A New Geometric Data Perturbation Method for Data Anonymization Based on Random Number Generators

Merve Kanmaz[1,*], Muhammed Ali Aydin[2] and Ahmet Sertbas[2]

[1]*Computer Programming Department, Istanbul University-Cerrahpasa, Istanbul, Turkey*
[2]*Computer Engineering Department, Istanbul University-Cerrahpasa, Istanbul, Turkey*
E-mail: merve.kanmaz@iuc.edu.tr; aydinali@iuc.edu.tr; asertbas@iuc.edu.tr
[*]*Corresponding Author*

Received 23 June 2021; Accepted 05 August 2021;
Publication 14 October 2021

Abstract

With the technology's rapid development and its involvement in all areas of our lives, the volume and value of data have become a significant field of study. Valuation of the data to this extent has produced some consequences in terms of people's knowledge. Data anonymization is the most important of these issues in terms of the security of personal data. Much work has been done in this area and continues to being done. In this study, we proposed a method called RSUGP for the anonymization of sensitive attributes. A new noise model based on random number generators has been proposed instead of the Gaussian noise or random noise methods, which are being used conventionally in geometric data perturbation. We tested our proposed RSUGP method with six different databases and four different classification methods for classification accuracy and attack resistance; then, we presented the results section. Experiments show that the proposed method was more successful than the other two classification accuracy, attack resistance, and runtime.

Journal of Web Engineering, Vol. 20_6, 1947–1970.
doi: 10.13052/jwe1540-9589.20613

Keywords: Privacy-preserving data mining, data anonymization, data perturbation, geometric perturbation, random number generators.

1 Introduction

In today's technology age, data is an essential source of information. With the development of technology, the increase in social media applications, and the introduction of concepts such as the internet of things, the volume and dimensionality of collected data have increased considerably. The value of the data has grown with the growing size [1]. We live in the age of data, and the oil of our age is data. As such, there are many different sources of big data. For example, an aircraft engine generates data for 5000 different components per second. According to the information obtained from a site that keeps the statistics of the shares on social media, 500 million tweets are posted daily only on Twitter, and 90 thousand videos are watched per minute on YouTube [2]. Contents produced due to video, picture applications, data read from medical devices, patient follow-up data, and most importantly, the data produced by the IoT devices are sufficient examples to define the many titles in the concept of big data.

With the acceleration of technological developments regarding collecting and storing such extensive data, data mining studies, which are the science of extracting meaningful information from data, have also gained importance [3]. In addition, it has become a new field of study that needs to be worked on to ensure the privacy of sensitive data about individuals and organizations in mining operations or in any area where data is accessible. The term confidentiality used here should not only be considered as the protection of data from threats from attackers, but it also means of preventing unauthorized access and use [4, 5].

There are many successful de-identification and data security solutions for traditional (relational) databases. However, with the concept of big data, different robust infrastructures such as Hadoop, Spark, Cloud that can store and process big data have been developed. Many organizations need innovative solutions that will not create performance problems in accessing data while ensuring data privacy in these distributed work environments.

De-identification is the process of making the data independent from the person who owns it [6]. Even if it matches with other data, it cannot be matched with a specific identity in any way or cannot be identified indirectly. While the de-identification process, each record is divided into the following attributes [7].

- **Identifier (ID):** Characteristics that can identify the person alone without the need for any other feature. Such as ID number, account number.
- **Quasi-identifier (QID):** The feature sets cannot identify the person alone but can identify by associating with other data sets. Like gender and zip code.
- **Sensitive attribute (SA):** Characteristics of the person that should remain hidden from other individuals. Such as salary, grade, or disease.
- **Non-Sensitive Attribute:** Features that will not disclose the person, even if they fall into the hands of people who do not have access. Like hobby, reviews, insurance company

Privacy-preserving data mining procedures in the literature can be categorized into three key groups. These are reconstruction-based methods, heuristic methods, and cryptographic methods. According to the algorithm they apply, these methods can de-identify data using descriptive, semi-descriptive, or sensitive attributes. Frequently used anonymization methods:

- **Generalization:** It is the process of replacing semi-descriptive attributes with values that will less describe the attribute. In other words, the relevant data is translated into a more general value rather than a personal value. Generalization types: It is known as full domain generalization, cell generalization, multidimensional generalization, sibling generalization, and subtree generalization.
- **Suppression:** Some known values in this method are replaced with special characters, hiding their actual value. The purpose of the hiding process is to reduce the probability of an accurate estimation. The hiding method can be done separately on the cell, on the record, and on the value.
- **Anatomization:** The table containing sensitive data is split into two separate tables; semi-descriptors and sensitive attributes. In other words, the data are not generalized or hidden; they are kept in two separate tables in a way that they cannot disclose personal information. Both tables have a common feature in order not to break the relationship between them.
- **Permutation:** In this method, data is divided into groups, then sensitive data within each group is mixed and de-identified.
- **Perturbation:** In the perturbation method, the values on the data set are replaced with synthetic values that have no real value. With this change, there will be less deterioration in the distribution of data and statistical

calculations. It has options such as data exchange (swapping records between pairs), adding noise (adding an equal amount of data to the original data), creating synthetic values (determining meaningless data).

Various studies using the above-mentioned methods have been done by now. And in this manuscript, we use the random number generators and geometric data perturbation based new method (RSUGP) for efficient privacy preserving. RSUGP is an irreversible input perturbation mechanism with a new noise model. Using the random number generators for noise addition is a novel method for anonymization and it is one of our contribution. Our main contribution is developing a privacy preserving data publishing algorithm which is independent of data set and can be applied all numerical attributes. We prove experimentally that RSUGP provides more privacy guarantee with better classification accuracy, better attack resistance and faster than comparable methods.

The rest of the paper is organized as follows. Section 2 provides a summary of existing related work about anonymization methods. Geometric data perturbation and the technical details of RSUGP are described in Section 3. Performance results of the proposed model are shared in Section 4. Furthermore, in the last section, the study was concluded by mentioning the contribution of the proposed method.

2 Literature Reviews

Privacy-preserving data mining methods are classified into three classes: reconstruction-based methods, heuristic approaches, and cryptographic methods [8]. In reconstruction-based techniques, the sensitive values on the original data set are removed and replaced with different values. Heuristic methods are models used to measure the level of privacy protection. Cryptographic methods allow data to be changed in a structure like encryption and have more time complexity than other methods [8]. Despite this complexity, it is not appropriate and practical to use as an anonymization technique because it reduces data availability against security [9].

Heuristic methods are mostly statistical methods used to determine the boundaries of confidential information. The best known of these methods are k-anonymity, l-diversity, and t-closeness. K-anonymity is a privacy model that ensures that each record in the published data cannot be distinguished from at least $k - 1$ records in the data if the attacker somehow obtains semi-descriptive values [10]. In the studies on the lack of $k = 4$ anonymity, it was determined that in cases where the sensitive attributes diversity is low, k-anonymity does not adequately protect the confidentiality and can be inferred from the data. Thereupon, the l-diversity principle, based on the relationship between semi-descriptive and sensitive qualities, was proposed. If there is at least 1 kind of sensitive feature in a semi-descriptive group selected according to the method, it is said that l-diversity has been achieved [11]. Because l-diversity prevents disclosures by providing diversity in personal data, it does not provide sufficient protection because it is not concerned with the content and sensitivity of the data, and the t-closeness principle has been put forward [12]. In t-closeness, the data are anonymized by dividing data into subclasses the according to calculated proximity degree.

Reconstruction-based methods are also called perturbation techniques. Perturbation is used to replace data or data sets with synthetic data with the same distribution as themselves. The most important and biggest challenge in data modification is to ensure that the quality and balance of the data are not compromised while maintaining privacy. Although there are many approaches in privacy-preserving data mining, data perturbation is one of the frequently used methods because it is a simple and effective method [13]. Data perturbation is examined in two classes as input perturbation and output perturbation. While the input perturbation uses one of the additive noise or multiplicative noise methods, the noise addition and rule hiding approaches [14–16] are applied together in output perturbation [17]. Input perturbation; can be applied one dimensional with additive perturbation [18, 19], random response [20] and swapping [21–24], while condensation [25], random rotation [26–30], random projection [31] and geometric perturbation [32], can be applied multi-dimensionally.

In the literature review in recent years, it is observed that the anonymization studies have focused on perturbation methods. In a study published Ph.D. thesis [33], the author proposed a k-anonymity method to keep data confidentiality at the highest level while minimizing data loss and latency on data flowing in the Apache Spark environment. In a different study [17], an efficient de-identification algorithm is proposed on data flowing over IoT devices, using the method they call P2RoCAl. In another study [34], the authors put forward a model that provides data privacy by using chaotic maps, which are frequently used in many fields.

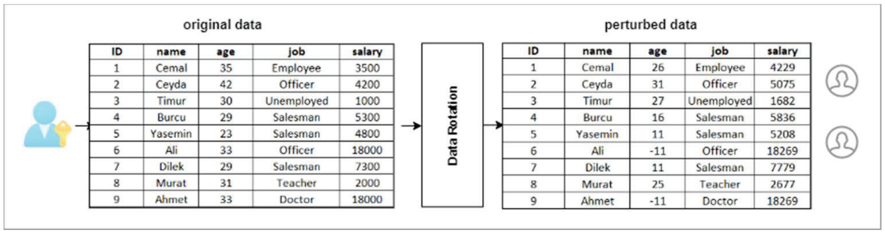
In the study on geometric data perturbation [35], the authors proposed a 3-stage model that works in harmony with different data mining methods. They also published detailed analyses on multi-column privacy in the study. In a different study [36], the data set containing personal health records was developed using the Gaussian noise model in geometric perturbation and compared with the AES encryption method in terms of operating times. In a similar study, geometric perturbation was performed using the Gaussian noise model, and healthcare data set [37]. The authors in [38] developed geometric data perturbation and a classification model accordingly. Classification and clustering methods are essential issues for privacy-preserving data mining. In the study in [39], the clustered data using the k-mean clustering method were perturbed using the rotation process with different angles on the cluster centers. In another study [40], a 4-dimensional rotation model has been proposed for anonymization. In this model, the data were split into two groups, and their values were changed by rotating them on the xy and then zw axes.

The confidentiality of data collected on the sensor node for transmitting in wireless sensor networks is also an important issue. In a study [41], geometric data perturbation was used to provide this confidentiality. Geometric data perturbation includes method steps. A different study about the order of these stages [42] examined changing the order of the steps; rotation, adding noise, scaling on the result was shared. In another large-scale study, geometric data perturbation was analysed in detail [43]. In the same study, the successful results obtained from the model proposed by the authors by combining the data separation method and the geometric approach were shared. In a different study of the same authors [44], an efficient and reliable perturbation method was proposed using the Laplace noise method.

When all these studies examined, geometric data perturbation models become special with step order or noise addition step. In most of the studies, it is seen that Gaussian noise model is a frequently used because of its ability to produce different noise each time [35–40]. And another one uses the Laplace noise [44]. And some of the studies uses random value in range 0–1. Our RSUGP method stands out here and uses a different noise model than others. In proposed model data is used for input for generating noise.

3 Proposed Method: RSUGP

With the current studies reviewed, it has been seen that perturbation methods continue to be used actively for de-identification studies. Geometric data perturbation is one of the perturbation methods with a high success rate. For this reason, in this study, we focused on geometric data perturbation and proposed a new model with random number generators.



original data

ID	name	age	job	salary
1	Cemal	35	Employee	3500
2	Ceyda	42	Officer	4200
3	Timur	30	Unemployed	1000
4	Burcu	29	Salesman	5300
5	Yasemin	23	Salesman	4800
6	Ali	33	Officer	18000
7	Dilek	29	Salesman	7300
8	Murat	31	Teacher	2000
9	Ahmet	33	Doctor	18000

perturbed data

ID	name	age	job	salary
1	Cemal	26	Employee	4229
2	Ceyda	31	Officer	5075
3	Timur	27	Unemployed	1682
4	Burcu	16	Salesman	5836
5	Yasemin	11	Salesman	5208
6	Ali	-11	Officer	18269
7	Dilek	11	Salesman	7779
8	Murat	25	Teacher	2677
9	Ahmet	-11	Doctor	18269

Figure 1 Rotation example.

3.1 Geometric Perturbation

Geometric Data Perturbation, one of the reconstruction-based anonymization methods, consists of rotation, translation, and noise addition steps. It is processed as RX-Rotation, T-Transform, Δ-Add Noise steps on the formula:

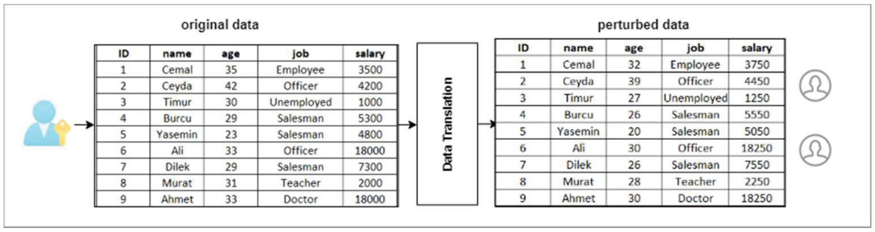
$$\mathrm{G(X) = RX + T} + \Delta \tag{1}$$

3.1.1 Multiplicative transformation

Data rotation is the perturbation of sensitive data on the D data set by the multiplicative transformation method. The data set consisting of sensitive numerical attributes are multiplied by a matrix for transformation. This matrix can be a random rotation matrix or a random projection matrix. While the rotation matrix preserves the distance exactly after transforming the data, the projection matrix maintains approximately. Therefore, the random rotation matrix is more preferred. A rotation matrix is a matrix used to rotate points or points on an N-dimensional coordinate system by a specified angle on a specified axis (Formula 1). The rotation matrix is special and must be generated according to some rules [26]. Figure 1 shows the result of rotating the age and salary sensitive attributes of the sample data set clockwise to $\alpha = 13.7$ degrees.

3.1.2 Translational transformation

Data translation is the perturbation of sensitive data on the D data set with the additive noise method. While the translation process, a constant value is added to the whole value of an attribute. The translation matrix must be the same positive or negative values for each same attribute in the data set. In Figure 2, the result of the (-3,250) values added to the sensitive attributes of the sample data set's age and salary are monitored.



original data

ID	name	age	job	salary
1	Cemal	35	Employee	3500
2	Ceyda	42	Officer	4200
3	Timur	30	Unemployed	1000
4	Burcu	29	Salesman	5300
5	Yasemin	23	Salesman	4800
6	Ali	33	Officer	18000
7	Dilek	29	Salesman	7300
8	Murat	31	Teacher	2000
9	Ahmet	33	Doctor	18000

perturbed data

ID	name	age	job	salary
1	Cemal	32	Employee	3750
2	Ceyda	39	Officer	4450
3	Timur	27	Unemployed	1250
4	Burcu	26	Salesman	5550
5	Yasemin	20	Salesman	5050
6	Ali	30	Officer	18250
7	Dilek	26	Salesman	7550
8	Murat	28	Teacher	2250
9	Ahmet	30	Doctor	18250

Figure 2 Translation example.

3.1.3 Noise addition

As a result of the investigated studies, similar rotation and translation methods have been used in the geometric data perturbation. Random noise has been added as a noise model in some studies. However, in most studies, Gaussian noise, which gives more effective results, was used because it uses the data, namely the standard deviation and the average of the data. Gauss generates a random noise value each time, depending on the data.

Based on this, we focused on the concept of randomness. Randomness is a concept that is needed for many fields such as statistics, game theory, simulation, numerical analysis, entertainment. Primarily cryptological applications mostly operate with randomly generated numbers. Session keys, signature keys and parameters, authentication protocols, temporary keys, zero-knowledge proof, initial vectors for block ciphers, and blinding and masking are protection measures against side-channel attacks are just some of the cryptological applications that require random numbers [45–46].

3.2 Random Number Generators and Linear Congruential Generator

To generate random numbers, generally a random number generator is used. Random number generators are divided into two categories according to the mechanism they use.

- **Hardware Random Number Generator** (*Real Random Number Generator, HRNG*): It is a method of generating random numbers without using a computer program by using the physical properties of the computer such as thermal noise, the photoelectric effect. Numbers produced by this method are entirely random; that is, they are unstable and cannot be predicted because they are not produced according to a rule.
- **Pseudo-Random Number Generator** (*Pseudo-random number generator, PRNG*): It generates random numbers starting from an initial

condition with an algorithm. It is such that no relationship can be established between the numbers produced. However, the numbers produced in this way are not entirely random as they are formed according to a specific rule and are therefore known as pseudo-random numbers. Despite its ease, speed, and inexpensive features, it is a widely used method. The most widely used model is Linear Congruential Generators. Linear Congruence Generators work like the system also called clock arithmetic.

$$X_{i+1} = aX_i + c(mod\, m) \tag{2}$$

m is modulus where $m > 0$, a is the multiplier where $0 = a < m$, c is the increment where $0 = c < m$ and initial value of sequence as seed $0 = X_0 < m$ where $X_0 \in \{1, 2 \ldots m - 1\}$. After Formula (2) is applied for the specified number of iteration, the random number value is obtained.

While Linear Congruential Generators process, the result changes with different values of the parameters a, c, and m. As a result of the studies, it has been seen Gaussian noise successful because of using data for creating noise. Accordingly, the data were chosen for the initial value, the maximum value of the data as the "*m*" parameter, the standard deviation as the "*a*" parameter, and the average of the data as the "*c*" parameter. Thus, a data-based suitable noise model was generated, and predictability was minimized.

Modelling and flowchart of the proposed random number generators-based noise generation method are shown in Figures 3 and 4. The pseudo-code of this model is given in Algorithm 1. In addition, the whole model applied is shown in Figure 5.

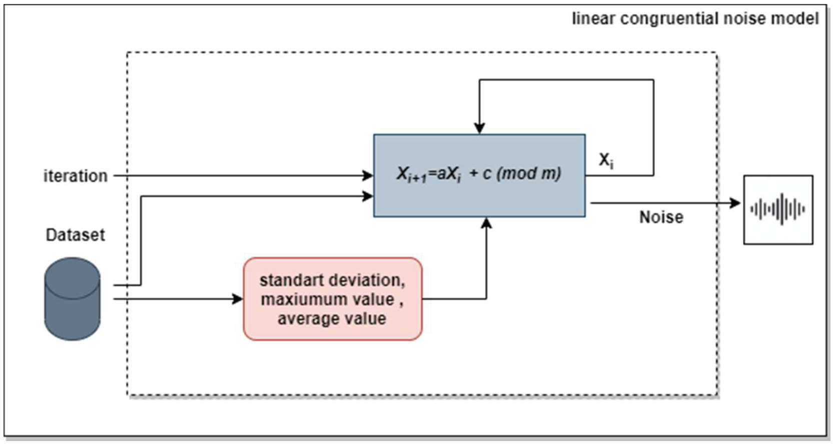


Figure 3 Linear congruential noise model.

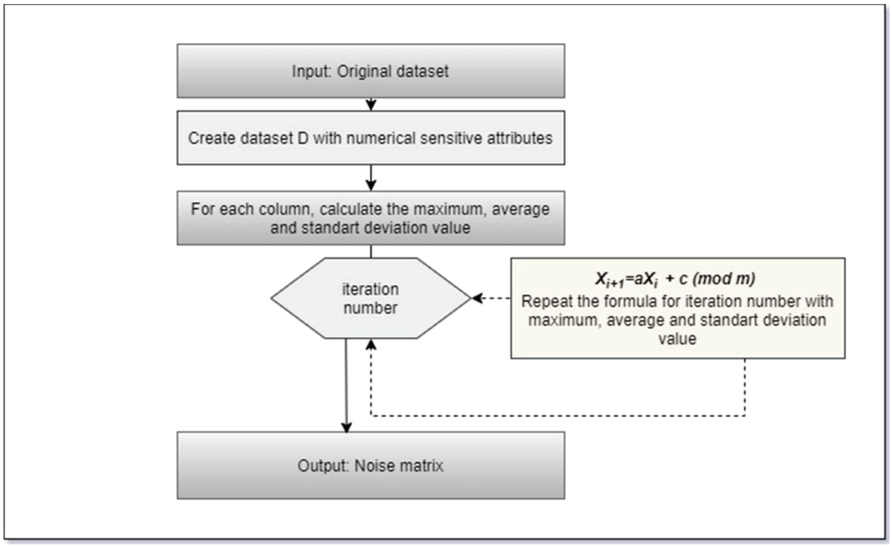

Input: Original dataset
Create dataset D with numerical sensitive attributes
For each column, calculate the maximum, average and standart deviation value
iteration number
$X_{i+1}=aX_i + c \ (mod\ m)$
Repeat the formula for iteration number with maximum, average and standard deviation value
Output: Noise matrix

Figure 4 Linear congruential noise model flowchart.

Algorithm 1 RSUGP noise

```
Input: Original dataset $D$, numerical sensitive attributes $SA(SA_1, SA_2, SA_3 \ldots SA_N)$
Output: RSU noise matrix $N$
Initial assignments: iteration
$d = |D|$ (number of data)
$s = |SA|$ (number of sensitive attribute)
for $i = 1$ to $s$ do
    $attStd_i =$ standart deviation of the sensitive attribute
    $attMean_i =$ average value of the sensitive attribute
    $attMax_i =$ maximum value of the sensitive attribute
    for $j = 1$ to $d$ do
        $N_i = D_i$
        for $k = 1$ to iteration do
            $N_{ij} = \mathrm{mod}\ (attStd_i * N_{ij} + attMean_i, attMax_i)$
        end for
        $D_i = N_i$
    end for
end for
generated noise matrix for output
$N = D$
```

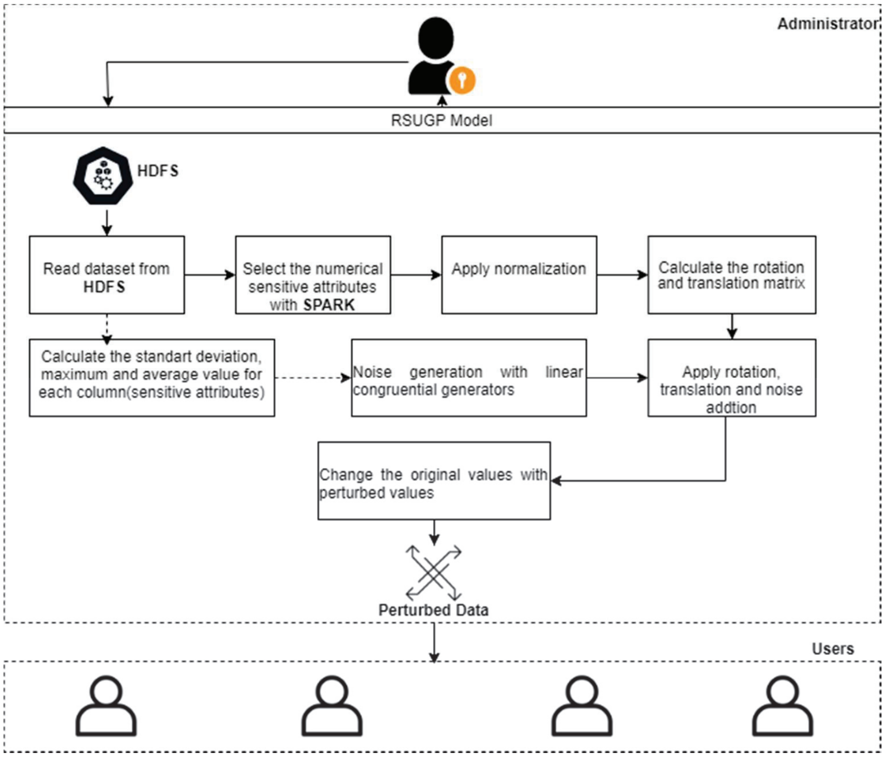

Administrator
RSUGP Model
HDFS
Read dataset from HDFS
Select the numerical sensitive attributes with SPARK
Apply normalization
Calculate the rotation and translation matrix
Calculate the standart deviation, maximum and average value for each column(sensitive attributes)
Noise generation with linear congruential generators
Apply rotation, translation and noise addtion
Change the original values with perturbed values
Perturbed Data
Users

Figure 5 RSUGP model diagram.

4 Performance Evaluation

In this chapter, the proposed algorithm has been measured with some performance metrics. These metrics are set classification accuracy, attack resistance, and runtime. Friedman's test was applied to evaluate the compared methods. The recommended method was implemented using Hadoop HDFS and Spark, installed in the virtual machine on a personal computer with 16GB RAM, Intel Core i5-1035G1 processor, and Windows 10 operating system, and Zeppelin was used as the editor. Weka 3.8 was used for classification methods, MATLAB R2018a for attack resistance, and SPSS for statistical evaluation of the results.

Table 1 Detailed description of used datasets

Datasets	Number of Records	Number of Attributes	Number of Classes	Used Attributes
Iris[1]	150	4	3	SepalLengthCm, SepalWidthCm PetalLengthCm, PetalWidthCm
Heart StatLog[2]	270	14	2	resting_blood_pressure, serum_cholestoral
Wine Quality-White[3]	4898	12	7	volatile acidity, citric acid, residual sugar,chlorides
Fried[4]	40768	11	2	x1, x2, x3, x4, x5, x6, x7, x8, x9, x10
Bank[5] Management	45212	17	2	age, balance
Electricity[6]	45313	9	2	nswprice, nswdemand

[1] https://archive.ics.uci.edu/ml/datasets/iris

[2] https://archive.ics.uci.edu/ml/datasets/statlog+(heart)

[3] https://archive.ics.uci.edu/ml/datasets/wine+quality

[4] https://www.openml.org/d/901

[5] https://archive.ics.uci.edu/ml/datasets/BankMarketing

[6] https://www.openml.org/d/151

4.1 Dataset Description

Performance results of the proposed model have been applied on six different data sets. These data sets were determined from the data sets of different sizes, widely used in data de-identification studies. Numerical sensitive attributes were determined while anonymizing the data sets. Detailed dataset description with used attributes is given in Table 1.

4.2 Classification Accuracy

Classification accuracy refers to the percentage of data sets that have been placed in the correct classes after classification. TP is the number of positive groups labelled positive, TN is the number of negative groups labelled as negative, T is all positives, and N is all negatives then:

$$\text{Classification Accuracy} = (TP + TN)/(P + N) \tag{3}$$

Classification accuracy of the proposed model is investigated using four different classifiers which are Naïve Bayes, J48, Decision Table and OneR.

2-fold, 5-fold, 10-fold cross validation are performed for all classifiers. For k-fold cross validation technique, the results of the proposed model with six different sized data sets are demonstrated in Table 2.

Classification accuracy of the model was compared with three different results: the original classification accuracy, the accuracy created by generating random noise, and the accuracy created using the Gaussian noise model. When the results in Table 2 are examined, a rise in k value causes small increase in accuracy generally. And on data sets which have numeric values only, all methods have smaller accuracy rate than other databases.

Closest result to the original accuracy rate is RSUGP almost every row. The comparisons were made using a nonparametric statistical comparison test: Friedman's rank test, which is analogous to a standard one-way repeated-measures analysis of variance [47]. Friedman's rank test further supports this argument by returning the highest mean rank for RSUGP's classification accuracy results on Table 2's last row. It can be also seen that a rise in k increase FMR value just as accuracy rate.

As seen from the table, the proposed algorithm shows better or equal performance in all cases of Naïve Bayes and J48 classification algorithms compared to the existing algorithms. In Decision Table classifier results of compared methods nearly same mostly. In OneR classifier, although RSUGP is the closest method, the difference between the RSUGP and the original result is much more. The J48 and Naïve Bayes classifier is better than two classifiers in terms classification accuracy for all algorithms.

4.3 Friedman's Mean Rank

Classification accuracy results were evaluated with non-parametric statistical testing methods, Friedman's test [47]. According to Friedman's test, repeated tests with different conditions are evaluated according to each other, and a value is produced.

$$F = \left[\frac{12}{Nk(k+1)} + \int_{i=1}^{k} R_i^2\right] - 3N(k+1) \tag{4}$$

The higher value indicates that the method shows a distinct difference compared to other methods. The last row in Table 2 shows the FMR value of the compared methods. The test statistics of the experiment had a $\chi 2$ value of 24, a degree of freedom of 2 and a p-value of 34,889. According to the following rank values, it is seen that the proposed method has a significant difference from other methods and gives better results.

Table 2 Classification accuracy results

		2-Fold Cross Validation				5-Fold Cross Validation				10-Fold Cross Validation			
Data Sets		NB	J48	DT	OneR	NB	J48	DT	OneR	NB	J48	DT	OneR
Iris	Original	96.6	93.3	94.0	94.0	96.0	96.0	92.6	92.6	96.0	96.0	92.6	92.0
	Random Noise	92.0	84.4	84.4	87.3	92.0	83.0	86.6	84.6	91.3	86.0	88.0	80.0
	Gauss Noise	88.6	87.5	83.3	83.3	88.7	87.5	83.5	83.5	88.8	87.7	83.5	83.5
	RSUGP	95.5	90.2	87.3	88.4	95.6	90.5	87.3	88.5	95.6	90.8	87.3	88.5
Heart (statlog)	Original	82.2	80.0	74.0	67.7	85.1	77.7	77.0	70.7	83.7	76.6	82.5	72.5
	Random Noise	81.8	80.3	73.8	67.7	84.0	75.7	75.0	70.7	84.0	75.7	82.0	72.4
	Gauss Noise	76.5	71.2	70.8	67.7	76.8	71.2	71.8	70.7	76.9	72.8	71.8	72.4
	RSUGP	82.9	79.6	74.0	67.7	84.8	78.2	77.7	70.7	84.4	76.5	82.5	72.5
Wine Quality	Original	53.6	44.8	48.7	39.5	56.7	44.8	47.1	44.0	56.7	44.8	44.5	45.0
	Random Noise	50.6	44.8	48.7	35.6	51.4	44.8	47.0	36.7	51.1	44.8	44.3	36.8
	Gauss Noise	50.4	44.8	48.7	34.5	51.3	44.8	47.0	34.6	51.1	44.8	44.2	34.7
	RSUGP	55.8	44.8	48.7	39.4	56.7	44.8	47.1	43.5	56.7	44.8	44.5	44.5
Fried	Original	86.4	88.3	83.0	68.8	86.5	88.9	83.3	69.1	86.5	89.4	83.4	69.2
	Random Noise	71.9	69.9	69.2	60.0	71.9	69.9	69.9	61.2	71.9	70.0	69.7	61.5
	Gauss Noise	70.6	68.7	69.2	60.2	70.7	68.7	69.3	60.5	71.2	68.8	69.5	60.7
	RSUGP	79.7	76.4	73.7	62.4	79.7	76.9	74.1	62.5	79.7	77.4	74.0	62.7
Bank Management	Original	87.8	90.1	90.0	88.4	87.9	90.4	90.0	88.5	88.0	90.3	90.0	88.5
	Random Noise	87.4	90.0	90.0	88.0	87.6	90.1	90.0	88.1	87.6	90.0	90.0	88.1
	Gauss Noise	85.5	88.9	89.1	88.0	85.5	89.0	89.1	88.2	85.5	89.5	89.1	88.2
	RSUGP	87.8	90.1	90.0	88.4	87.9	90.3	90.0	88.5	88.0	90.2	90.0	88.5
Electricity	Original	72.9	87.88	78.9	75.4	72.9	90.5	79.7	75.8	72.9	91.1	79.9	76.0
	Random Noise	70.6	81.2	77.5	70.1	70.7	82.5	76.1	71.8	71.3	88.2	76.4	71.5
	Gauss Noise	60.6	79.3	72.1	64.8	60.6	81.6	72.7	64.8	60.6	82.7	73.0	64.7
	RSUGP	72.7	87.7	78.8	72.3	72.6	89.3	79.4	72.8	72.8	90.2	79.5	72.8
FMR Values	Random Noise			1,94				1,92				1,90	
	Gauss Noise			1,25				1,19				1,21	
	RSUGP			2,81				2,90				2,90	

4.4 Attack Resistance

When the literature is reviewed, there are various attack methods to extract the original data from the data used against the matrix multiplication-based de-identification methods [48]. The most used are ICA (Independent Component Analysis), I/O attacks, and NI (Naive Interference). The measured value is considered the standard deviation of the difference between the original data and the perturbed data, the amount of change in the data. The results of the tests performed using these methods are shared in Table 3.

When Table 3 is examined, it indicates that the higher value is more resistant to attacks. At the end of the table, the comparison of the methods is again given the FMR values. When these values are examined, it is observed that the proposed method is more resistant to attacks.

Table 3 Attack resistance results

Datasets	Algorithms	ICA$_{\text{avg}}$	ICA$_{\text{min}}$	IO$_{\text{avg}}$	IO$_{\text{min}}$	NI$_{\text{avg}}$	NI$_{\text{min}}$
Iris	Random Noise	0,83	0,71	0,34	0,08	1,97	1,95
	Gaussian Noise	0,78	0,77	0,58	0,5	1,92	1,84
	RSUGP	0,85	0,78	0,48	0,19	1,98	1,96
Heart	Random Noise	0,82	0,73	0,06	0,02	1,89	1,83
	Gaussian Noise	0,75	0,73	0,67	0,55	1,93	1,92
	RSUGP	0,77	0,73	0,65	0,56	1,98	1,96
Wine Quality-White	Random Noise	0,72	0,67	0,02	0,46	1,8	1,77
	Gaussian Noise	0,72	0,72	0,5	0,2	1,89	1,85
	RSUGP	0,75	0,69	0,5	0,03	1,89	1,79
Fried	Random Noise	0,71	0,7	0,57	0,39	1,63	1,49
	Gaussian Noise	0,71	0,66	0,53	0,32	1,64	1,5
	RSUGP	0,74	0,69	0,45	0,4	1,61	1,47
Bank Management	Random Noise	0,99	0,99	0,01	0,01	1,95	1,95
	Gaussian Noise	0,82	0,7	0,53	0,35	1,98	1,97
	RSUGP	0,85	0,71	0,45	0,33	1,99	1,99
Electricity	Random Noise	0,8	0,51	0,57	0,51	1,93	1,9
	Gaussian Noise	0,84	0,77	1,51	0,4	1,91	1,89
	RSUGP	1,08	0,85	1,46	1,4	1,99	1,95
FMR Values	**Random Noise:** 1,64		**Gaussian Noise:** 1,97			**RSUGP:** 2,39	

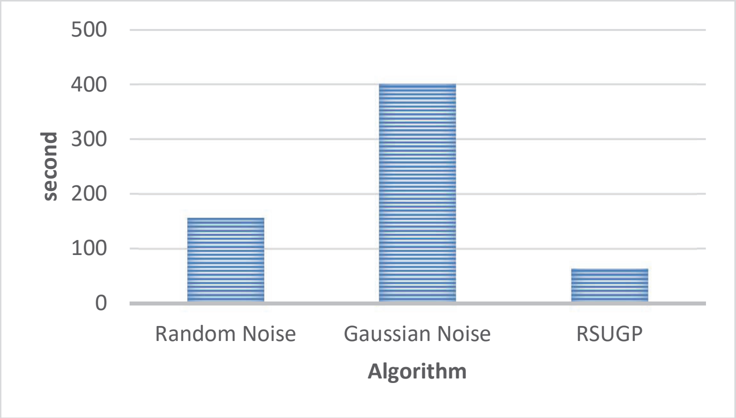


Figure 6 Average execution time.

4.5 Time

In this study, three different methods were run on six different size data sets. In comparison, the size of the data set and the selected number of sensitive attributes increases, the execution time of the methods increases accordingly.

Gaussian noise is defined as the statistical behaviour of random variables defined by the probability density function [49]. When the graph is examined, it is seen that the Gauss model, which produces noise according to the whole data, works longer than the others. The proposed model RSUGP, which produces noise based on random number generators, generates noise using only the selected sensitive attributes and iteration value to generate noise. In this way, it generates noise from the data and works in a much shorter time since it does not deal with the entire data. The average execution times of these methods, compared to each other, are shown on the graph in Figure 6.

5 Conclusion

In this study, we proposed a new geometric data perturbation method based on random number generators. For the performance evaluation of the proposed method, six different-sized data sets were used, and these data sets were compared with two different methods. The obtained values were compared with classification accuracy, attack resistance, and execution time, and for this comparison, the non-parametric statistical test method Friedman's test was used.

Our proposed method RSUGP with random number generators, gives better results classification accuracy and other criteria and is a suitable method for privacy protection for sensitive numerical data when the results are examined. Our future works will expand the dataset's volume and will use distributed data model to process big data.

Acknowledgment

This work is a part of the Ph.D. thesis titled "Development of data anonymization model based on random number generator for efficient privacy preserving" at the Institute of Graduate Studies, Istanbul University-Cerrahpasa, Istanbul, Turkey.

References

[1] Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165–1188.

[2] Internet Live Stats. Available at: https://www.internetlivestats.com/twitter-statistics/, [Accessed May. 28, 2021].

[3] Dodero, J. M., Rodriguez-Garcia, M., Ruiz-Rube, I., & Palomo-Duarte, M. (2019). Privacy-Preserving Reengineering of Model-View-Controller Application Architectures Using Linked Data. *Journal of Web Engineering*, *18*(7), 695–728.

[4] Canbay, Y., Vural, Y., & Sağiroğlu, Ş. (2020). Mahremiyet korumalı büyük veri yayınlama için kavramsal model önerileri. *Politeknik Dergisi*, *23*(3), 785–798.

[5] Zhang, X., Yang, L. T., Liu, C., & Chen, J. (2013). A scalable two-phase top-down specialization approach for data anonymization using MapReduce on cloud. *IEEE Transactions on Parallel and Distributed Systems*, *25*(2), 363–373.

[6] Ranjan, A., & Ranjan, P. (2016, April). Two-phase entropy-based approach to big data anonymization. In *2016 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 76–81). IEEE.

[7] Fung, B. C., Wang, K., Fu, A. W. C., & Philip, S. Y. (2010). Introduction to privacy-preserving data publishing: Concepts and techniques. CRC Press.

[8] Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *ACM Sigmod Record*, *33*(1), 50–57.

[9] Gai, K., Qiu, M., Zhao, H., & Xiong, J. (2016, June). Privacy-aware adaptive data encryption strategy of big data in cloud computing. In *2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud)* (pp. 273–278). IEEE.

[10] Sweeney, L. (1998). Datafly: A system for providing anonymity in medical data. In *Database Security XI* (pp. 356–381). Springer, Boston, MA.

[11] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), 3-es.

[12] Li, N., Li, T., & Venkatasubramanian, S. (2007, April). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering* (pp. 106–115). IEEE.

[13] Aldeen, Y. A. A. S., Salleh, M., & Razzaque, M. A. (2015). A comprehensive review on privacy preserving data mining. *SpringerPlus*, *4*(1), 1–36.

[14] Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., & Verykios, V. (1999, November). Disclosure limitation of sensitive rules. In *Proceedings 1999 Workshop on Knowledge and Data Engineering Exchange (KDEX'99)(Cat. No. PR00453)* (pp. 45–52). IEEE.

[15] Saygin, Y., Verykios, V. S., & Clifton, C. (2001). Using unknowns to prevent discovery of association rules. *ACM Sigmod Record*, *30*(4), 45–54.

[16] Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y., & Dasseni, E. (2004). Association rule hiding. *IEEE Transactions on knowledge and data engineering*, *16*(4), 434–447.

[17] Chamikara, M. A. P., Bertók, P., Liu, D., Camtepe, S., & Khalil, I. (2018). Efficient data perturbation for privacy preserving and accurate data stream mining. *Pervasive and Mobile Computing*, *48*, 1–19.

[18] Muralidhar, K., Parsa, R., & Sarathy, R. (1999). A general additive data perturbation method for database security. *management science*, *45*(10), 1399–1415.

[19] Agrawal, R., & Srikant, R. (2000, May). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 439–450).

[20] Tang, J., Korolova, A., Bai, X., Wang, X., & Wang, X. (2017). Privacy loss in apple's implementation of differential privacy on macos 10.12. *arXiv preprint arXiv:1709.02753*.

[21] McKenna, F. T. (1997). Object-oriented finite element programming: frameworks for analysis, algorithms and parallel computing. University of California, Berkeley.

[22] Fienberg, S. E., & McIntyre, J. (2004, June). Data swapping: Variations on a theme by dalenius and reiss. In *International Workshop on Privacy in Statistical Databases* (pp. 14–29). Springer, Berlin, Heidelberg.

[23] Hasan, A. T., Jiang, Q., Luo, J., Li, C., & Chen, L. (2016). An effective value swapping method for privacy preserving data publishing. *Security and Communication Networks*, *9*(16), 3219–3228.

[24] Estivill-Castro, V., & Brankovic, L. (1999, August). Data swapping: Balancing privacy against precision in mining for logic rules. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 389–398). Springer, Berlin, Heidelberg.

[25] Aggarwal, C. C., & Philip, S. Y. (2004, March). A condensation approach to privacy preserving data mining. In *International Conference on Extending Database Technology* (pp. 183–199). Springer, Berlin, Heidelberg.

[26] Chen, K., & Liu, L. (2005). A random rotation perturbation approach to privacy preserving data classification.

[27] Chen, K., & Liu, L. (2005, November). Privacy preserving data classification with rotation perturbation. In *Fifth IEEE International Conference on Data Mining (ICDM'05)* (pp. 4-pp). IEEE.

[28] Lin, Z., Wang, J., Liu, L., & Zhang, J. (2009, March). Generalized random rotation perturbation for vertically partitioned data sets. In *2009 IEEE Symposium on Computational Intelligence and Data Mining* (pp. 159–162). IEEE.

[29] Li, F., Zhang, R., Xu, Y., Liu, J., & Li, J. (2016, September). Privacy preservation based on rotation perturbation in weighted social networks. In *2016 16th International Symposium on Communications and Information Technologies (ISCIT)* (pp. 206–211). IEEE.

[30] Kadampur, M. A., & Somayajulu, D. V. (2010, September). Privacy preserving technique for Euclidean distance based mining algorithms using a wavelet related transform. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 202–209). Springer, Berlin, Heidelberg.

[31] Liu, K., Kargupta, H., & Ryan, J. (2005). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, *18*(1), 92–106.

[32] Chen, K., & Liu, L. (2011). Geometric data perturbation for privacy preserving outsourced data mining. *Knowledge and information systems*, *29*(3), 657–695.

[33] Sopaoğlu U. Privacy Preserving Anonymization of Big Data and Data Streams. PhD, TOBB University of Economics and Technology,2020

[34] Eyupoglu, C., Aydin, M. A., Zaim, A. H., & Sertbas, A. (2018). An efficient big data anonymization algorithm based on chaos and perturbation techniques. *Entropy*, *20*(5), 373.

[35] Chen, K., & Liu, L. (2009). Privacy-preserving multiparty collaborative mining with geometric data perturbation. *IEEE Transactions on Parallel and Distributed Systems*, *20*(12), 1764–1776.

[36] Balasubramaniam, S., & Kavitha, V. (2015). Geometric data perturbation-based personal health record transactions in cloud computing. *The Scientific World Journal*, *2015*.

[37] Reddy, V. S., & Rao, B. T. (2018). A combined clustering and geometric data perturbation approach for enriching privacy preservation of healthcare data in hybrid clouds. *International Journal of Intelligent Engineering and Systems*, *11*(1), 201–210.

[38] Darshna R., Avani J. (2015). Geometrıc Data Perturbatıon Usıng Clusterıng Algorıthm. *International Journal Of Advances In Cloud Computing And Computer Science (IJACCCS).* 1(1): 2454–4078.

[39] Dhiraj, S. S., Khan, A. M. A., Khan, W., & Challagalla, A. (2009, January). Privacy preservation in k-means clustering by cluster rotation. In *TENCON 2009-2009 IEEE Region 10 Conference* (pp. 1–7). IEEE.

[40] Javid, T., & Gupta, M. K. (2019, November). Privacy Preserving Classification using 4-Dimensional Rotation Transformation. In *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)* (pp. 279–284). IEEE.

[41] Sreekumar, K., & Baburaj, E. (2012). Privacy preservation using geometric data perturbation and fragmentation approach in wireless sensor networks.

[42] Oliveira, S. R., & Zaiane, O. R. (2010). Privacy preserving clustering by data transformation. *Journal of Information and Data Management*, *1*(1), 37–37.

[43] Chamikara, M. A. P., Bertók, P., Liu, D., Camtepe, S., & Khalil, I. (2019). An efficient and scalable privacy preserving algorithm for big data and data streams. *Computers & Security*, *87*, 101570.

[44] Yeliz GE. True Random Number Generation Based on Human Movements. *BEU Journal of Science.* 2019;8(1):261–9.

[45] Özkaynak F. Cryptographic Random Number Generators. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi*. 2015;8(2): 37–45.

[46] Okkalioglu, B. D., Okkalioglu, M., Koc, M., & Polat, H. (2015). A survey: deriving private information from perturbed data. *Artificial Intelligence Review*, *44*(4), 547–569.

[47] Friedman Test in SPSS Statistics. Available at: https://statistics.laerd.com/spss-tutorials/friedman-test-using-spss-statistics.php [Accessed May. 28, 2021].

[48] Küpeli C., Bulut F. Performance Analysis of Filters over Salt-Pepper and Gauss Noises in Images. *Haliç Üniversitesi Fen Bilimleri Dergisi.* 3(2):211–39. DOI: 10.46373/hafebid.768240

Biographies

Merve Kanmaz was born in İstanbul, Turkey. She received the B.S. and M.S. degrees in computer engineering from İstanbul University, İstanbul, in 2011 and 2016, respectively. She is currently pursuing the Ph.D. degree in computer engineering with İstanbul University-Cerrahpaşa, İstanbul. Since 2016, she has been working as a Lecturer at the Computer Programming Department, İstanbul University – Cerrahpaşa since 2014. Her research interests include data anonymization, information security, and big data. She has two journal articles and published and presented five international conference papers.

Muhammed Ali Aydin received the B.S. degree from İstanbul University, İstanbul, Turkey, in 2001, the M.Sc. degree from Istanbul Technical University, İstanbul, in 2005, and the Ph.D. degree from İstanbul University, in 2009, all in computer engineering. He was a Postdoctoral Research Associate with the Department of RST, Telecom SudParis, Paris, France, from 2010 to 2011. He has been working as an Associate Professor at the Computer Engineering Department, İstanbul University-Cerrahpaşa, since 2009. He has also been the Vice Dean of the Engineering Faculty and the Head of the Cyber Security Department, since 2016. He received ten research projects consisting of over Turkey from local industries in Turkey and the İstanbul University-Cerrahpaşa Research Foundation. He has authored 20 journal articles and published and presented 70 papers at international conferences. His research interests include cryptography, network security, information security, and optical networks.

Ahmet Sertbas was born in İstanbul, Turkey. He received the B.S. and M.S. degrees in electronic engineering from Istanbul Technical University, İstanbul, in 1986 and 1990, respectively, and the Ph.D. degree in electric-electronic engineering from İstanbul University, İstanbul, in 1997. Since 2000, he has been an Assistant Professor, an Associate Professor, and a

Professor with the Computer Engineering Department, İstanbul University, and a Professor with the Computer Engineering Department, İstanbul University-Cerrahpaşa, since 2018. His research interests include image processing, artificial intelligence, computer arithmetic, and hardware security. He has 19 articles in indexed SCI-SCIE journals, and many journal articles not indexed SCI-SCIE and international conference papers.