10. Balageas, D. Structural Health Monitoring [Text] / D. Balageas, C.-P. Fritzen, A. Gemes. – John Wiley & Sons, 2006. – 495 p. doi: 10.1002/9780470612071
11. Dworakowski, Z. Application of Artificial Neural Networks for Damage Indices Classification with the Use of Lamb Waves for the Aerospace Structures [Text] / Z. Dworakowski, L. Ambrozinski, P. Packo, K. Dragan, T. Stepinski, T. Uhl // Key Engineering Materials. – 2013. – Vol. 588. – P. 12–21. doi: 10.4028/www.scientific.net/kem.588.12
12. Shen, T. Damage location and identification of the wing structure with Probabilistic Neural Networks [Text] / T. Shen, F. Wan, B. Song, Y. Wu // 2011 Prognostics and System Health Managment Confernece. – 2011. doi: 10.1109/phm.2011.5939524
13. Palomino, L. V. Probabilistic Neural Network and Fuzzy Cluster Analysis Methods Applied to Impedance-Based SHM for Damage Classification [Text] / L. V. Palomino, V. Steffen, R. M. Finzi Neto // Shock and Vibration. – 2014. – Vol. 2014. – P. 1–12. doi: 10.1155/2014/401942
14. Bouraou, N. I. Syntez neironnoi merezhi dlia bahatoklasovoi diahnostyky elementiv konstruktsiy v ekspluatatsiy [Text] / N. I. Bouraou, A. H. Protasov, P. S. Myronenko, S. S. Rupich // Metody ta prylady kontroliu yakosti. – 2015. – Issue 2 (35). – P. 83–93.

Запропоновано модифікацію методу розв'язання задачі забезпечення групової анонімності на основі міметичного алгоритму, яка не передбачає участі експерта на етапі оцінювання розв'язків задачі. Автоматизація оцінювання розв'язків підвищує ефективність процесу групової анонімізації даних. Модифікацію методу проілюстровано шляхом розв'язання задачі анонімізації на основі реальних даних

Ключові слова: міметичний алгоритм, групова анонімність, мікрофайл, викид, модифікований метод тау Томпсона

Предложена модификация метода решения задачи обеспечения групповой анонимности на основе меметического алгоритма, которая не предусматривает участия эксперта на этапе оценивания решений задачи. Автоматизация оценивания решений повышает эффективность процесса групповой анонимизации данных. Модификация метода проиллюстрирована путем решения задачи анонимизации на основе реальных данных

Ключевые слова: меметический алгоритм, групповая анонимность, микрофайл, выброс, модифицированный метод тау Томпсона

UDC 004.62:004.023
DOI: 10.15587/1729-4061.2017.113046

IMPROVING EFFICIENCY OF PROVIDING DATA GROUP ANONYMITY BY AUTOMATING DATA MODIFICATION QUALITY EVALUATION

O. Chertov
Doctor of Technical Sciences, Associate Professor*
E-mail: chertov@i.ua
D. Tavrov
PhD*
E-mail: dan.tavrov@i.ua
*Department of Applied Mathematics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Peremohy ave., 37, Kyiv, Ukraine, 03056

1. Introduction

Around the world, amounts of digital data keep increasing with each year. A great share of these data are published in open access in their primary, non-aggregated form. Such data sets are called *microdata*. Microdata can be used for numerous purposes, including:

– dissemination of clinical data to facilitate medical research. E. g., in the U.S., this is regulated by the corresponding bills [1, 2];

– enforcing transparency of public policy. E. g., in the EU, protection of personal data is subject to the corresponding law [3];

– sharing census and other statistical research data to enable conducting economic, demographic, and other kinds of research.

At the same time, there is a certain risk that providing public access to the data in their unchanged form will not only achieve its primary goal but also lead to disclosing confidential information about an individual or a group thereof. E.g., open access to clinical data facilitates medical research. At the same time, publishing medical records can enable unique identification of a patient. Moreover, outliers in a regional distribution of patients might point to areas with exceeded sickness rate threshold.

Therefore, it is important to provide data anonymity at the stage of creating the content for open information resources. Anonymity of a subject can be seen as its property of being not identifiable (uniquely characterized) within a set of subjects [4]. Anonymity comes in two variants:

– *individual anonymity* concerns information about single respondents (persons, households, enterprises);

– *group anonymity* concerns distribution of information about a group of respondents.

Methods for providing individual anonymity have been a subject of research for more than 20 years and are developed

to a sufficient level. At the same time, methods for providing group anonymity began their development only in 2009. Methods described in the literature are population in their nature, i.e. they generate huge numbers of potential ways to provide anonymity. Not all of these ways solve the task of providing anonymity, therefore in practice, they need to be analyzed for satisfying imposed requirements. This process is time-consuming, so the research aimed at developing an automated method for evaluating ways of providing group anonymity is a topical one.

2. Literature review and problem statement

Primary non-aggregated data are typically being disseminated as *microfiles*, i. e. data tables, in which rows (*records*) correspond to respondents, and columns correspond to their *attributes*. Out of all microfile attributes, we can single out a set of *vital attributes*, whose values can be used to uniquely determine whether a respondent belongs to a certain *group* (subset of microfile records). Records that belong to the group are called *vital records*. In addition, we can select a *parameter attribute*, whose values determine a distribution of information about vital records. The microfile can be split into *parameter submicrofiles*, each of which contains records with the same parameter value.

Using the microfile data, we can build a *goal signal*, which shows the distribution of data about a given group along the values of the parameter attribute. In the literature, several kinds of goal signals are defined, i. e. quantity, concentration, and concentration difference signals. It is sufficient to discuss only quantity signal, elements of which are numbers of vital records in corresponding submicrofiles. Adaptation of the method described in this work to signals of other types is straightforward.

In practice, group anonymity can be violated by analyzing *outliers* in the quantity signal. Outliers are [5] values that differ from other signal values so much that they arouse suspicions that they were generated by a different mechanism. If outliers can be detected (either by visual inspection or with the help of an automatic procedure), it is said that there is a risk of violating group anonymity. In this paper, we will discuss the cases when the threat for anonymity violation comes from the outliers that *exceed* all other signal values. Such outliers can be interpreted as anomalous numbers of respondents belonging to the group in one parameter submicrofile.

To provide group anonymity, one needs to apply data modification in order to mask goal signal outliers, which satisfies the following conditions [6]:

1. After modification, the risk of anonymity violation is mitigated.

2. Results of analysis of the modified data are close to results of analysis of the primary data.

3. Data modification is cost-effective (in terms of time and other resources).

The simplest way to mask outliers is to remove the vital attribute from the microfile. E. g., if "Place of Work" is selected as a parameter attribute, and "Military Service" is selected as a vital one, then removing the latter one will hypothetically disable detecting outliers that correspond to sites of military bases. However, as was shown before [7], this approach satisfies only the third condition. The first two conditions will not be satisfied in general, because it is sometimes possible to build a model of a group that takes into consideration values of other core microfile attributes (such as "Age," "Sex," etc.), removing which is unacceptable. Using such a model, we can build a special type of distribution whose outliers match the ones in the quantity signal, thereby violating group anonymity. Thus, regardless of the decision to remove the vital attribute from the microfile, additional data modifications are necessary to provide group anonymity.

Detecting outliers is also important in the context of individual anonymity, because there exists a high risk of identifying an outlying record. To mask individual outliers, methods of providing k-anonymity, cell suppression, and generalization are used.

The idea of providing k-anonymity was introduced in [8]. It implies modifying data in such a way that combinations of microfile attribute values correspond to at least k microfile records. Typically, the corresponding modification leads to introducing certain statistical bias in the microfile attribute values distribution. In [9], enhanced methods of data generalization and suppression are proposed, which enable us to reduce the corresponding bias whilst providing k-anonymity. Methods of providing k-anonymity are implemented in the sdcMicro package for the R system [10].

In general, k-anonymity requirement is a weak one, so to increase data protection level, the concept of l-diversity was proposed [11]. According to this concept, each value of the sensitive attribute has to correspond to at least l microfile records. A development of this idea is the concept of t-closeness, whose basic tenet is that [12] the distance between the distribution of sensitive attribute values in a certain equivalence class inside the microfile and the distribution of this attribute values in the microfile as a whole doesn't exceed a certain threshold t.

Another group of methods for removing outliers from the microfile includes suppression and generalization methods. They imply [13] replacing certain values, which enable us to uniquely identify some records, with more general values, including interval ones.

In recent years, additional methods for providing individual anonymity have been proposed. In [14], an algorithm is described for disabling detecting classification rules, which can lead to leaking sensitive information from the microfile. An overview of methods for providing anonymity in social networks, including masking outliers, is given in [15].

All the discussed methods for removing outliers are not applicable to group anonymization, because they work at the level of single records, not quantity signals. Therefore, the task of developing a method for providing group anonymity with the automated procedure of detecting quantity signal's outliers remains unsolved in the literature.

3. The aim and objectives of the study

The aim of this work is to improve the efficiency of the data group anonymization process by developing an automated method for providing data group anonymity, which does not require expert's participation in evaluating solutions of the TPGA. In such a method, the expert's role boils down to choosing parameters of the problem at the outset. The rest is left up to the automatic procedure.

To achieve this aim, the following objectives were accomplished:

– to modify the memetic algorithm based method for solving the TPGA by automating the process of evaluating its solutions;

– to conduct a real data based experiment to validate the modified method in practice.

4. Materials and methods of researching the influence of the group anonymization modification on efficiency

4. 1. Formal definition of the task of providing group anonymity

Let us denote the microfile, in which it is needed to provide group anonymity, by **M**, its records by $\mathbf{r}^{(i)}$, i=1,..., ρ, and its attributes by $\mathbf{w}_j$, j=1,..., η. We will assume that the number of parameter values equals l_p. Let us denote parameter submicrofiles by $\mathbf{M}_1$,..., $\mathbf{M}_{lp}$. Let us denote the number of records in the ith submicrofile by ρ_i. We will denote the quantity signal by $\mathbf{q}=(q_1, q_2,..., q_{lp})$, where q_k is the number of vital values in $\mathbf{M}_k$. A set of indexes of **q**, which correspond to outliers, will be denoted by *OUT*(**q**).

The *task of providing group anonymity* (TPGA) is formulated as follows. Such data modification needs to be selected that masks outliers in **q** built based on the modified microfile **M*** (*modified quantity signal* **q***), but doesn't reduce data utility much (in terms of a given utility measure). At the same time, this modification should introduce as little distortion into the microdata as possible:

– to minimize data distortion at the level of single respondents, it is necessary to perform pairwise swapping of respondents between parameter submicrofiles (and changing their parameter values accordingly);

– to minimize data distortion at the level of the whole microfile, the respondents in each pair must be *similar* in some sense.

In the literature, a similarity measure widely used for this purpose is known as [16] *influential metric*:

$$\mathrm{InfM}(\mathbf{r}^{(i)},\mathbf{r}^{(j)}) = \sum_{k=1}^{n_{\mathrm{ord}}} \omega_k \left(\frac{r_{I_k}^{(i)} - r_{I_k}^{(j)}}{r_{I_k}^{(i)} + r_{I_k}^{(j)}}\right)^2 + \sum_{l=1}^{n_{\mathrm{cat}}} \gamma_l \chi^2\left(r_{J_l}^{(i)}, r_{J_l}^{(j)}\right), \quad (1)$$

where I_k (J_l) is the kth ordinal (lth categorical) *influential attribute* (attribute whose distribution is important for potential users of the microfile); $\chi(v_1, v_2)$ equals a certain number χ_1 if v_1 and v_2 fall into one category, and χ_2 otherwise; ω_k and γ_l are non-negative weights (the more important the attribute, the higher the weight).

As was shown in [16], the TPGA can be reduced to a well-known minimum cost network flow problem [17]. The architecture of the network in this task is uniquely determined by the choice of vital and parameter values, coefficients in (1), and the modified quantity signal **q***. This problem can be solved using algorithms of polynomical complexity, yielding a solution that can be interpreted as the level of distortion that must be introduced in the microdata in order to achieve the needed **q***.

Since different modified quantity signals define different networks, in practice, we face a *meta* problem of choosing such a modified quantity signal that the outliers are masked, and solution to the corresponding minimum cost flow problem corresponds to the minimal distortion introduced. Because we cannot specify exact values in **q*** beforehand that will lead to the smallest distortion, we can only impose certain *restrictions* on the values of the quantity signal **q** with indexes from *OUT*(**q**). Each such restriction is a function $\mu_i(x)$ defined for the ith value of **q** that possesses the following properties:

– equals 0 for $x \geq q_j$ (an outlier cannot grow);

– equals 1 for $x \leq \varepsilon_j$, where ε_j is a *threshold value*, below which the ith value of **q** should fall in the best case scenario. This value is set by the expert before solving the problem;

– monotonically falls to 0 when $\varepsilon_j \leq x \leq q_j$.

The value $\mu_i(q_i)$ is called *compatibility* of q_i with the restriction. The compatibility $\mu(\mathbf{q})$ of the whole signal **q** with a set of restrictions is defined as the *product* of individual compatibilities.

Threshold values can be picked according to the procedure described in [18]. Let us denote by $\mathbf{q}^{\boldsymbol{K\mathbf{max}}}$ the Kth biggest value of the subsignal $\mathbf{q}'=(q_j)$, j belongs to *OUT'*(**q**), i. e. a complement of *OUT*(**q**) to the index set {1,..., l_p}. The condition $\varepsilon_j = \mathbf{q}^{\boldsymbol{K\mathbf{max}}}$ requires that when the compatibility of the modified signal with the restrictions is high, the outliers fall below the level that does not exceed the Kth biggest value in the modified signal. Sometimes in practice, the threshold values can be selected as low as $\varepsilon_j = \mathbf{q}^{\boldsymbol{K\mathbf{max}}} - (q_j - \mathbf{q}^{\boldsymbol{K\mathbf{max}}}) \cdot 0{,}2$, to ensure the satisfying result.

Since data anonymization problems typically involve Big Data, solving TPGA for optimal modification is not warranted. Often, feasible but suboptimal solutions can provide anonymity and introduce sufficiently small data distortion.

Let us denote a feasible solution to the TPGA by

$$\mathbf{S} = \left(\left(\mathbf{r}^{(i_1)}, \mathbf{r}^{(j_1)}\right), ..., \left(\mathbf{r}^{(i_Q)}, \mathbf{r}^{(j_Q)}\right)\right),$$

where i_k, j_k, k=1,..., Q, are indexes of those microfile records that need to be swapped between submicrofiles. Then, the TPGA can be formally stated as follows: find such an ordered sequence of pairwise record swaps **S** that satisfies the following conditions:

$$\mu\left(q_1^*(\mathbf{S}), \ldots, q_{l_p}^*(\mathbf{S})\right) \geq \alpha_{comp},$$

$$\frac{\left|OUT(\mathbf{q}) \cap OUT\left(\mathbf{q}^*(\mathbf{S})\right)\right|}{\left|OUT(\mathbf{q})\right|} \leq K_{out},$$

$$\sum_{k=1}^{Q} \mathrm{InfM}\left(\mathbf{r}^{(i_k)}, \mathbf{r}^{(j_k)}\right) \leq K_{dist} \cdot C_{\max}, \quad (2)$$

where $\mathbf{q}^*(\mathbf{S})$ is the modified quantity signal built after swapping records from **S**; $\mu(q_1^*(\mathbf{S}), ..., q_{lp}^*(\mathbf{S}))$ is the compatibility of **q*** with the restrictions; α_{comp} is called a *compatibility threshold* (typically, $\alpha_{comp} \geq 0.5$); K_{out} is called a *sensitivity threshold*; K_{dist} is called a *distortion threshold*; $C_{\max}$ is the maximal possible cumulative value of (1) that can be attained for the current TPGA.

In [19], a method for solving the TPGA is proposed, which is based on memetic algorithms. Memetic algorithms are a dialect of evolutionary algorithms, coupled with local search techniques [20]. In many cases, using the local search procedure enhances [21] the algorithm's efficiency by incorporating specific knowledge about the task to be solved. In recent years, many novel applications of memetic algorithms have been proposed, especially in the area of

solving complex optimization problems [22]. Such algorithms aim at solving the shortest path routing problem [23], minimum dominating set problem [24], minimum graph cutwidth problem [25], etc.

Thus, in the general case, the method for solving TPGA consists of the following steps:

1. Determine parameter and vital attributes. Build a goal signal. Visually detect outliers in this signal with the help of an expert.

2. Formulate, with the help of the expert, restrictions on those values of the goal signal that correspond to the detected outliers. Determine the measure of similarity of respondents in a microfile.

3. Apply the memetic algorithm for solving the TPGA.

4. Visually analyze, with the help of the expert, the solutions obtained by the memetic algorithm (in order to confirm that the outliers have been indeed masked).

The last step of this method is necessary because evolutionary algorithms are known to be very effective at reaching the vicinity of the minimum of the objective function, but are quite ineffective at converging to this minimum. Due to the population nature of memetic algorithms, a large number (hundreds and even thousands) of candidate solutions to TPGA can be obtained, most of which may in fact be far from satisfying the desirable property of masking the outliers. Therefore, each of the solutions needs to be analyzed by the expert separately, which is very time-consuming, and therefore violates the third condition imposed on the data modification process.

4. 2. Memetic algorithm for solving the TPGA

Let us discuss the memetic algorithm for finding sequences of swaps that satisfy (2), described in [16]. In this algorithm, the population consists of matrixes U of dimension $Q\times 4$, where each row uniquely defines a pair of records to be swapped:

– element of the first column u_{i1}, i=1,..., Q, is the index of the submicrofile, from which it is necessary to remove a record;

– element of the second column u_{i2}, i=1,..., Q, is the index of the record in the corresponding submicrofile that needs to be removed;

– element of the third column u_{i3}, i=1,..., Q, is the index of the submicrofile, to which it is necessary to add a record;

– element of the fourth column u_{i4}, i=1,..., Q, is the index of the record in the corresponding submicrofile that needs to be swapped with the record defined by the first two columns.

Each submicrofile index i, i=1,..., l_p, can be present in the first column no more than q_i times. In the third column, index i can be present no more than $(\rho_i - q_i)$ times. Each microfile record can be present in U only once. Since rows of U are ordered, each individual in the population uniquely determines a candidate solution to the TPGA.

Fitness of an individual in the population is determined by the fitness function

$$f(U)=\Upsilon(U)\cdot\Phi(U)\cdot\Psi(U), \quad (3)$$

where $\Upsilon(U)$ is the measure of TPGA solution quality from the distortion minimization point of view; $\Phi(U)$ is the measure of TPGA solution quality from the masking outliers point of view (i.e. compatibility $\mu(\mathbf{q})$ of the signal with the restrictions); $\Psi(U)$ is the penalty against unbounded growth of a number of rows in individuals. Values of each factor in (3) must lie in [0, 1], because they are equally important for the overall solution quality.

The memetic algorithm proceeds along the following steps:

1. Randomly generate a population $P=\{U_i\}$ of μ individuals, i=1,..., μ.

2. Apply local search operator $S(U_i)$, i=1,..., μ.

3. Calculate fitness values (3) for each individual.

4. If the termination condition holds, stop the algorithm.

5. Select λ parent pairs and place them in set P'.

6. Apply recombination operator $R(U_{i1}, U_{i2})$ to each individual pair U_{i1}, U_{i2} from P'. Place the offspring in set P''.

7. Apply mutation operator $M(U_i)=(M_4\circ M_3\circ M_2\circ M_1)(U_j)$ $\forall U_j$, where each operator M_k, k=1, ..., 4, is applied separately to the kth column of U_j.

8. Apply $S(U_i)$ to each individual from P''.

9. Calculate fitness values (3) for each individual from P''.

10. Select μ fittest individuals from the union of P and P'' and place them in P, overwriting current individuals.

11. Go to step 3.

The first population should be initialized by randomly generating individuals with different numbers of rows. Elements of the first column in the individuals should be generated with probabilities proportional to corresponding elements of $\mathbf{q}$. Elements of the third column should be generated with probabilities proportional to sizes of corresponding submicrofiles.

Termination criterion can be chosen to be the number of generations elapsed. Choice of other algorithm parameters (μ, λ, recombination probability, mutation probability, selection method, etc.) depends on the task being solved.

4. 3. Modification of the method that doesn't involve the expert in evaluating solution quality

In order to improve data anonymization efficiency, in this work, we propose the modified method of solving the TPGA that doesn't involve the expert at the last stage. The modified method consists of the following steps:

1. Determine parameter and vital attributes. Build a goal signal. Visually *or automatically* detect outliers in this signal.

2. Formulate, with the help of the expert, restrictions on those values of the goal signal that correspond to the detected outliers. Determine parameters of the task being solved (coefficients in (1) and (2)).

3. Apply the memetic algorithm for solving the TPGA.

4. *Automatically select* solutions that satisfy (2).

To automatically select solutions that satisfy (2), we need to apply some method of automatic outlier detection in the modified goal signal. All the methods for outlier detection assume building some data model [26], deviations from which imply the presence of outliers. The simplest such model can be based on the assumption that signal values are normally distributed. However, in TPGA, goal signals typically contain at most several dozens of elements, so the assumption of the Student distribution of signal values is more justified. In this work, we will use the *modified Thompson tau technique* (MTTT) as the method recommended by the National Standard of the American Society of Mechanical Engineers PTC 19.1.

Let the values of the quantity signal be sorted in increasing order. To detect outliers, proceed as follows:

1. Calculate the median and pseudo-standard deviation (which, unlike mean and standard deviation, are more robust to the presence of outliers):

$$M_{\mathbf{q}} = \begin{cases} q_{(m_{\mathbf{q}}+1)/2}, & m_{\mathbf{q}} \text{ is odd}, \\ \dfrac{q_{m_{\mathbf{q}}/2} + q_{m_{\mathbf{q}}/2+1}}{2}, & m_{\mathbf{q}} \text{ is even}, \end{cases}$$

$$s_{ps\mathbf{q}} = \frac{q_{0.75} - q_{0.25}}{1.349},$$

where $m_{\mathbf{q}}$ is the number of elements in the signal, $q_{0.75}$ ($q_{0.25}$) is the upper (lower) quartile.

2. For each signal element q_i, i=1,..., $m_{\mathbf{q}}$, calculate absolute deviations from the median:

$$d_i = \left| q_i - M_{\mathbf{q}} \right|.$$

3. Calculate

$$\tau = \frac{t_{\alpha/2} \cdot \left(m_{\mathbf{q}} - 1 \right)}{\sqrt{m_{\mathbf{q}}} \sqrt{m_{\mathbf{q}} - 2 + t^2_{\alpha/2}}},$$

where $t_{\alpha/2}$ is the critical Student's t value based on $m_{\mathbf{q}}$ – 2 degrees of freedom and significance level α.

4. Apply the following criterion: if for some index $d_i > \tau s_{ps\mathbf{q}}$, then the ith signal value is an outlier. In this case, remove it from the signal and return to step 1. If the criterion is not satisfied for all i, stop.

4. 4. Description of the experimental research of the modified method for providing group anonymity

To illustrate how the automated method of providing group anonymity works in practice, let us discuss the task of masking the regional distribution of military personnel working in the state of New York (the U. S.). We used the 1 % sample of the American Community Survey conducted in the U.S. in 2013 [27]. The part of the microfile with respondents working in New York contains 91,398 records.

We selected the attribute "Place of work: PUMA, 2 000 onward" to be the parameter attribute. Each parameter value defines the code of the district where the respondent works. We selected the attribute "Occupation, SOC classification" as the vital attribute. Its values are occupation codes according to the U. S. Standard Occupational Classification system (SOC). Vital values were chosen to be 551010, 552010, 553010, 559830, which correspond to active military personnel of different ranks. The quantity signal **q** built for the group of military personnel in the state of New York defined this way is given in Fig. 2 (solid line). Elements 1–38 of the signal correspond to districts where respondents work, with the district having the lowest number corresponding to 1, and so on.

In this quantity signal, we can detect two outliers: in the 5th and 29th districts ($OUT(\mathbf{q})$={5, 29}). The outlier in the 5th district corresponds to the Fort Drum military base, whereas the outlier in the 29th district corresponds to the West Point Military Reservation, which can be checked by analyzing [28].

To mask these outliers, restrictions were imposed on the corresponding signal elements as follows: $\mu_5(x) = Z_{MF}(x, 10, 97)$ and $\mu_{29}(x) = Z_{MF}(x, 10, 45)$, where

$$Z_{MF}(x,a,b) = \begin{cases} 1, & x \le a, \\ 1 - 2\left(\dfrac{x-a}{b-a}\right)^2, & a \le x \le \dfrac{a+b}{2}, \\ 2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2} \le x \le b, \\ 0, & x \ge b. \end{cases}$$

In these functions, parameter a plays the role of the threshold value ε. Plots of the corresponding functions are given in Fig. 1.

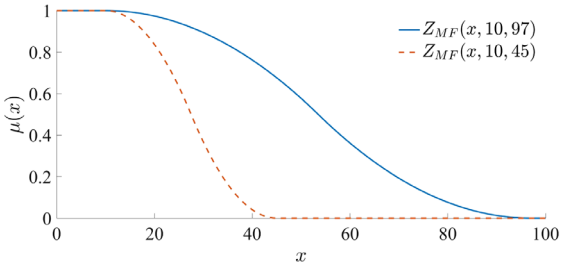

Fig. 1. Restrictions imposed on the values of the quantity signal **q** that correspond to outliers

Influential attributes were chosen according to Table 1. To simplify interpretation of metric (1), each attribute was considered categorical with the unit weight. The metric defined this way shows the total number of attribute values that need to be distorted after one swap takes place.

The fitness function (3) looks as follows:

$$f(U) = \frac{1846 - \sum_{i=1}^{Q} \sum_{k=1}^{13} \operatorname{sign} \left| \mathbf{M}_{u_{i1}}(u_{i2}, w_k) - \mathbf{M}_{u_{i3}}(u_{i4}, w_k) \right|}{1846} \times$$

$$\times ZMF\left(q_5(U), 10, 97\right) \cdot ZMF\left(q_{29}(U), 10, 45\right) \frac{1}{1 + e^{\frac{1}{2}(Q-25)}},$$

where w_k is the kth influential attribute, $\mathbf{M}_j(i, w_k)$, k=1,..., 13 are the values of the attribute w_k in the ith record from $\mathbf{M}_j$.

Recombination and local search operators were chosen as in [16]:

– recombination operator randomly generates two numbers k_1 and k_2, which lie from 0 to the number of rows in each of the parents, respectively, cuts the parents along the rows with corresponding indexes, and creates the offspring by exchanging the corresponding sections of individuals;

– local search operator for each row of individual U generates a random number r uniformly distributed on [0, 1]. If $r \le p_{mem}$, where p_{mem} is a parameter, then element u_{i4} is assigned an index of the record from the submicrofile with index u_{i3}, which is the most similar to the record u_{i2} from the submicrofile with index u_{i1}. If $r > p_{mem}$, then element u_{i2} is assigned an index of the record from the submicrofile with index u_{i1}, which is the most similar to the record u_{i4} from the submicrofile with index u_{i3}.

Mutation operators M_1 and M_3 were chosen to be swap mutation [29], mutation operators M_2 and M_4 were chosen to be random resetting mutation [21]. Selection method was chosen to be tournament selection [30]. Parameters of the algorithm and the TPGA are given in Table 2. Whenever the standard deviation of fitness values in the population fell below 0.03, mutation probability was increased tenfold in order to prevent premature convergence.

Table 1

Influential attributes for the TPGA

No.	Title	Values
1	Age	000 – less than 1 year old, 1..130 – 1 to 130 years, 135 – 135 years old
2	Educational attainment	00 – N/A or no schooling, 01 – nursery school to grade 4, 02 – grades 5–8, 03 – grade 9, 04 – grade 10, 05 – grade 11, 06 – grade 12, 07 – 1 year of college, 08 – 2 year of college, 09 – 3 year of college, 10 – 4 year of college, 11 – 5 year of college and higher
3	Sex	1 – male, 2 – female
4	Race	1 – White, 2 – Black/Negro, 3 – American Indian, 4 – Chinese, 5 – Japanese, 6 – other Asian, 7 – other race, 8 – two major races, 9 – three or more major races
5	Usual hours worked per week	00 – N/A, 01..98 – 1 to 98 hours per week, 99 – 99 hours and more
6	Hispanic origin	0 – not Hispanic, 1 – Mexican, 2 – Puerto Rican, 3 – Cuban, 4 – other, 9 – not reported
7	Marital status	1 – married, spouse present, 2 – married, spouse absent, 3 – separated, 4 – divorced, 5 – widowed, 6 – never married
8	Means of transportation to work	00 – N/A, 10 – automobile vehicle, 11 – auto, 12 – driver, 13 – passenger, 14 – truck, 15 – van, 20 – motorcycle, 30 – public transport, 31 – bus or trolley bus, 32 – streetcar, 33 – subway, 34 – railroad, 35 – taxicab, 36 – ferryboat, 40 – bicycle, 50 – walked, 60 – other, 70 – worked at home
9	Time of departure for work	0000 – N/A, other values report the time usually leaving for work (values 0001..2359 code time moments 00:01..23:59, respectively)
10	Travel time to work	000 – N/A, other values are amounts of time, in minutes, it took to get to work
11	Weeks worked last year	0 – N/A, 1 – 1–13 weeks, 2 – 14–26 weeks, 3 – 27–39 weeks, 4 – 40–47 weeks, 5 – 48–49 weeks, 6 – 50–52 weeks
12	Total personal income	A 7-digit numeric code reporting each respondent's income for the previous year, in USD
13	Speaks English	0 – N/A, 1 – does not speak English, 2 – speaks English, 3 – speaks only English, 4 – speaks very well, 5 – speaks well, 6 – speaks but not well, 7 – unknown, 8 – illegible

Table 2

Parameters of the memetic algorithm for solving the TPGA

Parameter	Value
Population size μ	100
Number of parent individual pairs λ	40
Recombination probability p_c	1
Mutation probability p_m	0.001
Local search parameter p_{mem}	0.75
Tournament size in the selection q	5
Compatibility threshold α_{comp}	0.5
Sensitivity threshold K_{out}	0.0
Distortion threshold K_{dist}	0.3
Number of algorithm runs	10
Number of generations in each run	1,000

Since visual analysis of 1,000 candidate solutions obtained in the final generations from all algorithm runs is very time-consuming and error-prone, MTTT with α=0.01 was used.

5. Results of the experiment for validating modification of the method for solving the TPGA

After conducting the experiment described above, only 24 solutions were selected as feasible. Corresponding solutions and their characteristics are given in Table 3 ordered by the cumulative value of (1).

The mean cumulative value of (1) for all 24 solutions is 467.958, which means that to provide group anonymity, it is sufficient to distort (on average) no more than 467.958/(13·91,398)≈0.04 % microfile attribute values.

Table 3

Modified quantity signals for the TPGA and their main characteristics

No.	Modified Quantity Signal	Fitness	Cumulative Metric (1)	μ(**q**)	*OUT*(**q**)
1	$\mathbf{q}^{*1}$=(2, 2, 1, 7, 14, 1, 14, 0, 5, 0, 2, 8, 0, 1, 1, 1, 4, 11, 1, 9, 0, 3, 2, 3, 2, 0, 3, 0, 13, 0, 4,14, 4, 5, 23, 7, 12, 6)	0.772	407	0.990	{35}
2	$\mathbf{q}^{*2}$=(2, 1, 1, 5, 15, 1, 13, 0, 4, 0, 2, 8, 0, 1, 1, 1, 4, 11, 1, 7, 0, 4, 2, 4, 3, 0, 5, 1, 12, 0, 4, 13, 4, 7, 21, 7, 13, 7)	0.772	410	0.993	{35}
3	$\mathbf{q}^{*3}$=(2, 2, 1, 3, 14, 1, 11, 0, 4, 0, 3, 9, 0, 1, 1, 1, 4, 11, 1, 7, 0, 3, 2, 5, 3, 0, 4, 0, 9, 0, 4, 10, 4, 7, 29, 7, 13, 9)	0.762	430	0.998	{35}
4	$\mathbf{q}^{*4}$=(1, 2, 3, 3, 13, 1, 10, 1, 6, 1, 3, 10, 0, 0, 1, 3, 5, 6, 2, 2, 0, 2, 1, 1, 3, 0, 3, 2, 8, 1, 8, 8, 14, 3, 29, 7, 12, 10)	0.757	432	0.999	{35}
5	$\mathbf{q}^{*5}$=(1, 1, 2, 0, 11, 1, 10, 0, 6, 0, 4, 9, 0, 1, 1, 1, 5, 8, 3, 5, 0, 4, 0, 0, 2, 1, 4, 2, 10, 0, 7, 7, 8, 4, 30, 7, 15, 15)	0.755	437	1.000	{35, 37, 38}
6	$\mathbf{q}^{*6}$=(0, 0, 1, 1, 11, 1, 11, 0, 5, 0, 3, 12, 0, 1, 2, 1, 6, 7, 2, 5, 0, 2, 1, 2, 1, 1, 2, 3, 10, 0, 7, 4, 11, 2, 32, 8, 18, 12)	0.748	450	1.000	{35, 37}
7	$\mathbf{q}^{*7}$=(1, 0, 1, 1, 10, 1, 12, 0, 6, 0, 4, 12, 0, 1, 2, 1, 6, 8, 2, 6, 0, 2, 1, 2, 2, 1, 2, 2, 8, 0, 6, 4, 10, 2, 33, 8, 17, 11)	0.716	459	1.000	{7, 12, 35, 37}
8	$\mathbf{q}^{*8}$=(0, 0, 1, 1, 7, 1, 8, 0, 9, 0, 2, 17, 1, 0, 0, 5, 3, 6, 4, 5, 0, 0, 4, 7, 0, 0, 2, 1, 11, 1, 5, 3, 14, 2, 38, 6, 12, 9)	0.714	461	1.000	{12, 33, 35}
9	$\mathbf{q}^{*9}$=(1, 1, 2, 0, 8, 1, 11, 0, 9, 0, 5, 11, 0, 1, 2, 1, 5, 10, 3, 5, 0, 2, 0, 0, 1, 1, 3, 2, 8, 0, 7, 6, 11, 3, 28, 8, 17, 12)	0.659	464	1.000	{35, 37}
10	$\mathbf{q}^{*10}$=(0, 4, 1, 7, 10, 2, 16, 1, 8, 0, 7, 11, 1, 2, 1, 1, 3, 6, 1, 5, 1, 5, 1, 1, 1, 1, 1, 2, 9, 0, 6, 2, 9, 1, 29, 5, 13, 11)	0.726	466	1.000	{7, 35, 37}
11	$\mathbf{q}^{*11}$=(0, 2, 3, 6, 10, 2, 12, 2, 7, 0, 2, 11, 1, 1, 1, 2, 3, 6, 3, 6, 0, 2, 1, 1, 4, 0, 1, 0, 7, 1, 6, 3, 7, 5, 29, 6, 16, 16)	0.691	466	1.000	{7, 35, 37, 38}
12	$\mathbf{q}^{*12}$=(2, 1, 1, 6, 9, 1, 15, 0, 5, 0, 3, 8, 0, 0, 1, 1, 4, 7, 1, 5, 0, 5, 1, 6, 2, 1, 6, 0, 7, 0, 5, 13, 5, 5, 28, 7, 13, 11)	0.658	468	1.000	{7, 32, 35, 37}
13	$\mathbf{q}^{*13}$=(1, 2, 4, 5, 9, 1, 10, 1, 6, 1, 4, 10, 0, 0, 1, 3, 6, 6, 2, 2, 0, 0, 1, 2, 4, 0, 4, 3, 8, 1, 6, 5, 15, 5, 27, 7, 14, 9)	0.689	470	1.000	{33, 35, 37}
14	$\mathbf{q}^{*14}$=(0, 0, 1, 8, 8, 1, 14, 0, 6, 1, 2, 11, 2, 3, 2, 2, 3, 7, 2, 5, 0, 0, 0, 2, 0, 0, 1, 4, 12, 1, 9, 8, 13, 3, 29, 6, 9, 10)	0.729	472	0.997	35
15	$\mathbf{q}^{*15}$=(0, 3, 5, 7, 10, 2 ,12, 1, 7, 0, 2, 9, 0, 2, 0, 2, 3, 6, 2, 5, 0, 2, 1, 2, 2, 0, 1, 0, 7, 1, 6, 2, 11, 6, 29, 6, 17, 14)	0.688	472	1.000	{7, 35, 37, 38}
16	$\mathbf{q}^{*16}$=(2, 2, 1, 7, 9, 1, 15, 0, 4, 0, 2, 7, 0, 1, 1, 1, 4, 11, 1, 8, 0, 3, 2, 4, 3, 0, 4, 0, 9, 0, 4, 15, 5, 5, 26, 7, 15, 6)	0.708	475	1.000	{7, 18, 32, 35, 37}
17	$\mathbf{q}^{*17}$=(2, 1, 1, 8, 8, 1, 14, 0, 4, 0, 3, 7, 0, 0, 1, 1, 4, 9, 1, 8, 0, 3, 1, 4, 3, 0, 5, 0, 8, 0, 4, 11, 4, 7, 32, 7, 14, 9)	0.652	479	1.000	35
18	$\mathbf{q}^{*18}$=(4, 0, 1, 3, 9, 1, 10, 0, 6, 1, 3, 14, 1, 1, 0, 1, 3, 7, 1, 2, 0, 3, 1, 0, 3, 6, 4, 0, 8, 2, 10, 19, 9, 6, 22, 6, 13, 5)	0.682	483	1.000	{12, 32, 35, 37}
19	$\mathbf{q}^{*19}$=(0, 2, 5, 3, 10, 1, 8, 1, 8, 1, 3, 10, 0, 0, 1, 4, 6, 6, 2, 2, 0, 1, 1, 1, 4, 0, 3, 3, 7, 1, 8, 5, 15, 5, 33, 7, 11, 7)	0.681	485	1.000	{33, 35}
20	$\mathbf{q}^{*20}$=(1, 0, 1, 1, 9, 1, 13, 0, 8, 0, 6, 10, 0, 0, 1, 1, 4, 9, 3, 5, 0, 2, 1, 2, 5, 1, 3, 3, 5, 0, 7, 6, 9, 5, 28, 7, 16, 12)	0.539	486	1.000	{35, 37}
21	$\mathbf{q}^{*21}$=(0, 0, 1, 7, 4, 1, 16, 0, 6, 1, 2, 9, 2, 3, 4, 3, 3, 6, 2, 4, 0, 1, 0, 2, 0, 0, 1, 3, 10, 1, 10, 6, 16, 3, 33, 6, 10, 9)	0.537	490	1.000	{7, 33, 35}
22	$\mathbf{q}^{*22}$=(2, 1, 1, 8, 4, 1, 17, 0, 4, 0, 3, 7, 0, 0, 1, 1, 4, 9, 1, 10, 0, 4, 1, 2, 3, 0, 4, 0, 8, 0, 4, 13, 3, 7, 32, 7, 13, 10)	0.363	505	1.000	{7, 35}
23	$\mathbf{q}^{*23}$=(2, 0, 1, 5, 8, 1, 10, 0, 5, 3, 3, 11, 1, 1, 0, 1, 3, 7, 1, 2, 0, 2, 1, 0, 4, 6, 6, 0, 5, 2, 10, 13, 12, 6, 30, 6, 13, 4)	0.450	513	1.000	{32, 35, 37}
24	$\mathbf{q}^{*24}$=(2, 2, 1, 6, 0, 1, 13, 0, 4, 0, 2, 9, 0, 2, 2, 1, 4, 13, 1, 11, 0, 5, 3, 4, 4, 0, 5, 0, 4, 0, 5, 17, 4, 5, 23, 8, 15, 9)	0.013	551	1.000	{7, 18, 20, 32, 35, 37}

Two solutions with the lowest cumulative value of (1) are given in Fig. 2. The following observations are true:

– compatibilities of these solutions with imposed restrictions satisfy (2) because

$$\mu(\mathbf{q}^{*1})=Z_{MF}(13, 10, 97)\cdot Z_{MF}(8, 10, 45)=0.990>\alpha_{comp};$$
$$\mu(\mathbf{q}^{*2})=Z_{MF}(11, 10, 97)\cdot Z_{MF}(10, 10, 45)=0.993>\alpha_{comp};$$

– all initial outliers are masked, because the set of initial outliers $OUT(\mathbf{q})=\{5, 29\}$ doesn't intersect the sets $OUT(\mathbf{q}^{*1})$ and $OUT(\mathbf{q}^{*2})$;

– values 407 and 410 are sufficiently lower than $K_{dist}\times C_{\max}=0.3\cdot 1846=553.8$.

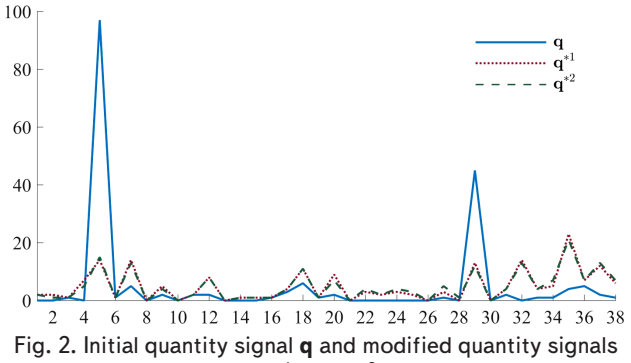

Fig. 2. Initial quantity signal **q** and modified quantity signals $\mathbf{q}^{*1}$ and $\mathbf{q}^{*2}$

Based on these observations, we can conclude that the solutions presented in Fig. 2 are feasible. Similar thoughts are true for the rest of 22 selected solutions.

6. Discussion of the results of solving the TPGA using the automated method

In the given experiment, 24 solutions to the TPGA obtained by using the memetic algorithm that satisfy the imposed requirements were selected automatically. Selection of these solutions was done by:

– automatic calculation of each modified signal's compatibility with the imposed restrictions and comparing it to the compatibility threshold α_{comp}=0.5. Solutions were selected whose compatibility exceeded 0.5;

– automatic calculation of each modified signal's cumulative value of (1) and comparing this value to the product of the distortion threshold and the maximum possible cumulative value of (1) for the given task $K_{dist} \cdot C_{\max}$=553.8. Solutions were selected whose cumulative value of (1) was less than 553.8;

– automatic detection of each modified signal's outliers using MTTT and comparing the set of outliers with the set $OUT(\mathbf{q})$={5, 29}. Solutions were selected whose sets did not intersect.

Calculations of modified signals' characteristics from the first two items of the list are performed inside the memetic algorithm when calculating their fitness values. Therefore, the main influence on the improvement in group anonymization efficiency is exerted by the automation of outlier detection using MTTT. Moreover, in the given experiment, only 24 out of 1,000 solutions turned out to be feasible, which constitutes only 2.4 % of the total number thereof. Visual analysis of such a number of solutions in order to select a small share of them is time- and resource-consuming. Therefore, applying the automated method for selecting TPGA solutions, compared to visual analysis of each and every signal, enables us to significantly reduce the total time for anonymization.

At the same time, in the proposed approach, an open question remains as to the influence of the parameter α in MTTT on the number and quality of the TPGA solutions. In fact, this parameter is, aside from compatibility, sensitivity, and distortion thresholds, an additional TPGA parameter. However, unlike the rest of the parameters, to correctly interpret it for the specialist in data anonymization, additional explanations and training are needed.

Additional research is needed for developing the instructions for selecting parameter α in MTTT. For this, the influence of selecting its value on the number of detected outliers in the initial and modified signals needs to be researched. A special attention needs to be paid to the possibility of artificial boosting of anonymization performance by hand-picking values of α, for which there will be no outliers in the modified signal in the first place.

7. Conclusions

1. It is established that existent methods of providing group anonymity are time-consuming, because they require visual analysis of solutions being obtained by the expert.

2. A modification of the method for solving the TPGA is proposed, which lies in automating outlier detection in the initial and modified quantity signals, and checking the outlier masking criterion. This modification enables us to improve data anonymization efficiency, which is reached by automating the process of outlier detection in the modified signals and by automatic selection of feasible solutions to the task.

3. Practical application of the method is demonstrated by solving the real data based task of masking the regional distribution of military personnel in the state of New York. It is established that only 24 out of 1,000 candidate solutions satisfy requirements of masking all the outliers. Visual analysis of these solutions requires much effort and time, which makes automation of the solution evaluation process cost-effective.

References

1. Health Insurance Portability and Accountability Act of 1996 (HIPAA) [Text] / E. M. Rafalski (Ed.) // Encyclopedia of Health Services Research. doi: 10.4135/9781412971942.n180
2. Patient Safety and Quality Improvement Act of 2005 (PSQIA) [Text]. – Federal Register. – 2001. – No. 73 (266).
3. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016: May 4, 2016 [Text] // Official Journal of the European Union. – 2016. – L 119. – P. 1–88.
4. Pfitzmann, A. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management. Version v0.34 [Electronic resource] / A. Pfitzmann, M. Hansen // Privacy and data security. – 2010. – Available at: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
5. Hawkins, D. Identification of Outliers [Text] / D. Hawkins. – Springer, 1980. – 198 p. doi: 10.1007/978-94-015-3994-4
6. Chertov, O. Group Anonymity [Text] / O. Chertov, D. Tavrov // Communications in Computer and Information Science. – 2010. – P. 592–601. doi: 10.1007/978-3-642-14058-7_61
7. Chertov, O. Microfiles as a Potential Source of Confidential Information Leakage [Text] / O. Chertov, D. Tavrov // Studies in Computational Intelligence. – 2014. – P. 87–114. doi: 10.1007/978-3-319-08624-8_4
8. Sweeney, L. k-Anonymity: A Model for Protecting Privacy [Text] / L. Sweeney // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. – 2002. – Vol. 10, Issue 05. – P. 557–570. doi: 10.1142/s0218488502001648
9. Angiuli, O. Statistical Tradeoffs between Generalization and Suppression in the De-identification of Large-Scale Data Sets [Text] / O. Angiuli, J. Waldo // 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC). – 2016. doi: 10.1109/compsac.2016.198
10. Templ, M. Statistical Disclosure Control for Micro-Data Using the R Package sdcMicro [Text] / M. Templ, B. Meindl, A. Kowarik // Journal of Statistical Software. – 2015. – Vol. 67, Issue 4. doi: 10.18637/jss.v067.i04

11. Machanavajjhala, A. L-Diversity: Privacy Beyond k-Anonymity [Text] / A. Machanavajjhala, D. Kifer, J. Gehrke, M. Venkitasubramaniam // ACM Transactions on Knowledge Discovery from Data. – 2007. – Vol. 1, Issue 1. doi: 10.1145/1217299.1217302
12. Domingo-Ferrer, J. From t-closeness to differential privacy and vice versa in data anonymization [Text] / J. Domingo-Ferrer, J. Soria-Comas // Knowledge-Based Systems. – 2015. – Vol. 74. – P. 151–158. doi: 10.1016/j.knosys.2014.11.011
13. Salazar-González, J.-J. Statistical confidentiality: Optimization techniques to protect tables [Text] / J.-J. Salazar-González // Computers & Operations Research. – 2008. – Vol. 35, Issue 5. – P. 1638–1651. doi: 10.1016/j.cor.2006.09.007
14. Parmar, A. A. Blocking Based Approach for Classification Rule Hiding to Preserve the Privacy in Database [Text] / A. A. Parmar, U. P. Rao, D. R. Patel // International Symposium on Computer Science and Society. – 2011. doi: 10.1109/isccs.2011.103
15. Singh, A. Privacy Preserving Techniques in Social Networks Data Publishing – A Review [Text] / A. Singh, D. Bansal, S. Sofat // International Journal of Computer Applications. – 2014. – Vol. 87, Issue 15. – P. 9–14. doi: 10.5120/15282-3880
16. Chertov, O. Two-Phase Memetic Modifying Transformation for Solving the Task of Providing Group Anonymity [Text] / O. Chertov, D. Tavrov // Studies in Fuzziness and Soft Computing. – 2016. – P. 239–253. doi: 10.1007/978-3-319-32229-2_17
17. Kleinberg, J. Algorithm Design [Text] / J. Kleinberg, E. Tardos. – Pearson, 2005. – 864 p.
18. Tavrov, D. Memetic approach to anonymizing groups that can be approximated by a fuzzy inference system [Text] / D. Tavrov // 2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) held jointly with 2015 5th World Conference on Soft Computing (WConSC). – 2015. doi: 10.1109/nafips-wconsc.2015.7284189
19. Chertov, O. Memetic Algorithm for Solving the Task of Providing Group Anonymity [Text] / O. Chertov, D. Tavrov // Studies in Fuzziness and Soft Computing. – 2014. – P. 281–292. doi: 10.1007/978-3-319-03674-8_27
20. Neri, F. A Primer on Memetic Algorithms [Text] / F. Neri, C. Cotta // Studies in Computational Intelligence. – 2012. – P. 43–52. doi: 10.1007/978-3-642-23247-3_4
21. Eiben, A. E. Introduction to Evolutionary Computing [Text] / A. E. Eiben, J. E. Smith. – Berlin, Heidelberg: Springer-Verlag, 2015. – 287 p. doi: 10.1007/978-3-662-44874-8
22. Zhang, Y. A multi-objective memetic algorithm based on decomposition for big optimization problems [Text] / Y. Zhang, J. Liu, M. Zhou, Z. Jiang // Memetic Computing. – 2016. – Vol. 8, Issue 1. – P. 45–61. doi: 10.1007/s12293-015-0175-9
23. Turky, A. A multi-population memetic algorithm for dynamic shortest path routing in mobile ad-hoc networks [Text] / A. Turky, N. R. Sabar, A. Song // 2016 IEEE Congress on Evolutionary Computation (CEC). – 2016. doi: 10.1109/cec.2016.7744313
24. Wang, Y. A Memetic Algorithm for Minimum Independent Dominating Set Problem [Text] / Y. Wang, J. Chen, H. Sun, M. Yin // Neural Computing and Applications. – 2017. doi: 10.1007/s00521-016-2813-7
25. Jain, P. Minimizing cyclic cutwidth of graphs using a memetic algorithm [Text] / P. Jain, K. Srivastava, G. Saran // Journal of Heuristics. – 2016. – Vol. 22, Issue 6. – P. 815–848. doi: 10.1007/s10732-016-9319-4
26. Aggarwal, C. C. Outlier Analysis [Text] / C. C. Aggarwal. – New York: Springer-Verlag, 2013. – 461 p. doi: 10.1007/978-1-4614-6396-2
27. Ruggles, S. Integrated Public Use Microdata Series: Version 6.0 [Electronic resource] / S. Ruggles, K. Genadek, R. Goeken, J. Grover, M. Sobek. – Minneapolis: University of Minnesota, 2015. – Available at: https://usa.ipums.org/usa/
28. Base Structure Report Fiscal Year 2014 Baseline – A Summary of the Real Property Inventory [Electronic resource]. – Available at: https://www.acq.osd.mil/eie/Downloads/BSI/Base%20Structure%20Report%20FY14.pdf
29. Syswerda, G. Schedule Optimization Using Genetic Algorithms [Text] / G. Syswerda // Handbook of Genetic Algorithms. – New York: Van Nostrand Reinhold, 1991. – P. 332–349.
30. Brindle, A. Genetic Algorithms for Function Optimization [Text]: Doctoral Dissertation and Tech. Rep. TR81-2 / A. Brindle. – Edmonton: University of Alberta, Department of Computer Science, 1981. – 93 p.