on synthetic data are likely to differ substantially from the results that would be obtained on the original data.

In [4,18,6] *hybrid* data generation methods were proposed that combine features of masking and synthetic methods and neutralize their pitfalls.

1.1 Contribution and Plan of This Paper

We present in this paper a hybrid data generation method. Similar to [6], we combine clustering of the original data and local data synthesis within each cluster. However, we claim that, to obtain best results, the clustering procedure and the synthesizer used should fit each other to reduce information loss. We use a model-based clustering method that tries to obtain clusters such that the within-cluster distribution is approximately multivariate normal. This allows us to use then a simple multivariate normal synthesizer to do the local within-cluster data synthesis. We show that our method is able to outperform fully synthetic data obtained via multiple imputation with sequential regressions and other SDL methods as well. We present an empirical comparison for numerical data based on propensity scores.

The idea of our method is described in Section 2. Some theoretical properties of our scheme are shown in Section 3. The results of a numerical experiment are reported in Section 4. Finally, Section 5 provides a concluding discussion and sketches lines for future work.

2 Hybrid Data Based on Model-Based Clustering

Several methods of hybrid data generation have been proposed in the literature. One example is [18], where a regression-like scheme is used to generate hybrid values as

$$\boldsymbol{y}_i = \boldsymbol{\gamma} + \boldsymbol{x}_i\boldsymbol{\alpha}^T + \boldsymbol{s}_i\boldsymbol{\beta}^T + \boldsymbol{e}_i, i = 1, \cdots, n \tag{1}$$

where $\boldsymbol{X}$ are the confidential variables and $\boldsymbol{S}$ are the non-confidential ones, and the coefficients $\boldsymbol{\gamma}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are computed in such a way that the first two moments of the hybrid data are the same as in the original data (see the details in the paper). According to the authors, continuous variables $\boldsymbol{S}$ are released unperturbed. However, in such a case the re-identification disclosure risk can be very high, because most of the values of continuous variables are unique. In addition to that, when the regression standard error is small, for example when $\boldsymbol{X}$ and $\boldsymbol{S}$ are strongly correlated, then attribute disclosure can be very high, because the synthesized values would be very close to the original ones.

Another hybrid scheme was proposed in [6]. This method uses microaggregation followed by a synthetic data generator. In the first step, microaggregation partitions the records in the original data file into clusters containing between k and $2k - 1$ records (where k is specified by the data protector) and, in the second step, a synthetic data generator generates records independently for each cluster. According to the authors, it can be any synthesizer which preserves some set of parameters or statistics of the original data set. However, if we want to maximize data utility, it is better to choose a clustering procedure and a synthesizer that "fit" each other in the best possible way. In other words, ideally we would like our clustering/microaggregation algorithm to find such clusters that the distribution of the records in these clusters would be of the same type as the distribution modeled by the synthesizer. For example, if our clustering algorithm is searching for a partition into approximately multivariate normal clusters, then the sensible synthesizer on the next step would be a multivariate normal generator (with the same means and covariance matrices as the original clusters).

There is another important issue about such a procedure: the model and the synthesizer for each cluster should not be overly complex, otherwise the whole procedure may become computationally very intensive (if not infeasible) for the multidimensional data, because we have to run it for each cluster. On the other side, each cluster's model does not need to be the best in order to achieve good utility properties for the overall data, because the synthesizer is not used as a single tool for data generation; in fact, the clustering done on the first step may have captured those characteristics that are hard to model. For example, suppose we have a numerical data set where some variables have skewed distributions and we use as a synthesizer a multivariate normal generator that preserves the mean vector and the covariance matrix. Obviously, this synthesizer alone would not do a good job as all the non-normal characteristics of the data would be lost. However, together with the proper clustering, the resulting hybrid data would be able to reproduce skewness and other distributional characteristics (besides

overall mean and covariance matrix) provided that the number of clusters is sufficient.

As such "proper" clustering, we will use model-based clustering. Its goal is to identify a set of k subpopulations in the data and provide a model (density distribution) of each subpopulation. This is a problem of mixture model estimation. The expectation-maximization (EM) [17,5] algorithm is an effective and popular technique for estimating mixture model parameters or fitting the model to the data.

A possible approach for continuous variables is a Gaussian mixture model. There are two reasons for such a choice: (1) density estimation theory guarantees that any distribution can be effectively approximated by a mixture of Gaussians [29,30]; (2) the multivariate normal synthesizer is simple. So the density of the entire data set can be presented as

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x|\boldsymbol{\theta_k}) \tag{2}$$

where π_k is the probability that an observation belongs to the k-th cluster ($\pi_k \geq 0$; $\sum_{k=1}^{K} \pi_k = 1$), and f_k is the multivariate normal density of the k-th cluster, where the distribution parameters $\boldsymbol{\theta_k}$ are represented by the within-cluster mean vector μ_k and the covariance matrix Σ_k.

Data generated by multivariate normal densities can be represented by groups of ellipsoid clusters centered at mean vectors μ_k. The geometric characteristics of the cluster are determined by the covariance matrices Σ_k.

Note, however, that if there are no restrictions on the covariance matrix structure, then the number of model parameters to be estimated is high, specifically $K(d+d(d+1)/2+1)-1$, where d is the dimensionality of the data. Constraining covariance matrices results in less parameters to be estimated. For example, the following constraints can be introduced: $\Sigma_k = \lambda I$, where all clusters are spherical and of the same size; or $\Sigma_k = \Sigma$, where all clusters have the same covariance and size, but do not need to be spherical. It is of course possible to use an unrestricted covariance matrix Σ_k, where each cluster may have a different geometry [11,3].

The EM algorithm can be used to find maximum likelihood estimates of μ_k, Σ_k and π_k.

In EM for mixture models, the "complete data" are considered to be $\boldsymbol{y}_i = (\boldsymbol{x}_i, \boldsymbol{z_i})$, where $\boldsymbol{z_i} = (z_{i1}, \cdots, z_{iK})$ is the unobserved portion of the data, with $z_{ik} = 1$ if $\boldsymbol{x}_i$ belongs to the cluster k and $z_{ik} = 0$ otherwise.

The E step of the EM algorithm for mixture is given by

$$\hat{z}_{ik} \leftarrow \frac{\hat{\pi}_k f_k(\boldsymbol{x}_i|\hat{\theta}_k)}{\sum_{j=1}^{K} \hat{\pi}_k f_j(\boldsymbol{x}_i|\hat{\theta}_j)} \tag{3}$$

For the M step, estimates of the means, covariance matrices $\hat{\Sigma}_k$ and probabilities have closed-form expressions which can be found in the literature (see [3]).

To find the best model (that is, the number of clusters and the covariance matrix parameterizations) we will use the Bayesian approach and, in particular, the BIC criterion for model selection:

$$BIC_k = 2 \log p(D|\hat{\theta}_k, M_k) - \nu_k \log(n) \tag{4}$$

where D is the data and ν_k is the number of independent parameters to be estimated in model M_k. The literature on model-based clustering suggests that the model choice based on BIC has given good results from the data utility perspective [1,2,10,31]. On the other hand, because the formula for BIC includes a penalty term on the number of estimated parameters, which is a function of the number of components, the model with very many small clusters will not (necessarily) be chosen, which is good from the disclosure risk perspective.

The EM algorithm is sensitive to the initial solution. We will use the approach described in [11], where EM is initialized with the result of model-based hierarchical agglomerative clustering, which approximately maximizes the classification likelihood. As for the number of clusters, we consider from 2 to κ clusters, where κ depends on the number of records in the data set. In our experiments, we set $\kappa = 10$ and computed BIC for each such model. The mixture model which maximizes BIC is chosen. This procedure is implemented in the mclust R package [12].

3 Some Analytical Properties of the Method

The proposed method preserves the mean vector and the covariance matrix within the clusters. Preservation of the mean vector for the overall data set follows from its preservation within each cluster and from the fact that we generate the same number of records for each cluster in the masked data as in the original data.

The covariance matrix for the overall data set is related to the covariance matrix of the clusters as

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \pi_k (\boldsymbol{\Sigma_k} + \mathbf{MDIF_k}) \tag{5}$$

where $\boldsymbol{\Sigma_k}$ is the covariance matrix of the masked data in the cluster k and $\mathbf{MDIF_k}$ is the following matrix:

$$\begin{pmatrix} (\mu_{k_1} - \mu_1)^2 & (\mu_{k_1} - \mu_1)(\mu_{k_2} - \mu_2) & \cdots & (\mu_{k_1} - \mu_1)(\mu_{k_d} - \mu_d) \\ (\mu_{k_2} - \mu_2)(\mu_{k_1} - \mu_1) & (\mu_{k_2} - \mu_2)^2 & \cdots & (\mu_{k_2} - \mu_2)(\mu_{k_d} - \mu_d) \\ & & \ddots & \\ (\mu_{k_d} - \mu_d)(\mu_{k_1} - \mu_1) & & \cdots & (\mu_{k_d} - \mu_d)^2 \end{pmatrix}$$

where μ_i is the mean of variable $\boldsymbol{X}_i$ for the overall data set and μ_{k_i} is the mean of variable $\boldsymbol{X}_i$ over cluster k. Because in the original and masked data the overall means μ_i and the cluster means μ_{k_i} are preserved, matrices $\mathbf{MDIF_k}$

are the same in the original and masked data. Cluster covariances $\boldsymbol{\Sigma_k}$ are also preserved, so the overall covariance $\boldsymbol{\Sigma}$ will be preserved too.

Now let us consider the higher-order moments. Preservation of the higher-order moments depends on the distribution within the clusters of the original data. So, let us consider a generic central moment $E[(\boldsymbol{X}_1 - \mu_1)^{s_1}(\boldsymbol{X}_2 - \mu_2)^{s_2} \cdots (\boldsymbol{X}_d - \mu_d)^{s_d}]$. Let $\boldsymbol{X_{m}}_i$ be the variable i in the masked data. Because the cluster and overall means are preserved, we will omit the index m (denoting masked data) in the expressions for the means. So for the masked data this moment is:

$$
\begin{aligned}
&E\Big[\prod_{i=1}^{d}(\boldsymbol{X_{m}}_i - \mu_i)^{s_i}\Big] = \sum_{k=1}^{K}\pi_k E\Big[\prod_{i=1}^{d}(\boldsymbol{X_{m}}_i - \mu_{k_i} + \mu_{k_i} - \mu_i)^{s_i}\Big] = \\
&= \sum_{k=1}^{K}\pi_k E\Big[\prod_{i=1}^{d}(\sum_{l_i=0}^{s_i}\binom{s_i}{l_i}(\boldsymbol{X_{m}}_i - \mu_{k_i})^{l_i}(\mu_{k_i} - \mu_i)^{s_i-l_i})\Big] = \\
&= \sum_{k=1}^{K}\pi_k E\Big[\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d=0}^{s_d}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(\boldsymbol{X_{m}}_i - \mu_{k_i})^{l_i}(\mu_{k_i} - \mu_i)^{s_i-l_i}\Big)\Big] = \\
&= \sum_{k=1}^{K}\pi_k \sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d=0}^{s_d}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(\mu_{k_i} - \mu_i)^{s_i-l_i}\Big) NormM_k(l_1, l_2, \cdots, l_d) \quad (6)
\end{aligned}
$$

where $NormM_k(l_1, l_2, \cdots, l_d) = E[\prod_{i=1}^{d}(\boldsymbol{X}_i - \mu_i)^{l_i}]$ is the normal mixed central moment for $\boldsymbol{X} \sim N(\mu_{k_i}, \boldsymbol{\Sigma}_k)$ over cluster k. The expression for $NormM_k(l_1, l_2, \cdots, l_d)$ can be found in the literature; for example in [21].

Note that $NormM_k(l_1, l_2, \cdots, l_d)$ should be computed only for those moments for which $\sum_{i=1}^{d} l_i$ is even, because all other moments are equal to 0.

Taking into account that our method preserves the first two moments, the difference between the corresponding moments computed on the original and masked data is the following:

$$
\begin{aligned}
&E\Big[\prod_{i=1}^{d}(\boldsymbol{X_{m}}_i - \mu_i)^{s_i}\Big] - E\Big[\prod_{i=1}^{d}(\boldsymbol{X_{o}}_i - \mu_i)^{s_i}\Big] = \\
&= \sum_{k=1}^{K}\pi_k\Big(\Big(\underbrace{\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d=0}^{s_d}}_{\sum l_i > 2 \text{ and even}}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(\mu_{k_i} - \mu_i)^{s_i-l_i}\Big)\times \\
&\times\big(NormM_k(l_1, l_2, \cdots, l_d) - M_k(l_1, l_2, \cdots, l_d)\big)\Big) - \\
&- \Big(\underbrace{\sum_{l_1=0}^{s_1}\sum_{l_2=0}^{s_2}\cdots\sum_{l_d}^{s_d}}_{\sum l_i > 1 \text{ and odd}}\prod_{i=1}^{d}\Big(\binom{s_i}{l_i}(\mu_{k_i} - \mu_i)^{s_i-l_i}\Big) M_k(l_1, l_2, \cdots, l_d)\Big)\Big) \qquad (7)
\end{aligned}
$$

where subscripts "o" and "m" denote the original and masked data correspondingly, and $M_k(l_1, l_2, \cdots, l_d)$ is the moment computed in the k^{th} cluster of the original data.

The difference between the original and masked moments depends on the non-normal properties of the clusters of the original data. Obviously, if all the clusters in the original data are normally distributed then all the moments will be preserved by our masking scheme. We want to note that, since the clustering step is done by the EM algorithm, the distribution of the clusters should tend to normality because the probability of assigning record i to cluster k is estimated using Expression (3), which has the normal density function in the nominator. That is, roughly speaking, the records which by the criterion of normal distribution are far away from the cluster mean have very little chance of being assigned to this cluster. Hence, for approximately normal clusters, the first term on the right hand side of Expression (7), which reflects the difference between "even" moments, should not be very big, and the second term should be close to zero. We have to note though, that clusters found by EM are not always normally distributed, of course. In our experiments, some clusters were non-normal and outliers especially in the "outside clusters" contributed to non-normality of the clusters, however, overall utility of the resultant data (according to the metric we used) was still considerably high, as it is described in the section 4. In the future work we will investigate possible remedies to this problem, including data transformation.

4 Numerical Experiments with Continuous Data

The procedure described above was implemented and evaluated on two real data sets:

- The first data set, called DIABETES ([25] and also available from the R package mclust [12]), contains the following three continuous variables: glucose intolerance, insulin response to oral glucose and insulin resistance (quantified by determining the steady-state plasma glucose (SSPG) concentration in response to an infusion of octreotide, glucose, and insulin). There are 145 individuals in this data set.
- The second data set was extracted from the 1995 U.S. Current Population Survey; specifically, it contains some variables of the CENSUS test data set [22]. There are 1080 records in this data set, containing the following three numerical variables: adjusted gross income (agi), employer contribution for health insurance (emcontrb) and federal income tax liability (fedtax).

We applied the procedure described above to these data sets. For DIABETES, a model with three clusters (33, 33 and 82 records) and unconstrained covariance matrices was chosen by BIC. We want to note that respondents in this data set can actually be classified as patients with chemical diabetes, patients with overt diabetes and normal subjects. So our method was able to correctly identify these meaningful subpopulations and reproduce the first two moments in these subgroups.

The CENSUS data set did not have any obvious clustered structure, and the application of our method led to the partition of the records into 9 clusters with different covariance matrices. The sizes of clusters ranged from 62 to 240 records (again this was the model with the best BIC value).

We call our method above as Hybrid. For the sake of comparison we also generated hybrid data sets where the clustering step was done by MDAV multivariate microaggregation [8], and the data synthesis within each cluster was done using a synthesizer that preserves means and covariance matrices, as described in [6]. Denote this method as Hybridmicro. For the CENSUS data set, the microaggregation parameter k was set to 120 records per cluster, and for the DIABETES data set we took $k = 48$ (such cluster sizes are the average cluster sizes in the Hybrid method for the corresponding data set). This yielded 9 clusters for CENSUS and 3 clusters for DIABETES. The obtained hybrid data sets were compared with fully synthetic data. For synthetic data generation we used a method based on multivariate sequential regressions as described in [24,26,16]. A free multiple imputation software IVEware [15] was used to generate them.

Other methods used for comparison were plain multivariate microaggregation MDAV [8], denoted as Micro, and noise addition, which are perturbation methods. Multivariate microaggregation was done with $k = 10$ records per cluster for CENSUS and $k = 5$ for DIABETES. The choice of k was made empirically to reach a reasonably fair comparison with the other methods. Since Hybrid and HybridMicro restore the variance within the clusters and Micro does not, it would be unfair to compare them with Micro with $k = 120$ records per cluster for CENSUS data, because such microaggregated data would have only 9 distinct records. With $k = 10$ there are 108 different records, which is a much better case. Similar considerations apply to justify the $k = 5$ used in DIABETES.

We used the implementation of MDAV microaggregation available in the R package sdcMicro [32] for our method Micro and the first step of HybridMicro. Regarding noise addition, we used a version that preserves the mean vector and the covariance matrix. This method was implemented in the following way:

$$\boldsymbol{X_m} = E[\boldsymbol{X_o}] + \frac{(\boldsymbol{X_o} - E[\boldsymbol{X_o}]) + \boldsymbol{E}}{\sqrt{1+c}}, \tag{8}$$

where $\boldsymbol{X_m}$ is the masked data, $\boldsymbol{X_o}$ is the original data, $E[\boldsymbol{X_o}]$ denotes the expectation of $\boldsymbol{X_o}$, $\boldsymbol{E}$ is the random noise with $N(\boldsymbol{0}, c\boldsymbol{\Sigma_o})$, $\boldsymbol{\Sigma_o}$ is the covariance matrix of the original data, and c is the parameter of the method which regulates the amount of the noise added to the data. We used $c = 0.15$, as recommended in the literature [19,20,35]. We call this method Noise.

To evaluate the data quality provided by these methods, we chose a measure of data utility which can be suitable for a number of analyses: the propensity score-based measure [35]. This measure is based on discrimination between the

original and masked data: masked data that are difficult to distinguish from the original data have relatively high utility.

Propensity-based information loss is computed in two steps. First, the original and masked data sets are merged and an indicator variable T equal to one for masked records, and to zero for original records, is added. Second, for each record in the original and masked data, the propensity score —the probability of being in the masked data set— is computed. It was shown in [35] that, if the propensity scores of all records are close to 0.5, then the original and masked data have the same distributions. The utility measure is computed as

$$\mathbf{Propen} = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{p}_i - 0.5\right]^2, \tag{9}$$

where N is the total number of records in the merged data set and $\hat{p}_i$ is the estimated propensity score for record i.

The propensity score utility measure depends on the specification of the model used to estimate propensity scores (see [35]). The model that we used contains all main effects, first-order interactions among all the variables and also quadratic effects.

Table 1. Propensity score utility for various methods (lower values mean better utility)

	Method				
Data set	Hybrid	HybridMicro	Synthetic	Micro	Noise
Diabetes	4.26	10.27	19.2	19.79	7.02
Census	4.080	6.515	48.83	200.040	11.977

The results for different methods are shown in Table 1. These are average values of data utility for 30 realizations of masked data sets obtained from the same original data set by the application of Hybrid, HybridMicro, Synthetic and Noise; for Micro a single realization was enough, because it is a deterministic method. We see that both hybrid methods, Hybrid and HybridMicro, outperform the fully synthetic method Synthetic. Further, Hybrid is the best method in terms of utility. We noticed that Hybrid performs better than HybridMicro even if we increase the number of clusters for HybridMicro. For example, in the case of the CENSUS data set, when we changed the aggregation level from 120 records per cluster to 60 for HybridMicro, thus increasing the number of clusters from 9 to 18, the average utility for HybridMicro was about 5.05, which is still worse than the utility of Hybrid with only 9 clusters. A similar behavior was observed for the DIABETES data set. Remember that reducing the number of clusters without losing utility is highly desirable, because higher levels of aggregation can be expected to result in lower disclosure risk.

5 Concluding Discussion and Future Work

Model-based clustering followed by generation of synthetic records using parameters estimated in the clustering step seems to be a quite promising and flexible approach for disclosure limitation of continuous data.

In our experiments, this approach outperformed all other disclosure limitation methods which were considered for comparison. In particular, for both of our data sets, our method was also considerably better than the fully synthetic data generator based on sequential regressions. This suggests that global synthesis of data sets with complex structure may not give good results in what regards the utility of the resulting synthetic data. In contrast, *local synthesis*, which is the essence of our method, may be the best option, even when using a local synthesizer as simple as the multivariate normal one. Indeed, a proper combination of clustering and synthesis may capture those properties of the data which are hard to model on the global data set.

Our method is flexible also in the sense that, by increasing or decreasing the number of clusters, we can obtain data that resembles more or less to the original data. This is reflected by the utility scores: for a fixed method, lower values of the propensity score measure (better utility) were obtained for the models with greater number of clusters. We want to note, however, that using a very large number of clusters may be dangerous from the point of view of disclosure risk, because clusters become smaller and the synthesized records become quite similar to the original records.

Another reason why we should not try to maximize the number of clusters is because, for the data sets that have an underlying mixture distribution, like DIABETES, the data user may be interested in estimating the properties of the meaningful underlying subgroups. The Hybrid method allows the user to do that when that the best model is chosen by BIC. However, the HybridMicro method based on microaggregation followed by synthesizer may fail to allow this: HybridMicro will create approximately spherical clusters with $2k$ to $2k - 1$ records each. Note that natural subpopulations are not necessarily of size between k and $2k - 1$ records. Hence, microaggregation may produce too many clusters (depending on the choice of k) and even when the first two moments are preserved in these clusters, this may be of little use for the user, as the clusters do not represent the meaningful subgroups. In addition, maximizing the number of clusters most likely will result in adding unnecessary noise to the data.

For those data sets which do not have a clear clustered structure, it seems that we can have more clusters. However, some parsimony criterion like BIC is still necessary to prevent the formation of too many clusters, which could lead to the synthesized records being too close to the original ones.

The results presented in this paper are just the first steps in the development of a local synthesis approach which, in our opinion, is flexible and powerful. More investigation and experimental work is necessary. In particular, in the future experiments, we will use other utility measures as well in addition to propensity scores and we will include more synthetic/ hybrid methods in the comparative experiments, in particular those, that are based on nonparametric

machine-based approach [28], [27]. Regarding to future methodology research, we intend to address the following questions:

- Investigate and compare different mixture models with not necessarily normal components for disclosure limitation.
- Quantify disclosure risk for hybrid methods. Re-identification disclosure is not very relevant for our method, because all records are synthesized and there is no one-to-one correspondence between the original and hybrid records. There is, however, a correspondence between the groups. What seems to be more relevant is attribute disclosure, which should be investigated and quantified.
- Extend the method to the cases when data sets have continuous and also categorical variables. Latent Class Analysis (LCA) seems to be appropriate for modeling categorical variables. A proper correspondence should be established between the model-based clustering techniques for continuous and categorical variables in order to preserve the relationships between these types of variables.

Acknowledgments and Disclaimer. The second author is with the UNESCO Chair in Data Privacy, but the views expressed in this paper do not necessarily reflect the position of UNESCO nor commit that organization. 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", TIN2011-27076-C03-01 "CO-PRIVACY" and CONSOLIDER INGENIO 2010 CSD2007-00004 "ARES", and by the European Comission under FP7 project "DwB". The second author is partially supported as an ICREA Acadèmia researcher.

References

1. Campbell, J.G., Fraley, C., Murtagh, F., Raftery, A.E.: Linear flaw detection in woven textiles using model-based clustering. Pattern Recognition Letters 18, 1539–1548 (1997)
2. Campbell, J.G., Fraley, C., Stanford, D., Raftery, A.E.: Model-based methods for real-time textile fault detection. International Journal of Imaging Systems and Technology 10, 339–346 (1999)
3. Celeux, G., Govaert, G.: Gaussian parsimonious clustering models. Pattern Recognition 28, 781–793 (1995)
4. Dandekar, R.A., Cohen, M., Kirkendall, N.: Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique. In: Domingo-Ferrer, J. (ed.) Inference Control in Statistical Databases. LNCS, vol. 2316, pp. 117–253. Springer, Heidelberg (2002)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood for incomplete data via the em algorithm (with discussion). Journal of the Royal Statistical Society, Ser.B 39, 1–38 (1977)
6. Domingo-Ferrer, J., González-Nicolás, U.: Hybrid microdata using microaggregation. Information Sciences 180, 2834–2844 (2010)
7. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogeneous k-anonimity through microaggregation. Data Mining and Knowledge Discovery 11, 195–212 (2005)

8. Domingo-Ferrer, J., Torra, V.: Ordinal, continuous and heterogenerous k-anonymity through microaggregation. Data Mining and Knowledge Discovery 11(2), 195–212 (2005)
9. Edwards, A.W.F., Cavalli-Sforza, L.L.: A method for cluster analysis. Biometrics 21, 362–375 (1965)
10. Fraley, C., Raftery, A.E.: How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal 41, 578–588 (1998)
11. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97(458), 611–631 (2002)
12. Fraley, C., Raftery, A.E.: MCLUST Version 3 for R: Normal mixture modeling and model-based clustering. Technical Report no. 504, Department of Statistics, University of Washington (September 2006), http://cran.r-project.org/web/packages/mclust/index.html
13. Hansen, P., Jaumard, B., Mladenovic, N.: Minimum sum of squares clustering in a low dimensional space. Journal of Classification 15, 37–55 (1998)
14. Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte-Nordholt, E., Seri, G., DeWolf, P.-P.: Handbook on Statistical Disclosure Control (version 1.2). ESSNET SDC project (2010), http://neon.vb.cbs.nl/casc
15. IVEware. Imputation and Variance Estimation software, http://www.isr.umich.edu/src/smp/ive/ (accessed July 12, 2012)
16. Little, R.J., Liu, F., Raghunathan, T.: Statistical disclosure techniques based on multiple imputation. In: Gelman, A., Meng, X.-L. (eds.) Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, vol. 18, pp. 141–152. Wiley, New York (2004)
17. McLachlan, G.J., Krishnan, T.: EM Algorithm and Extensions. Wiley, New York (1997)
18. Muralidhar, K., Sarathy, R.: Generating sufficiency-based non-synthetic perturbed data. Transactions on Data Privacy 1(1), 17–33 (2008), http://www.tdp.cat/issues/tdp.a005a08.pdf
19. Oganian, A.: Security and Information Loss in Statistical Database Protection. PhD thesis, Universitat Politecnica de Catalunya (2003)
20. Oganian, A., Karr, A.F.: Combinations of SDC Methods for Microdata Protection. In: Domingo-Ferrer, J., Franconi, L. (eds.) PSD 2006. LNCS, vol. 4302, pp. 102–113. Springer, Heidelberg (2006)
21. Phillips, K.: R functions to symbolically compute the central moments of the multivariate normal distribution. Journal of Statistical Software, Code Snippets 33(1), 1–14 (2010)
22. Mateo-Sanz, J.M., Brand, R., Domingo-Ferrer, J.: Reference data sets to test and compare SDC methods for protection of numerical microdata (2002), http://neon.vb.cbs.nl/casc/CASCtestsets.htm
23. Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., Solenberger, P.: A multivariate technique for multiply imputing missing values using a series of regression models. Survey Methodology 27, 85–96 (2001)
24. Raghunathan, T.E., Reiter, J.P., Rubin, D.B.: Multivariate imputation for statistical disclosure limitation. Journal of Official Statistics 19(1), 1–16 (2003)
25. Reaven, G.M., Miller, R.G.: An attempt to define the nature of chemical diabetes using multidimensional analysis. Diabetologica 16(1), 17–24 (1979)
26. Reiter, J.P.: Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics 18, 531–544 (2002)

27. Reiter, J.P.: Using CART to generate partially synthetic public use microdata. Journal of Official Statistics 21, 441–462 (2005)
28. Caiola, G., Reiter Random, J.P.: forests for generating partially synthetic, categorical data. Transactions on Data Privacy 3(1), 27–42 (2010)
29. Scott, D.W.: Multivariate Density Estimation. John Wiley & Sons, New York (1992)
30. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman & Hall (1986)
31. Stanford, D., Raftery, A.E.: Principle curve clustering with noise. IEEE Transactions on Pattern Analysis and Machine Inteligence, 601–609 (2000)
32. Templ, M.: Statistical disclosure control for microdata using the R-package sdcMicro. Transactions on Data Privacy 1(2), 67–85 (2008)
33. Torra, V.: Microaggregation for Categorical Variables: A Median Based Approach. In: Domingo-Ferrer, J., Torra, V. (eds.) PSD 2004. LNCS, vol. 3050, pp. 162–174. Springer, Heidelberg (2004)
34. Ward, J.H.: Hierarchical grouping to optimize an objective function. Journal of American Statistical Association 58, 236–244 (1963)
35. Woo, M.-J., Reiter, J.P., Oganian, A., Karr, A.F.: Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1(1), 111–124 (2009)