Mathematical Problems of Computer Science 41, 81--92, 2014.

Pair Correlations Preserving Model in Synthetic Data Generation

Vardan H. Topchyan

Institute for Informatics and Automation Problems of NAS RA
e-mail: vardan.topchyan@gmail.com

Abstract

The risk of disclosure of confidential information increases by the statistical organizations, due to the large volume of data released to the public. The most common methods of limiting the risk of dicloure are synthetic data genaretion methods. Unfortunately, these methods have a heuristic nature, because they do not have a clear theoretical basis. In this work presented a formal model of synthetic data generation for pair correlation preservation

Keywords: Synthetic data, **C**onfidentiality, Disclosure limitation.

1. Introduction

While providing state, economic or social data to public organizations it may be necessary to hide/protect the confidential component of the provided information. For this purpose, publishing organizations often resort to modification of initial data or their replacement by other synthetic data. New synthetic data [1] generated based on different models and algorithms; they should provide the analyzing organizations with the achievement of adequate conclusions. Synthetic data clearly distort the true picture of the data, but in addition, which is very important, they may cause distortion of links between different segments of the data sets, which, in turn, can lead to rougher erroneous conclusions at the data analysis stage. Therefore, in such cases, it is necessary together with the protection of personal information, also to ensure the safety of the same functional relations between the corresponding segments of data sets. The specified field intensively studied in the literature [2], [3] and [4]; there exist typical approaches and solutions. It often characterized as a problem of statistical disclosure limitation. This is because the data provided are mainly the object of statistical data analysis. Theoretically, as it is observed by some authors [1], the created tasks and methods of their solutions are similar to probabilistic problems with recovery of missing values [5], [6]. Analysis of the available methods for generating the synthetic data [7], [8], [9] indicates their heuristic nature. And this, in turn, means that the consistency of these methods substantiated by simulation methods and there is no theoretical validity of the use of a particular approach. In this paper we will make an attempt to formulate a formal model of the problem under consideration to identify and examine its very essence, in the case of saving of paired correlations, i.e. the study of structures of the input data

themselves, natural limitations imposed on them by the methods of data processing as well as by various types of privacy requirements.

2. Model Description

Consider themainmodel of the problem. Let U= $U_1, U_2, \dots, U_n$ be a set of certain elements from which the considered input data tasks *(information units set)*composed.Let a certain data element be characterized by a set of attributesA= $A_1, A_2, \dots, A_m$:

$$U_i = a_{i1}, a_{i2}, \dots, a_{im} , \quad 1 \le i \le n .$$

Without loss of generality, we assume that confidential information contains in first p columns of data table. Given that the order of attributes not fixed within the meaning of our problem, by rearranging them we can ensure that the confidential information were only in the first p columns,

$$A_{conf} = A_1, A_2, \dots, A_p , \ A_{conf} \subseteq \text{A}.$$

In principle, as the so-called “categorical” as well as continuous attributes are considered, but in our study as confidential attributes we will restrict ourselves only to the consideration of continuous attributes. Let the intervals $D_1, D_2, \dots, D_p$ of the real axis be the range of values of the attributes $A_1, A_2, \dots, A_p$, respectively. For these attributes introduce a set of threshold conditions which determines the degree of their confidentiality,

$$\text{C} = C_1, C_2, \dots, C_p .$$

The condition C_j $1 \le j \le d$ defines the critical area of attribute values A_j. For the consideration simplicity assume that the conditions C_j specify numerical intervals in the range of definition of the corresponding attribute A_j, although the consideration of other restriction structures may be quite natural and useful. Let C_j define the interval $(\bar{c}_j, \bar{\bar{c}}_j)$ of critical values in the range $(\bar{a}_j, \bar{\bar{a}}_j)$ of attribute $A_j$definition, $\bar{a}_j \le \bar{c}_j \le \bar{\bar{c}}_j \le \bar{\bar{a}}_j$.

Further, we assume that as additional information the setRis given, the elements of which represent (announce, declare) correlatedness (the quality of being correlated) between certain pairs of attributes of the set A,

$$\text{R} = R_1, R_2, \dots, R_t .$$

R_k $1 \le k \le t$ is a subset of A,$R_k \subset$A, which indicates the existence of correlation (or puts forward a demand of maintenance of the correlation form and degree) between the elements of this set of attributes.

U=

	A_1	...	A_p	...	A_m
U_1	a_{11}	...	a_{1p}	...	a_{1m}
U_2	a_{21}	...	a_{2p}	...	a_{2m}
⋮					
U_n	a_{n1}	...	a_{np}	...	a_{nm}

C = $C_1, C_2, \dots, C_p$

R = $R_1, R_2, \dots, R_t$

Fig. 1. The structure of the original data, critical intervals of values and system of correlativeness of the attributes.

Thus, our task is protection / concealment of confidential attributes A_{conf} critical values, that are determined based on the threshold conditions set C = $C_1, C_2, \dots, C_p$, with the condition of maintaining pairing correlations of attributes A based on system R= $R_1, R_2, \dots, R_t$.

3. Analysis of Critical Areas of Confidential Attributes in Data Table

When analyzing the critical areas of confidential attributes A_{conf} in the considered data table it is important to evaluate the predictability of values in these areas. So long as for some attribute $A_i \in A_{conf}$ $1 \le i \le p$ the number (volume) of different from each other critical values is relatively small, then during their imputations the reduction of disclosure risk of confidential information contained in this attribute will be insignificant. In this regard, it is expedient to evaluate the information entropy [10] of critical values for each set of attribute A_{conf}. The results of simulations indicate that for each confidential attribute with a value of not less than 7.5 - 8 entropy the generation of synthetic data categorical in terms of limiting the risk of disclosure of confidential information is possible. For further analysis of critical areas of the attributesA_{conf}introduce necessary definitions.

Definition 1: *The attribute A_j $1 \le j \le m$ is called single (isolated) if it is not correlated withone of the other attributes of the set A by the system of constraints R. The set of single attributes is denoted by A_{sng}, $A_{sng} \subseteq A$.*

Definition 2: *The attribute A_j $1 \le j \le m$ is called linked if it is correlated with at least one of the otherattributes of the setA. The set of linked attributes is denoted by A_{lnk}, $A_{lnk} \subseteq A$.*

It is easy to notice that the set of confidential attributes A_{conf} can be represented as a union of the corresponding subsets of single and linked attributes:

$$A_{conf} = A'_{sng} \cup A'_{lnk},$$

where A'_{sng} is the set of confidential single attributes, $A'_{sng} \subseteq A_{sng}$, and A'_{lnk} – the set of confidential linked attributes $A'_{lnk} \subseteq A_{lnk}$.

To analyze the permissible areas of modification/imputation of critical values of the attributes from A_{conf}, consider separately the sets A'_{sng} and A'_{lnk}. Without loss of generality, assume that the first l attributes of the set A_{conf} are single, $A'_{sng} = A_1, A_2, \dots, A_l$, and the rest $(p - l)$- are linked $A'_{lnk} = A_{l+1}, A_{l+2}, \dots A_p$.

Consider the set $A'_{sng} = A_1, A_2, \dots, A_l$. Changes in the critical values of the elements of the set A'_{sng}in the construction of synthetic data may be carried out independently from each other, as they do not correlate with any of the attributes of the setA. Let A_j $1 \le j \le l$ be acurrent attribute under consideration. We can assume that the critical values of this attribute are located in the upper part of the corresponding column; otherwise this representation can be obtained by consecutive relocations of certain rows of the table (Fig. 2). The above presented grouping may serve as a basis for consideration of changes in the critical values of the set A_j in two separate areas: (1) change of values in critical area of the column, (2) change of values in the whole column.

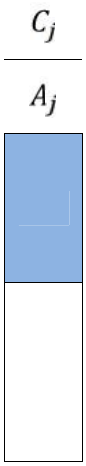

Fig. 2. Location scheme of single attribute values and areas of their confidential values.

Specific change in the attribute values will depend on the supposed procedures of data analysis. Consider a simple example of calculating the mean value of the attribute under consideration. If there are not other limitations, then ***relocations***can be considered in the whole area (2). Relocations do not change the objective values, and they change their distribution by individuals - rows. In our simpleexample, there isgreater scope forchange. One can just take another arbitrary column in (2) with the same mean value, or, - in the area (1) if there is a condition of preserving noncritical values. Saving the mean value is a weak condition and it does not appear separately in practice, so that real changes will maintain the character of the receivable values of the attribute under consideration. Thus, the concealment of critical values of the attributes $A'_{sng} = A_1, A_2, \dots A_l$ can be carried out independently of the rest of attributes of the set A, within the relevant field.

To analyze the attributes$A'_{lnk} = A_{l+1}, A_{l+2}, \dots A_p$, consider the set R= $R_1, R_2, \dots, R_t$. Further analysis will be based on the assumption that all attributes of the set A_{conf}presented in the system R and each of its elements contains at least one confidential attribute. Since, otherwise, if some element R_k $1 \le k \le t$ does not contain any confidential attribute, then its

consideration does not make sense, for the changes in the critical attribute values A_{conf}will nowise affect the connection represented by this element.

Let $(p - l) > 2$. Assume also that the attributes taking part in the definition of the pair correlation for $R_1, R_2, \dots, R_t$, contain intersection. Consider the subsets $R_{k_1}, R_{k_2} \in R$ represent the correlation between the attributes A_{j_1}, A_{j_2}иA_{j_2}, A_{j_3} $l < j_1 \neq j_2 \neq j_3 \leq m$, respectively,$R_{k_1} = A_{j_1}, A_{j_2}$, $R_{k_2} = A_{j_2}, A_{j_3}$. Assume also that the correlation between the attributes A_{j_1} and A_{j_3} is not additionally defined. In order to preserve communication over R_{k_1}, it is necessary that the change of the valuesA_{j_1} and A_{j_2} were agreed. Namely, the values of A_{j_1} should be changed in view of the appropriate attribute values A_{j_2}, and vice versa. Similar judgments hold for R_{k_2} and the attributes A_{i_2}, A_{i_3}. Obviously, the attribute A_{j_2} depends as on A_{j_1}, as well as on A_{j_3}. Therefore, to preserve the correlation byR_{k_1}, R_{k_2} the attribute values $A_{j_1}$and A_{j_3}should also be changed in agreement with each other. As a result, between the attributes $A_{j_1}$and A_{j_3} an interrelation arises subject to consideration of the attribute A_{j_2}. The foregoing data allow us to introduce the following natural definition.

Definition 3: *Let's say that the attributes A_{j_1} and $A_{j_v}$are conditionally correlated, subject to consideration of the attributes$A_{j_2}, A_{j_3}, \dots, A_{j_{v-1}}$, if there exists a set of paired correlations $R_{k_1}, \dots, R_{k_{v-1}}$ so that$R_{k_1} = A_{j_1}, A_{j_2}$, $R_{k_2} = A_{j_2}, A_{j_3}$, $\dots, R_{k_{v-1}} = A_{j_{v-1}}, A_{j_v}$.*

The conditional correlation of attributes A_{j_1}, A_{j_v} we denote by $R_{A_{j_1}, A_{j_2}, \dots, A_{j_v}} = A_{j_1}, A_{j_v}$.

The next stage of analysis of the set $A'_{lnk} = A_{l+1}, A_{l+2}, \dots A_p$ was the study of a binary relation between those attributes for which the communication preserved by the system R= $R_1, R_2, \dots, R_t$. Consider the set of these attributes, denoted by A_{corr}*(correlated).*

Definition 4: *We say that the attribute A_{j_1}enters into the binary relation $\alpha$with the attribute A_{j_2} , $A_{j_1} \alpha A_{j_2}$, if they meet one of the following conditions:*

- *Attributes $A_{j_1}$and A_{j_2}coincide: $A_{j_1} = A_{j_2} \Rightarrow A_{j_1} \alpha A_{j_2}$,*
- *Attributes $A_{j_1}, A_{j_2}$are correlated: $\exists R_{k_t} \in R_k \in R, R_{k_t} = A_{j_1}, A_{j_2} \Rightarrow A_{j_1} \alpha A_{j_2}$,*
- *AttributesA_{j_1}, $A_{j_2}$are conditionally correlated: $\exists A_{j_3}, \dots, A_{v_v} \in A_{corr}$, $R_{A_{j_3}, \dots, A_{j_v}} = A_{j_1}, A_{j_2} \Rightarrow A_{j_1} \alpha A_{j_2}$.*

It is obvious that α satisfies the properties of reflexivity and symmetry:

$$\forall A_{j_k} \in A_{corr} \Rightarrow A_{j_k} \alpha A_{j_k},$$

$$\forall A_{j_k}, A_{j_r} \in A_{corr}, A_{j_k} \alpha A_{j_r} \Rightarrow A_{j_r} \alpha A_{j_k}.$$

Let us show that this relation also satisfies the transitivity property, namely:

$$\forall A_{j_k}, A_{j_r}, A_{j_s} \in A_{corr}, A_{j_k} \alpha A_{j_r}, A_{j_r} \alpha A_{j_s} \Rightarrow A_{j_k} \alpha A_{j_s} .$$

Since $A_{j_k} \alpha A_{j_r}$, $A_{j_r} \alpha A_{j_s}$, then from the definition of the relationαit follows that between the attributes A_{j_k}, A_{j_r} and A_{j_r}, A_{j_s} there exists either a direct or a conditional correlation. Then by virtue of Definition 3 the attributes A_{j_k} and A_{j_s} will be conditionally correlated. And this, in turn, means that A_{j_k} enters into the relation α with the attribute A_{j_s}:$A_{j_k} \alpha A_{j_s}$.

Thus, the relation α satisfies the properties of reflexivity, symmetry and transitivity, hence it is an equivalence relation. In this caseα divides the set A_{corr} into disjoint equivalence classes:

$$A_{corr} = A^1_{corr} \cup A^2_{corr} \cup \ldots \cup A^d_{corr},$$

$$A^i_{corr} \cap A^j_{corr} = \emptyset,\ 1 \le i \ne j \le d.$$

Moreover, any two attributes of one and the same class are interconnected to each other, and between the attributes of various classes the correlation is missing.

Thus, in order to conceal the confidential information contained in the attributes of the set $A'_{lnk} = A_{l+1}, A_{l+2}, \ldots A_p$, on the assumption of preservation of the pair relations in the system R= $R_1, R_2, \ldots, R_t$, changes in critical values of the attributes $A_{l+1}, A_{l+2}, \ldots A_p$ may be carried out in the obtained equivalence classes $A^1_{corr}, A^2_{corr}, \ldots, A^d_{corr}$ separately.

For convenience, consider a particular case when an equivalence class $A^k_{corr} = A_{j_1}, A_{j_2}, A_{j_3}, \ldots, A_{j_r}$ $1 \le k \le d$ contains two confidential attributes $A_{j_1}, A_{j_2} \in A_{conf}$. Since A^k_{corr} does not have common attributes with other classes of equivalence, then for data analysis it is worthwhile to group the values of its attributes as shown in Figure 4.

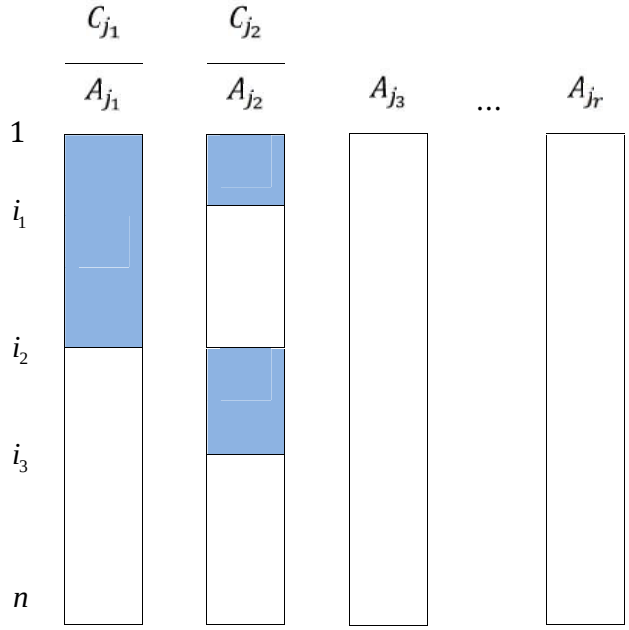

Fig. 4. Location scheme of linked attribute values and areas of their confidential values.

As shown in Figure 4, the critical values of the attribute A_{j_1} are situated in the interval of rows $1, i_2$ and at their concealment the appropriate values should be taken into account A_{j_2}. Moreover, the rows of the given area apart from the other values of the attribute A_{j_2}may contain also critical ones. Therefore, to ensure consistency during the imputations of the attribute valuesA_{j_1}, A_{j_2}, it is necessary that the rows in the interval $1, i_1$ should be considered separately from the rest of the rows of the interval $1, i_2$. In order to generate categorical synthetic data, the imputations of critical values in the interval $1, i_1$ should be carried out between the rows close / homogeneous by the attribute values $A_{j_1}$and A_{j_2}, on condition that the correlations between the elements of the class A^k_{corr} are preserved. Since, in our studies we restrict ourselves to the

consideration of only continuous attributes as confidential ones, then to group the rows of this range the technique of decomposing hierarchical cluster analysis can be applied [11]. In this case, the data elements are considered as r-dimensional vectors consisting of values of the class attributes A^k_{corr}, which enables partitioning based on correlations between the attributes of this class. As a measure of distance between the data elements the Euclidean distance is considered, and as a measure of homogeneity of the obtained subsets - the measure of RMSSTD (Root-Mean-Square Standard Deviation) [12], which is equal to the mean square deviation of critical values of confidential attributes. In addition, it is more appropriate to carry out imputations by collection of these attributes instead of successive imputations for each of them. Due to this, the correlation between the attributes A_{j_1}, A_{j_2} will be preserved in the best possible way.

Then, as for the other critical values of the attribute A_{j_1}, they may be considered together with the values from the interval $i_2 + 1, n$ or without them, depending on the restrictions imposed on noncritical values of confidential attributes. In any case, to preserve the correlations between the elements of the class A^k_{corr} the imputations of these values should be made taking into account the conditional distribution of A_{j_1} for the rest of the class attributes A^k_{corr}. According to the literature, [4], one of the most appropriate methods for determining the conditional distribution in the generation of synthetic data are CART trees (Classification and Regression Trees) [13]. CART trees are used to predict the values of the dependent variable based on a set of predictors. In this case, as a dependent variable is considered the attribute A_{j_1}, and as predictors - the rest $r - 1$ attributes of the class A^k_{corr}. The principle of constructing the CART trees consists in a recursive partition of the set of data elements under consideration into subsets that are homogeneous with respect to the dependent variable. Namely, at each step the best condition is determined for some predictor and a partition of the current set is produced (by growing it). As a result, in the leaves of the obtained tree the data elements will be contained with the same value of the dependent variable. Since the obtained tree may consist of unjustifiably large number of nodes and branches, then to reach an acceptable size of these trees their pruning is made on the basis of some optimality criterion. In essence, the leaves of the CART tree represent a conditional distribution of the dependent variable for the set of predictors under consideration. Subsequent imputations of the attribute values A_{j_1} will consistently be carried out in each leaf. Similar statements hold also for the critical values of the attribute A_{j_2} in the interval $i_2 + 1, i_3$.

Thus, from the above simple example one may conclude that if the equivalence class A^k_{corr} consists of m_k attributes, $A^k_{corr} = A_{j_1}, A_{j_2}, A_{j_3}, \dots, A_{j_{m_k}}$, the first s of which are confidential, then the rows of the data table are divided into not more than 2^s separate areas, each of which contains a specific combination of critical values of confidential attributes and is processed by one of the methods presented above. These model structures and analysis confirmed by computational experiments presented in the next section.

4. Simulation Studies

The presented model has been approved based on the data of 2012 Household's Integrated Living Conditions Survey, provided by the "National Statistical Service" of the Republic of Armenia. From the entire set of attributes characterizing these data, we are interested only in the following six: *FoodPurchased, FoodConsumed, NonFoodPurchased, Expenditure, MonitoryIncome, TotalIncome* (Table 1).

Table 1: Description of attributes of interest

Name	*Label*	*Type*	*Description*
FoodPurchased	FP	Numeric(18,10)	Foodpurchased of household per month in AMD.
FoodConsumed	FC	Numeric(18,10)	Foodconsumed of household per month in AMD.
NonfoodPurchased	NFP	Numeric(18,10)	Nonfood purchased of household per month in AMD.
Expenditure	E	Numeric(18,10)	Expenditures of household permonth in AMD.
MonitoryIncome	M	Numeric(18,10)	Monetary income of household per month in AMD.
TotalIncome	I	Numeric(18,10)	Total income of household per month in AMD.

In our experiments we assume that the confidential information is contained in the attributes *Expenditure* and *TotalIncome*, $A_{cndf} = E, I$, and as threshold conditions are considered$E > 200000$ and$I > 250000$, respectively. As for pair correlations, which should be saved, they are as follows: $R_1 = I, M$, $R_2 = I, E$, $R_3 = E, NFP$ и$R_4 = E, FP$, i.e. R= R_1, R_2, R_3, R_4 . It is obvious that in this case, when generating the synthetic data, only one equivalence class $A_{corr}^1 = I, M, E, FP, NFP$ will be considered. In addition, the imputations of critical attribute values$E$and I are implemented due to the methods of ***relocation*** and ***reevaluation*** of values. First of all relocation of values is carried out using the method of Bayesian bootstrapping [14]. After that if some values remain unchanged, then their reevaluation is made: (i) probabilistic density of these values is determined using the Gaussian kernel density estimator; (ii) new values are set using the inverse-cdf method. Finally, in the result of experiment,$m = 5$ sets of synthetic data are generated and as resulting values of statistical quantities, their average values are taken calculated on these sets.

Table 2: Mean and standard deviation of confidential attributes

	Mean		Standard deviation	
	I	E	I	E
Estimated value on original data set	132862.38	109818.56	105518.35	95060.952
Average of estimated values on synthetic data sets	132523.88	108960.52	100335.54	78792.5834

Table 3: Correlation coefficients

	Correlations coefficient			
	I, M	I, E	E, NFP	E, FP
Estimated value on original data	0.974	0.414	0.708	0.579
Avg. of estimated values on synthetic data	0.880	0.511	0.845	0.673

Table 4: Coefficient in regression of total income on monitory income, food purchase, expenditures

	Value on original data set	Avg. of values on synthetic data sets
Coefficients		
Constant	11279.38	17607.02
M	0.966	0.800
FP	-0.268	-0.297
E	0.165	0.303
R	0.984	0.900

Table 5: Coefficient in regression of total income on monitory income and expenditures

	Value on original data set	Avg. of values on synthetic data sets
Coefficients		
Constant	8868.75	13756.59
M	0.964	0.800
E	0.072	0.214
R	0.980	0.894

Table 6: Coefficient in regression of expenditure on total income, food purchased, food consumed, nonfood purchased.

	Value on original data set	Avg. of values on synthetic data sets
Coefficients		
Constant	-1664.49	1813.41
I	0.026	0.041
FP	1.017	0.919
FC	0.992	0.9262
NFP	1.089	1.110
R	0.984	0.957

Table 7: Coefficient in regression of expenditure on food purchased, food consumed, nonfood purchased

	Value on original data set	Avg. of values on synthetic data sets
Coefficients		
Constant	-168.794	3692.55
FP	1.033	0.943
FC	1.016	0.962
NFP	1.108	1.140
R	0.983	0.957

The above presented tables contain the results of experiments conducted. As shown in Tables 2 and 3, the mean values of simple statistical quantities (mathematical expectation, mean square deviation) of confidential attributes and the coefficients of the corresponding pair correlations calculated on the sets of synthetic data, are close to the original ones. And this, in turn, indicates that the numerical characteristics of the attributes *Totalincome*and Expenditure, as well as the primary relations are saved also in the synthetic data. Then, in Tables 4 - 7 the coefficients of linear regressions constructed on the original and synthetic data sets are shown. On the sets of synthetic data the values of the parameter R (indicating the degree of correctness of interpretation of a dependent variable from the independent ones) are in the vicinity of 0.9 and more, indicating that they are correct. In addition, the corresponding values of the regression coefficients are close to the original. Consequently, conclusions drawn from the analysis of synthetic data will correctly reflect the results of analysis of the original data.

5. Conclusion

Problem of maintaining confidentiality of state, economic and social data in distributed computing related to new theoretical and applied research. As of today, the existing algorithms of their analysis are of a heuristic nature. In the present work the structure of these data was studied and a model was presented limiting the risk of disclosure of confidential information with preservation of paired connections between various segments of data.

References

[1] D. B. Rubin, “Discussion: statistical disclosure limitation”, *Journal of Official Statistics*, vol. 9, pp. 462–468, 1993.

[2] T. E. Raghunathan, J. P. Reiter and D. B. Rubin, “Multiple imputation for statistical disclosure limitation”, *Journal of Official Statistics*, vol. 19, pp. 1–16, 2003.

[3] J. Drechsler, *Synthetic Datasets for Statistical Disclosure Control. Theory and Implementation*, Springer, 2011.

[4] J. Drechsler and J. P. Reiter, “An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets”, *Computational Statistics & Data Analysis*, vol. 55, no. 12, pp. 3232--3243, 2011.

[5] T. E. Raghunathan, J. M. Lepkowski, J. van Hoewyk and P. Solenberger, ”A multivariate technique for multiply imputing missing values using a series of regression models”, *Survey Methodology*, vol. 27, pp. 85--96, 2001.
[6] D. B. Rubin, *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons, 1987.
[7] J. P. Reiter, “Using CART to generate partially synthetic, public use microdata”, *Journal of Official Statistics*, vol. 21, pp. 441-462, 2005.
[8] G. Caiola and J. P. Reiter, “Random forests for generating partially synthetic, categorical data”, *Transactions on Data Privacy*, vol. 3, pp. 27--42, 2010.
[9] J. Domingo-Ferrer, J. Magkos, (eds.), *Privacy in Statistical Databases*, New York: Springer, pp. 148--161, 2010.
[10] В. Лидовский, *Теория Информации*, Москва, Спутник+, 2004
[11] L. Aslanyan and V. Topchyan, “Hierarchical cluster analysis for partially synthetic data generation”, *Transactions of IIAP NAS RA, Mathematical Problems of Computer Science*, vol. 40, pp. 55--67, 2013.
[12] M. Halkidi, Y. Baristakis and M. Vazirgiannis, “On clustering validation techniques”, *Journal of Intelligent Information Systems*, vol. 17, no. 2-3, pp. 107--145, 2001.
[13] L. Breiman, J. H. Friedman, R. A. Olshen and C.J. Stone, *Classification and Regression Trees*, Belmont, CA: Wadsworh, Inc., 1984.
[14] D. B. Rubin, “The Bayesian bootstrap”, *The Annals of Statistics*, vol. 9, pp. 130–134, 1981.

Submitted 20.12.2013, accepted 05.03.2014.

Զույգ կորելյացիաների պահպանման մոդելը սինթետիկ տվյալների գեներացման միջոցով

Վ. Թոփչյան

Ամփոփում

Կոնֆիդենցիալ տեղեկատվության բացահայտման ռիսկը մեծանում է ի շնորհիվ վիճակագրական կազմակերպությունների կողմից հանրությանը տրամադրվող մեծ քանակությամբ տվյալների: Այս խնդիրի լուծման ամենատարածված մեթոդներից են սինթետիկ տվյալների գեներացումը: Ցավոք, այդ մեթոդներն ունեն էվրիստիկ բնույթ, քանի որ նրանք չունեն հստակ տեսական հիմնավորում: Այս աշխատանքում ներկայացված է սինթետիկ տվյալների գեներացման ֆորմալ մոդելը, որն ապահովում է զույգ կորելյացիաների պահպանությունը:

Модель сохранения парных корреляций при генерации синтетических данных

В. Топчян

Аннотация

Риск раскрытия конфиденциальной информации увеличивается в связи с большим объемом данных, предоставляющимися статистическими организациям и общественности. Наиболее распространенными методами для решения данной проблемы являются методы генерации синтетических данных. К сожалению эти методы имеют эвристический характер, потому что они не имеют четкой теоретической основы. В этой работе представлена формальную модель генерации синтетических данных, обеспечивающих сохранение парных корреляций.