mathematics

MDPI

Article

Privacy Protection Practice for Data Mining with Multiple Data Sources: An Example with Data Clustering

Pauline O'Shaughnessy *,† and Yan-Xia Lin †

School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW 2522, Australia
* Correspondence: poshaugh@uow.edu.au; Tel.: +61-02-4221-4241
† These authors contributed equally to this work.

Abstract: In the age of data, data mining provides feasible tools with which to handle large datasets consisting of data from multiple sources. However, there is limited research on retrieving statistical information from data when data are confidential and cannot be shared directly. In this paper, we address this problem and propose a framework for performing data analysis using data from multiple sources without revealing true values for privacy purposes. The proposed framework includes three steps. First, data custodians individually mask data before publishing; then, the masked data collection is used to reconstruct the density function of the original dataset, from which resampled values are generated; last, existing data mining techniques are applied directly to the resampled data. This framework utilises the technique of reconstructing an original density function from noise-masked data using the moment-based density estimation method, which plays an essential role. Simulation studies show that the proposed framework performs well; analysis results from the resampled data are comparable to those of the original data when the density of the original data is estimated well. The proposed framework is demonstrated in data clustering analysis using the example of a real-life Australian soybean dataset. Results from the k-means algorithms with two and three fitted clusters are presented to show that cluster analysis using resampled data can well replicate that of the original data.

Keywords: data masking; multiplicative noise; data mining; sample size calculation

MSC: 68P27; 92B15

Citation: O'Shaughnessy, P.Y.; Lin, Y.-X. Privacy Protection Practice for Data Mining with Multiple Data Sources: An Example with Data Clustering. *Mathematics* **2022**, *10*, 4744. https://doi.org/10.3390/math10244744

Academic Editors: Niansheng Tang and Shen-Ming Lee

Received: 28 October 2022
Accepted: 9 December 2022
Published: 14 December 2022

1. Introduction

With the explosive evolution of information technology and computer science, it is easier and less expensive to collect and store data, and the databases containing this information are often massive. While technological evolution makes access to voluminous data feasible, it also brings many challenges in how to turn big data into big knowledge. Data mining is a key component in big data analytics. It is an inductive process for extracting hidden and potentially useful patterns and information without a priori hypotheses, where traditional hypothesis-driven methods, such as online analytic processing and most statistical methods, fall short [1]. This feature makes data mining techniques ideal when hypotheses are difficult to determine or define.

Given its nature, big data can consist of data from multiple sources, and require a sophisticated information systems for storage and access, often being stored off-site or in systems managed by a third party (e.g., cloud storage). When the control of data access is no longer in the hands of the data owners, there are potential threats to data security. In practice, data access control protocols are implemented to secure data privacy [2]. One of the more extreme ways is to indiscriminately restrict public access to data. This method is often chosen by the data owners for data containing sensitive commercial values. Access restriction provides reasonable data security in this case, as it solely relies on the safekeeping

of datasets [3]. However, data access restriction is usually not an optimal solution, as it is restrictive for general data use and data sharing.

The main interest in data privacy research is to develop methods for protecting data privacy that also allow the preservation of statistical information. This topic has been studied separately in the fields of statistics (statistical disclosure) and computer science (privacy-preserving data mining and privacy-enhancing technologies) [4]. Torra and Navarro-Abrribas [5] provided an overview of the existing data privacy methods, categorised by the types of data and the types of analyses applied to these data. They summarised that when data are published for a general purpose, masking (statistics) and anonymisation (computer science) are the two available methods which can be used to protect the privacy of the data values. Masking and anonymisation methods systematically transform datasets prior to release. They can be classified into three categories: perturbative, non-perturbative and synthetic data generators. Perturbative methods alter data values by introducing predetermined errors, including noise addition or multiplication, substitution, rank swapping, etc. Non-perturbative methods generally refer to data generalisation and suppression, which make data less detailed. Synthetic data generators replace original data with values generated from an underlying model, ideally retaining the desired statistical information of the original data.

Data privacy simultaneously requires that data values are well protected from disclosure and that statistical inference is accurate about the population of interest [4]. Figure 1 demonstrates these two processes for a set of data published for a general purpose. (i) Data protecting techniques are used by data providers to protect original data values to ensure a certain level of privacy protection before publicly releasing datasets. (ii) Once the protected data are available to the general public, suitable procedures are then performed to retrieve the statistical information of the unpublished, original data.

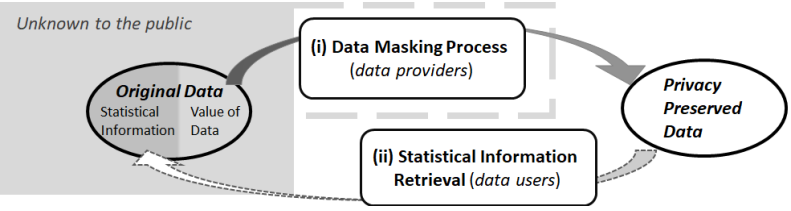

Figure 1. Overview of the data privacy process. The grey area is unavailable to the public for privacy purposes. Information regarding the data masking process performed by data providers can be made partially available to the public.

Current data privacy research focuses on developing methods of ensuring a desirable amount of disclosure with some guarantee on the utility loss for a given statistic (process (i) in Figure 1). An overview of practical privacy protection methods and their applications can be found in [6]. However, there is very limited work on investigating the statistical information retrieval process ((ii) in Figure 1), i.e., how to apply various statistical techniques to a privacy-protected dataset to obtain inferential results other than that considered by the utility loss. Consider the following scenario. Data collection consists of data collected independently from K institutions. Due to the issue of data sensitivity, all institutions require a guarantee of a certain level of privacy protection upon releasing data. Meanwhile, all of them are interested in the statistical inference given by the clustering analysis using the entire data collection. To the best of our knowledge, there is no literature on discussing how to apply existing data mining techniques, particularly clustering analysis, directly to published protected data from multiple data sources when the original data values are not accessible. This topic is the focus of this paper.

Note that there exists a strand of research within the data mining community that addresses privacy issues, namely, privacy-preserving data mining. Privacy-preserving data mining involves modifying existing data mining algorithms to ensure the privacy of the outcomes of algorithms. Reference [7] gives a detailed review of these methods

and discusses developments in this area. The focus of privacy-preserving data mining is on the protection of the outcomes of algorithms, not the data themselves [8], which is a fundamentally different situation from the confidentiality-related privacy issues discussed in this paper. Furthermore, privacy-preserving data mining methods generally require access to original data and need to be customised depending on the analysis. This is different from the problem of the original data being inaccessible, which is discussed in this paper; therefore, we will not consider these privacy-preserving data mining methods.

This paper proposes a framework for data clustering analysis, assuming that the underlying true data are confidential and that it is impossible to directly share data between multiple data sources. In this framework, confidential quantitative data are firstly protected using the noise-multiplicative masking method. Then, the density function of the original data are reconstructed from the noise-protected data using the moment-based density estimation method. Resampled values are then drawn from the reconstructed density and analysed directly for modelling and inferential results.

The paper is organised as follows. In Section 2, we introduce the multiplicative noise masking method for data value protection and the moment-based density function reconstruction for statistical information retrieval. We also introduce an application of the Kolmogorov–Smirnov test for determining the sample size in the context of the reconstructed density function. This is the basic knowledge required for the clustering analysis discussed in this paper. The proposed framework and its performance evaluation through simulations are presented in Section 3. In Section 4, we present the application of the framework to a real-life dataset and evaluate it empirically.

2. Data Publishing and Information Retrieval

Let X be a random variable. Sometimes, we also call it the data population. Assume that there are K institutions. Each of them independently and randomly draws a sample from the population X. Denote $\boldsymbol{x}^{(k)} = \{x_i^{(k)}\}_{i=1}^{N_k}$, as the data collected by the kth institution, where N_k is the size of sample and $k = 1, \cdots, K$. We merge those datasets and form a large sample from the population X. This paper assumes that the K institutions want to carry out clustering analysis based on the large sample. However, all institutions consider their data confidential and do not wish to share them with others without any privacy protection measures in place.

In current data privacy literature, information retrieval is often treated as a part of the data masking strategy. Depending on the parameters of interest and the methods used for data analysis, a specific data masking method is chosen not only for data value protection but also for obtaining reasonable estimates for the parameters. For example, in differential privacy [9], which is a widely-used data privacy mechanism, a zero-mean Laplace noise is used to ensure the unbiased estimation of a group sum, and the infinity divisibility property of the Laplace distribution is utilised to achieve a certain level of privacy when running queries [9–12]. A differential privacy mechanism ensures that no single observation is identifiable from differentiating queries. However, in practice, the level of perturbation needed to ensure a statistically level of privacy protection often is high, which leads to a low statistical utility. Additionally, when the parameter of interest is no longer the sum, the masking techniques or noise distributions must be customised accordingly [13]. Currently, there is no discussion on how to retrieve the accurate statistical information beyond simple statistics in the differential privacy framework, when a dataset is masked and published for general use and the intended analyses are unspecified. Simply applying cluster analysis to masked data protected by a differential privacy mechanism cannot guarantee the results from masked values represent those from the original values.

In this section, we propose a general framework for perturbing data values for privacy, then retrieving relevant statistical information by reconstructing the density function of original data. The basic idea is motivated by Fisher's likelihood principle, which is arguably one of his greatest contributions to the foundation of statistical science. It states that the likelihood function contains all the evidence in a data sample relevant to model

parameters [14]. The likelihood principle implies that statistical information is fully stored in the density function, and data are a representation of the density function. If we can reconstruct the original density function from masked data, we can generate a new dataset from the reconstructed density function, and this new dataset will contain the same statistical information as the original data. Here we discuss the details relevant to the two data privacy processes described in Figure 1, specifically for the general framework for masking and analysing data from multiple data sources.

2.1. Data Masking and Reconstructing the Density Function

2.1.1. Data Masking at Publishing

Data masking protects data by altering values at the individual observation level. Given that data mining techniques are traditionally performed at the individual data level, we only consider the data masking methods that allow for releasing the protected individual data. In particular, we propose to use the multiplicative noise method in this framework, which has desirable properties for masking a wild range of datasets. The multiplicative noise method can be applied to both numerical and categorical data. In addition, the multiplicative noise method provides uniform protection in terms of the coefficient of variation of the noise. This means that the required variation of noise to achieve a desirable level of certainty in estimation does not depend on the values of data, providing an effective way of using small variance for noise distribution to significantly alter large-value data, especially in datasets with large spreads [15].

To protect the values of $\mathbf{x}^{(k)}$ for $k = 1, \cdots, N_K$, firstly, data owners agree on an appropriate random noise C, which is independent of X. Then, each data owner selects a random sample $\{c_i^{(k)}\}$ from the noise population C; a new dataset $\mathbf{x}^{*(k)} = \{x_i^{*(k)}\} = \{x_i^{(k)} c_i^{(k)}\}$ is calculated for the kth institution and can be released to others.

Note that all data owners are required to use the same random noise C to mask their data. Data owners often choose to release certain characteristics of the noise distribution, C, i.e., the shape of the distribution or moment information, etc. (shown as the dashed line around (i) data masking process in Figure 1). When this partial information about the noise distribution is known to the public, the values of $\{x_i^{(k)}\}$ will still be well protected and unable to be recovered from $\{x_i^{*(k)}\}$. For the relevant discussion, see [15].

2.1.2. Reconstructing Density Function

After masked data are publicly available, we use the masked data to reconstruct the density function of the original data in order to accurately obtain the data's statistical information. In practice, there is often no additional information about the underlying distribution beyond actual observations. A robust estimation method with less prior information on reference density is preferred, even though it may be computationally expensive. References [16–18] were the first to independently introduce the fundamental methods for estimating the density function of original data from masked data for a single variable. Lin and Krivitsky [19] gave a detailed review and pointed out that the algorithms proposed in the first three papers have several technical problems, including non-convergence and slow computation. These problems are pronounced in skewed data. Lin [18] exclusively discussed density estimation using a moment-based polynomial approach for noise-multiplied data. This research showed that, for a random variable X with a density function defined on a finite interval $[a, b]$, the density function of X can be approximated by

$$f_{X,P}(x) = \sum_{p=0}^{P} a_p(x) \frac{\mu_{X*}(p)}{\mu_C(p)},$$

with an appropriate integer P, where $X^* = XC$ is the masked random variable, noise C is the independent multiplicative random variable, $\mu_{X^*}(p) = E(X^{*p})$, and $\mu_C(p) = E(C^p)$. $a_p(x)$ is a continuous polynomial function of x.

Lin [18] also pointed out that, given the noise-multiplied data $\{x_i^*\}_{i=1}^N$ and sample moments' information on the multiplicative noise C, $f_{X,P}(x)$ can be empirically approximated by

$$f_{X,P|\{x_i^*,c_i\}}(x) = \sum_{p=0}^{P} a_p(x) \frac{\overline{(X^*)^p}}{\overline{C^p}}, \tag{1}$$

where $\overline{(X^*)^p} = \sum_{i=1}^N (x_i^*)^p/N$ and $\overline{C^p} = \sum_{i=1}^N c_i^p/N$ are the empirical pth moment for masked data X^* and noise distribution C, respectively. This means that we can use the moment information about the masked data and the noise distribution to reconstruct the density function of the original data. Lin [20] subsequently developed a computational algorithm and built an R package called `MaskDensityBM` using the moment-based density estimation method. In this study, we used the method proposed by Lin [18] and utilised the existing software packages for density reconstruction. After reconstructing the density function, we can generate resamples to perform analysis.

2.2. Determining Sample Size for Resampled Data

Since the original data are confidential, we cannot directly use the data for clustering analysis. Our approach uses a sample drawn from the constructed density function (sometimes called the simulated data or resampled data below) to replace the original data to avoid this problem. Based on the approach we propose, the quality of clustering analysis will rely on two factors. One factor is the closeness of the reconstructed density function to the actual density function. We applied the R package `MaskDensityBM` [20] to determine the reconstructed density function. The other factor is the size of the sample drawn from the reconstructed density function, which ensures the statistical information of the reconstructed density can be well retrieved.

We assume that the reconstructed density function captures the main characterises of the density function of the original data. Even under a perfect scenario, the outputs of data analysis given by the original data and those of the simulated data are likely different if the size of the simulated data is small. The main reason is that when the size of the resamples is too small, the information on certain characteristics of the distribution may be missing from the simulated samples, especially in the two tail-end regions. Even if the size of the simulated data is the same as that of the original data, due to randomness in data generating process, there is no guarantee that the set of simulated data has a similar density to the original data.

We suggest an analytic solution to determine an appropriate data size through a sequence of Kolmogorov–Smirnov tests. Denote $\{x_i\}_{i=1}^N$ as the set of the underlying original data with a sample size of N; $\hat{f}_{X,N}$ is the estimated smoothed density function determined by the original data. Let $\{\tilde{x}_i\}_{i=1}^M$ be a set of resampled data with a size of M. Verifying if the smoothed density function given by $\{\tilde{x}_i\}_{i=1}^M$ is statistically equivalent to $\hat{f}_{X,N}$ is the same as checking whether the empirical cumulative probability distribution function given by $\{\tilde{x}_i\}_{i=1}^M$ is close to the cumulative probability distribution determined by $\hat{f}_{X,N}$. The hypotheses are defined as:

$$H_0 : \hat{F}_{\tilde{X}} = \hat{F}_{X,N}$$

and

$$H_1 : \hat{F}_{\tilde{X}} \neq \hat{F}_{X,N}\ ,$$

where $\hat{F}_{\tilde{X}}$ is the empirical cumulative function given by the simulated data $\{\tilde{x}_i\}_{i=1}^M$ and $\hat{F}_{X,N}$ is the cumulative distribution related to $\hat{f}_{X,N}$. The test statistic is

$$D_M = \max_{1 \le i \le M} \left\{ \left| \hat{F}_{X,N}(\tilde{x}_{(i)}) - \frac{i-1}{M} \right| \right\}, \tag{2}$$

where $\tilde{x}_{(i)}$ represents the ith ordered values in the dataset $\{\tilde{x}_i\}_{i=1}^{M}$. A small D_M suggests similarity between the smoothed density function of the original data and the empirical density of the resamples. We considered 0.007 as a critical value for this test and solved for M.

Example 1. *We generated 1000 data from a random variable X following a mixture of normal distributions with density function*

$$f_X = 0.25 \times N(0, 1^2) + 0.75 \times N(4, 2^2)\ . \tag{3}$$

Using Criterion (2), the resampled data with a size of 37,000 has a sufficiently small D_M (=0.0056), and the smoothed density function of the resampled data is shown to be a reasonable estimation of the density of the original data (Figure 2).

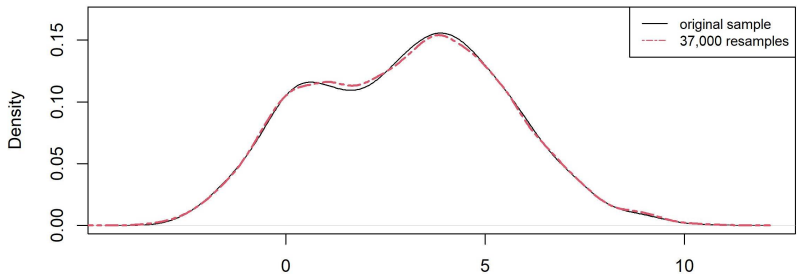

Figure 2. The plots of the smoothed density functions given by $\{x_i\}$ (solid line) for $i = 1, \ldots, 1000$ and a set of resampled values $\{\tilde{x}_i\}$ determined by $D_M < 0.007$ (dash line).

Figure 3 illustrates the relationship between sample size and the information lost, measured by D_M. We generated 500 samples of 1000 data from the model in (3); then we calculated D_M for the resampled data with various sizes of 300, 700, 1000, 3000, 6000, and 12,000. The larger size of resampled data preserves the information of the cumulative distribution of the density function in (3) better with a smaller mean of D_M. It also shows that the variations of the test statistics are much larger in the resampled data with smaller sizes.

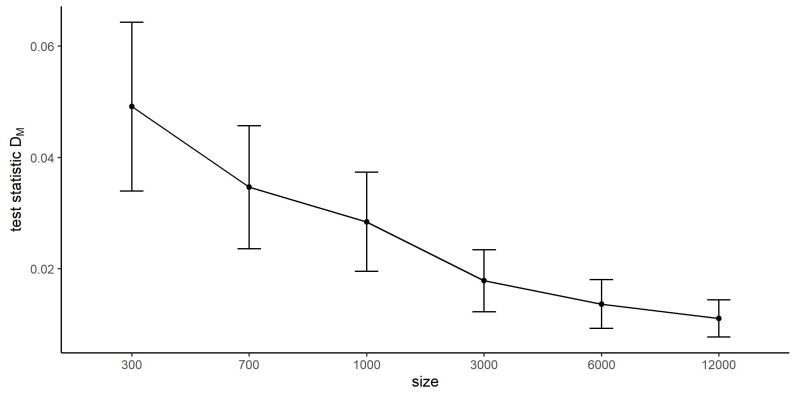

Figure 3. Relationship between D_M and the size of the resampled data for M = 300, 700, 1000, 3000, 6000, and 12,000. The bars extend to one standard deviation above and below the average D_M values from 500 simulation samples.

3. Proposing a Framework and Simulation Studies

In this section, we propose a general framework for publishing data for general use and retrieving statistical information by generating resamples using reconstructed density functions. We consider all data from the K institutions as a whole. Let $\sum_{k=1}^{K} N_k = N$ be the total number of observations from the K institutions and $\mathbf{x} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K)}\} = \{x_j\}$ for $j = 1, \ldots, N$ be a collection of N original data from the K institutions.

Framework for publishing and mining data from multiple data sources:

(i) Publishing data masking:
   (a) Data owners across the K institutions agree on an appropriate noise distribution C. Then, information about the noise distribution and parameter values are released to the public;
   (b) For the kth institution, independently generate a sample $\{c_i^{(k)}\}_{i=1}^{N_k}$ from C, and produce a masked dataset $\mathbf{x}^{*(\mathbf{k})} = \{x_i^{(k)} c_i^{(k)}\}$;
   (c) Each institution publishes the masked data $\mathbf{x}^{*(k)}$ separately. Considering $\mathbf{x}^* = \{\mathbf{x}^{*(1)}, \ldots, \mathbf{x}^{*(K)}\}$, this new collection $\mathbf{x}^*$ with sample size of N is the masked data of the original collection $\mathbf{x}$.

(ii) Generating simulated samples:
   (a) Calculate the moments of the masked data $\mathbf{x}^*$ and the noise C;
   (b) Reconstruct the density function of the original data $\mathbf{x}$ based on the masked data collection $\mathbf{x}^*$ using the moment-based estimation method implemented in the `R` package `MaskDensityBM`;
   (c) Generate a large set of resampled data $\tilde{\mathbf{x}} = \{\tilde{x}_i\}$ from the reconstructed density function. Use these data to replace the original data $\mathbf{x}$ for retrieving the statistical information of $\mathbf{X}$. The size of the simulated data M can be determined iteratively and must satisfy the $D_M < 0.007$ criterion (2).

(iii) Analysing resampled data:
   Apply a data mining technique directly to the resampled data $\tilde{\mathbf{x}}$ to obtain statistical information about the original data.

This framework covers several possible practical scenarios. If data owners want to combine their data with those of others to perform data analysis, they will need to follow all three steps, starting with masking their data (i). If a data owner is only interested in publishing his data but still wants to allow others to perform analysis, only Step (i) needs to be followed to release the masked data and the relevant information on the noise distribution. If a masked data collection $\mathbf{x}$ and relevant information about the noise distribution are already available, data users can start from the resampled data generation (ii).

3.1. Simulation Study

To evaluate the performance of the proposed framework, we conducted a simulation study under four different scenarios for cluster analysis, each representing different compositions of locations of means and proportions of samples. First, we present a short introduction to clustering. Cluster analysis or clustering is the task of grouping a set of objects so that objects in the same group (called a "cluster") are more similar to each other than to those in other groups. K-means, introduced by MacQueen [21], is a classic and still-popular algorithm for clustering analysis in data mining.

The K-means algorithm is a special case of the expectation-maximization (EM) algorithm for Gaussian mixture analysis, which decides cluster assignment based on posterior probabilities. Bishop [22] demonstrated that in the limit, the EM algorithm for the Gaussian mixture reduces to the K-mean result. In general, mixture model analysis aims to identify individual base distributions, which are used to form a mixture distribution for the underlying mixture model. Those individual base distributions are usually unimodal probability distributions. If the centres of those unimodal probability distributions can be identified with statistical significance, these centres are considered to be the centres of clusters. In other words, if we can sufficiently estimate the mixture distribution of data, we can use the EM algorithm for the clustering exercises using the mixture distribution. In this study, we used the commonly used `R` software `mclust` [23] (k-means clustering analysis tool) to carry out clustering analysis.

The simulation settings considered in this paper represent factors relevant to cluster analysis, which are how close the means are to each other and the proportions of the sizes of clusters. Let X be the random variable of the sample data; the four simulation settings are given as follows.

Setting 1. Two-group unequal proportions with a large difference in means: $f_X = 0.25 \times N(0, 1^2) + 0.75 \times N(6, 2^2)$.
Setting 2. Two-group equal proportions with a large difference in means: $f_X = 0.5 \times N(0, 1^2) + 0.5 \times N(6, 2^2)$.
Setting 3. Two-group equal proportions with close means: $f_X = 0.5 \times N(0, 1^2) + 0.5 \times N(4, 2^2)$.
Setting 4. Three-group equal proportions with two close means and one large mean: $f_X = 1/3 \times N(0, 1^2) + 1/3 \times N(6, 2^2) + 1/3 \times N(10, 2^2)$.

We generated 400 Monte Carlo simulations for each simulation setting; each simulation sample contained 900 data. For each simulation, we reconstructed the density function then generated $M = 18,000$ resampled data. We ensured that the sample size criterion $D_M < 0.007$ (2) was satisfied in all simulations so that the density of the resampled data was close to that of the original data. Both sample data and resampled data were then analysed using the `R` function `kmeans` for the cluster analysis results.

The performance of the proposed framework was evaluated in terms of estimation accuracy and variation for cluster analysis, under the condition that the cumulative density of resampled sample estimates that of the original sample well. For estimation accuracy, we examined the sampling bias, which is the average value of the differences in cluster means between the resampled data and sample data. To examine the estimation variability, we considered two measures, the sampling standard deviation (s.d.) and the coefficient of variation (CV). Sampling standard deviation is the standard deviation of the cluster means of the resamples, and CV measures the dispersion of the estimation by taking the ratio of the sampling standard deviation and sampling cluster means. We also included the root mean-square error (RMSE), which is the root average of the sum of the square of differences in cluster means between resampled values and sample data. RMSE can be used to directly compare the performance of the proposed method under different simulation settings.

Table 1 shows the results from the Monte Carlo simulations for different simulation settings, with various compositions of mean locations and cluster sizes. In terms of estimation accuracy, average biases were relatively small for all simulation settings, and better performance was achieved from the settings with equal proportions between clusters (Settings 2 and 3). Setting 4, with a smaller cluster size, is slightly more biased than others, possibly due to the reduced quality of fit from a smaller cluster size.

Estimation variability was relatively stable across different simulation settings. The slightly larger sampling standard deviation in Setting 4 indicates that estimation variations elevated when the original cluster sizes are small. Coefficients of variations are generally larger for smaller clusters with low means. This is consistent with the stable sampling standard deviation results, as the CV is the ratio between standard deviation and means; i.e., when the standard deviations are similar, smaller CVs are caused by smaller means.

Results of RMSE, which measures the dispersion of cluster means between resampled values and original sample data, can be used directly to compare the performance of the proposed framework for the four different settings. The most ideal scenario, Setting 2, with two groups of equal cluster sizes and a large difference in cluster means, has the smallest RMSE. This means that the dispersion is smallest and the proposed framework's performance was best in Setting 2. Equal cluster size and larger clusters also contribute to low dispersion (smaller RMSE). Dispersion was elevated in the case of smaller cluster sizes (Setting 4), which is consistent with the conclusion observed from the measure for estimation accuracy.

Table 1. Monte Carlo simulation results comparing the cluster analysis results from the resampled data to those of the sample data under four simulation settings, including average bias, sampling standard deviation (s.d.), coefficient of variation (CV), and root mean-square error of the estimation differences (RMSE).

		Average Bias	Sampling s.d.	CV	RMSE
Setting 1	Cluster 1 mean = 0	0.044	0.105	0.162	0.873
	Cluster 2 mean = 6	0.037	0.083	0.013	0.733
Setting 2	Cluster 1 mean = 0	-0.009	0.056	0.325	0.182
	Cluster 2 mean = 6	0.019	0.093	0.015	0.386
Setting 3	Cluster 1 mean = 0	-0.026	0.055	0.268	0.525
	Cluster 2 mean = 4	0.015	0.108	0.023	0.298
Setting 4	Cluster 1 mean = 0	-0.034	0.073	0.330	0.682
	Cluster 2 mean = 4	0.050	0.173	0.035	1.003
	Cluster 3 mean = 10	0.061	0.138	0.013	1.221

4. Real-Life Data Application

This section illustrates how to implement the framework proposed in Section 2.1 for data clustering and apply it to a real dataset. We applied the proposed framework to the Australian soybean dataset (more information on the study design and the data download link are available at http://three-mode.leidenuniv.nl/, accessed on 1 October 2021) [24], which contains data for 58 different genotypes of soybeans collected from eight experiments for six different soybean attributes. In the dataset, the 58 different lines (genotypes) of soybeans are 43 Australian lines and 15 other lines, of which 12 are from the US. Line 1–40 are local Australian selections from Mamloxi (CPI 172) and Avoyelles (CPI 15939).

In this example, we considered that each genotype of soybean is owned by a data provider and clustered the soybean genotypes based on the attribute seed size. The total number of data providers was 58, and there were 8 data points from each of the providers (genotype). The total number of observations for seed-size data N is $58 \times 8 = 464$. Each provider wants to know which cluster his/her data belongs to, when there is no access to the actual values of data from other providers. In particular, they are interested in which clusters their data can be classified into if there are two or three clusters.

Following the framework proposed in Section 3, all 58 data providers first agree on a noise distribution C. Assume that the probability density function of C is

$$f_C = 0.6 \times Uni(2,5) + 0.4 \times Uni(4,6).$$

Then, the data providers independently mask their raw seed-size data using the multiplicative noises C and publish their own masked values to create a collection of masked data for see size from 58 data providers. Figure 4 plots the masked data of seed size against the original values, showing the effectiveness of data masking. A given masked value corresponds to a large range of possible values of original data. This indicates that it is hard to accurately estimate the values of the original data from the masked values.

The second step is to reconstruct the density function of the original data based on the masked data and the information of the noise distribution C to generate resampled data. We applied the R package `MaskDensityBM` to the masked data collection and obtained the estimated density function associated with the set of the original seed-size data. The density function of the original seed size and its reconstructed density function are presented in Figure 5, which shows that the reconstructed density function preserves the two-mode feature and follows the pattern of the original density reasonably well. Then, the resampled data were generated from the reconstructed density function. The sample size required to satisfy the $D_M < 0.007$ criterion is 1856, approximately four times the original sample size of 464.

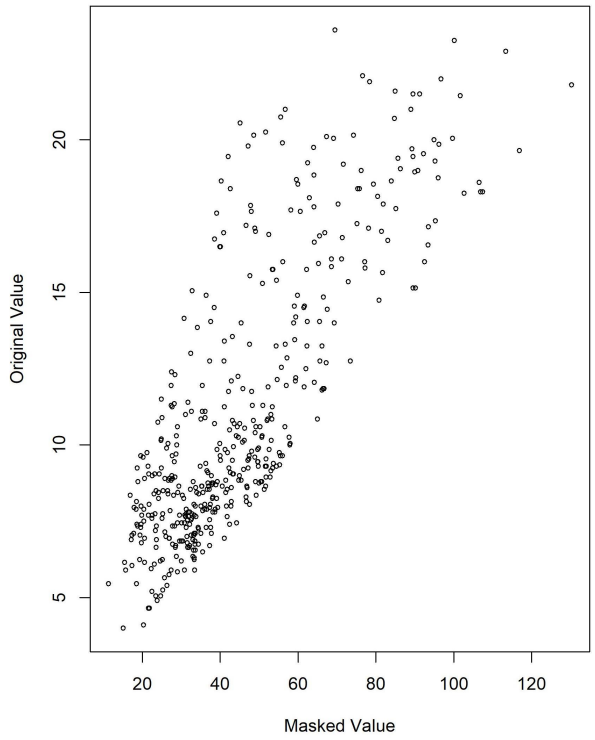

Figure 4. The scatter plot of the masked values and the original values for seed size.

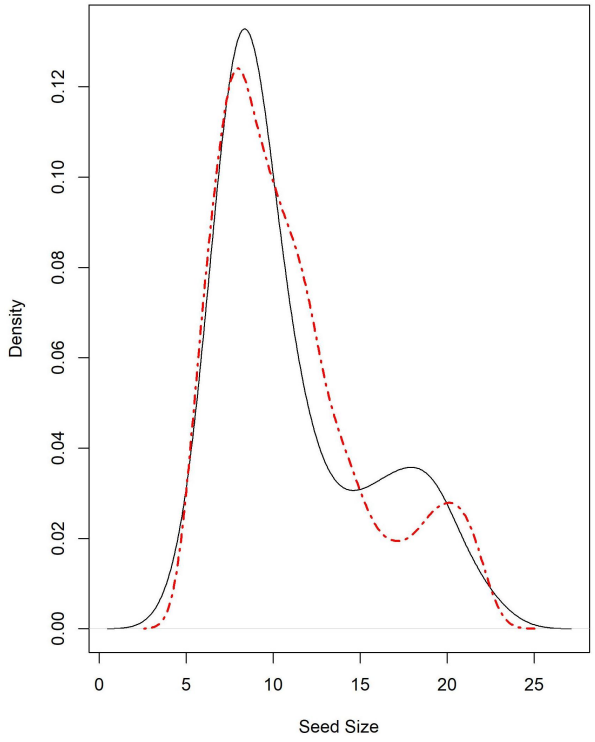

Figure 5. The plot of smoothed original density function (solid line) and reconstructed density function from 1856 simulated samples (dotted line).

We fitted both datasets—the resampled soybean seed-size data and the original data—using the *k*-means clustering analysis tool in R. We obtained the outputs of clustering analysis for $k = 2$ and 3 clusters. The outputs are given in Table 2. For $k = 2$, centres from the two datasets are similar to each other, having similar standard deviations. A

higher cluster mean in the resampled data was observed, though the difference in means was not statistically significant. This is consistent with what we observe in Figure 5; the reconstructed density for the resampled data has a second mode at larger seed-size values. A similar conclusion can be drawn for the $k = 3$ output, except that the larger mean only occurs in the last cluster of the resampled dataset, and there are two groups with a significant difference in means. However, this is not a surprise, as the density plot suggests that the original data are more likely to have two clusters (not three), so the k-means with $k = 3$ clusters may not fit the data well. Furthermore, the clusters show similar allocation (in proportion) between the two datasets for both $k = 2$ clusters, but not as good of a fit for $k = 3$ clusters.

Table 2. Comparison of outputs of the clustering analysis from the resampled dataset and original dataset. Here, the results include cluster mean, cluster standard deviation (s.d.), cluster size, and its corresponding proportion.

Number of Clusters $k = 2$		
	Resampled data	**Original data**
Cluster centre (*s.d.*)	8.882 (2.111 #) 17.631 (2.709)	8.686 (1.837) 17.503 (2.530)
Cluster size (*proportion*)	1412 (76.1%) 444 (23.9%)	335 (72.2%) 129 (27.8%)
Number of Clusters $k = 3$		
	Resampled data	**Original data**
Cluster centre (*s.d.*)	7.650 * (1.235) 11.896 (1.457) 18.942 * (1.949)	7.901 (1.253) 12.016 (1.526) 18.476 (1.980)
Cluster size (*proportion*)	902 (48.6%) 633 (34.1%) 321 (17.3%)	256 (55.2%) 108 (23.3%) 100 (21.6%)

* Sample means are statistically different between groups at the 5% significance level. # Sample standard deviations are statistically different between groups at the 5% significance level. Details of statistics and associated p values for comparing group means and variances in given in Table A1 in Appendix A.

We conclude that the resamples generated from reconstructed density can produce statistically equivalent results of the original data. However, this may not be guaranteed when the model fit is not appropriate. The analysis results from resampled data must be used with caution, as they depend not only on the quality of the reconstructed density, but also on the appropriate use of a data analysis technique for making inferences about the population parameter of interest.

5. Closing Remarks

The issues of data privacy are currently receiving widespread and significant attention. In general, methods for the statistical analysis of confidential data should be different from traditional methods. This paper proposes a data clustering analysis method for scenarios where data are independently collected from various data sources. These data are confidential and cannot be shared across data sources directly. The approach proposed is supported by the technique of reconstructing density functions based on noise-multiplied data. The method ensures that an original density function can be closely approached by the reconstructed density function. Therefore, we can retrieve accurate statistical information of the original data from the samples generated from its reconstructed density function. We detailed the application of the approach to a real-life dataset, assuming that the data have privacy issues.

The proposed framework is feasible in practice. Few traditional data analysis R tools can be directly applied for confidential data analysis due to privacy issues. The sample generated from the reconstructed density function plays the role of a "bridge", linking the

confidential data and the existing R tools. The proposed approach brings great convenience to realistic data analysis practices when data privacy is of concern, avoiding the need to develop special R tools for data analysis.

The framework developed in this paper is not limited to cluster analysis. Its applications extend to a broad range of data mining analyses. This paper only focuses on univariate data. However, we can apply the framework of the approach to multivariate data once a technique of reconstructing joint density function based on multivariate masked data is available. This technique for multivariate density estimation is under development and will be introduced soon in another paper.

Author Contributions: Conceptualization, P.O. and Y.-X.L.; methodology, Y.-X.L.; software, Y.-X.L. and P.O.; validation, P.O.; writing, P.O. and Y.-X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

It is important to evaluate the accuracy of the clustering analysis results from the resampled values. We compared the analysis outcomes to examine whether the clustering groups in the resampled and original data have statistically equal means and variances. We firstly tested the spread of the data from each allocated group using a F test of equal variance between the resampled data and original data groups. Depending on the F test result, either a pooled or unpooled t test was then used to test whether group means between the resampled data and the original data are statistically equivalent; for the groups with statistically different variance (p values from F test less than 0.05), an unpaired unpooled t test was performed to test the difference in group means, assuming a difference in group variances. Otherwise, an unpaired pooled t test was performed for the groups with statistically equivalent variance ($p > 0.05$ in F test of equal variance) in the resampled data and in the original data.

Table A1. Test statistics and associated p values for comparing the cluster means (t-tests) and cluster standard deviations (F tests) between the resampled and original Australian soybean data.

	F Tests for Equal Variance		***t* Tests for Equal Means**	
	F Statistic	***p***	**T Statistic**	***p***
Number of clusters k = 2	1.3206	0.0018	1.7041	0.0889
	1.1465	0.3554	0.4793	0.6318
Number of clusters k = 3	0.9684	0.734	−2.8607	0.004
	0.9116	0.5046	−0.7856	0.4324
	0.9689	0.8236	2.0799	0.03814

Table A1 includes the test statistics and p values from the relevant F tests and t-tests to compare the cluster variances and cluster means between the resampled and original Australian soybean data. The equal-variance tests (F tests) show that the allocated groups in the resampled and original data have nonsignificant difference (similar) variances, except for the first group in the cluster analysis with two clusters. The equal-mean tests (t tests) for the cluster analysis with two clusters are non-significant, suggesting that the resampled data and the original data produce the same groups means. However, when the model

fit is less appropriate (cluster analysis with three clusters), two groups show a significant difference in cluster means between the resamples and the original data.

References

1. Zhao, C.-M.; Luan, J. Data mining: Going beyond traditional statistics. *New Dir. Institutional Res.* **2006**, *131*, 7–16. [CrossRef]
2. Colombo, P.; Ferrari, E. Access control technologies for Big Data management systems: literature review and future trends. *Cybersecurity* **2019**, *2*, 1–13. [CrossRef]
3. Bertino, E.; Ghinita, G.; Kamra, A. Access Control for Databases: Concepts and Systems. *Found. Trends® Databases* **2011**, *3*, 4–7.
4. Torra, V. *Data Privacy: Foundations, New Developments and the Big Data Challenge*; Springer International: Cham, Switzerland, 2017.
5. Torra, V.; Navarro-Arribas, G. Big Data Privacy and Anonymization. In *Proceedings of the Privacy and Identity Management. Facing up to Next Steps. Privacy and Identity 2016*; IFIP Advances in Information and Communication Technology; Springer: New York, NY, USA, 2016.
6. Templ, M. *Statistical Disclosure Control for Microdata: Methods and Applications in R*; Springer International: Cham, Switzerland, 2017; pp. 99–132.
7. Aldeen, Y.; Sallen, M.; Razzqque, M. A comprehensive review on privacy preserving data mining. *Springerplus* **2015**, *4*, 1–36. [CrossRef] [PubMed]
8. Sachan, A.; Roy D.; Arun, P.V. An analysis of privacy preservation techniques in data mining. *Adv. Comput. Inf. Technol.* **2013**, *3*, 119–128.
9. Dwork, C. Differential privacy. In Proceedings of the 33rd International Colloquium on Automata, Languages and Programming, Venice, Italy, 10–14 July 2006; pp. 1–12.
10. McSherry, F.; Talwar, K. Mechanism Design via Differential Privacy. In Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science, Providence, RI, USA, 20–23 October 2007; pp. 94–103.
11. Ács G.; Castelluccia, C. I Have a DREAM! (DiffeRentially privatE smArt Metering). *Inf. Hiding* **2011**, *6958*, 118–132.
12. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **2014**, *9*, 211–407. [CrossRef]
13. Bambauer, Y.; Jane, R.; Muralidhar, K.; Sarathy, R. Fool's Gold: An Illustrated Critique of Differential Privacy. *Vanderbilt J. Entertain. Technol. Law* **2014**, *16*, 13–47.
14. Fisher, R. On the Mathematical Foundations of Theoretical Statistics. *Philos. Trans. R. Soc. A* **1922**, *222*, 594–604.
15. Nayak, T. K.; Sinha, B.; Zayatz, L. Statistical properties of multiplicative noise masking for confidentiality protection. *J. Off. Stat.* **2011**, *27*, 527–541.
16. Agrawal, R.; Srikant, R. Privacy-preserving data mining. *ACM Sigmod Rec.* **2000**, *29*, 439–450. [CrossRef]
17. Kargupta, H.; Datta, S.; Wang, Q.; Sivakumar, K. On the privacy preserving properties of random data perturbation techniques. In Proceedings of the Third IEEE International Conference on Data Mining, Washington, DC, USA, 19–22 November 2003; pp. 99–106.
18. Lin, Y.X. Density approximant based on noise multiplied data. In Proceedings of the International Conference on Privacy in Statistical Databse, Ibiza, Spain, 17–19 September 2014; Lecture Notes in Computer Science Series; pp. 89–104.
19. Lin, Y.X.; Krivitsky, P. Reviewing methods for estimating density function based masked data. In Proceedings of the International Conference on Privacy in Statistical Databse, Valencia, Spain, 26–28 September 2018; Lecture Notes in Computer Science Series; pp. 231–246.
20. Lin, Y.X. Mining the Statistical Information of Confidential Data from Noise-Multiplied Data. In Proceedings of the 3rd IEEE International Conference on Big Data Intelligence and Computing, Orlando, FL, USA, 6–11 November 2017.
21. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, UAS, 21 June – 18 July, 1965 Volume 1: Statistics.
22. Biship, C. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006; p. 423.
23. Scrucca, L.; Fop, M; Murphy, T.B.; Raftery, A. mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R J.* **2016**, *8*, 289–317. [CrossRef] [PubMed]
24. Shorter, R.; Byth, D.; Mungomery, V. Genotype by environment interactions and environmental adaptation. ii. Assessment of environmental contributions. *Aust. J. Agric. Res.* **1977**, *28*, 223–235. [CrossRef]