Широке впровадження в галузі офіційної статистики методів, що забезпечують анонімність даних про окремі групи (колективи) респондентів, стримується відсутністю відповідних промислових інформаційних технологій та систем. Запропоновано трирівневу клієнт-серверну архітектуру інформаційної технології забезпечення групової анонімності даних, у якій виділено клієнтів, сервери застосунків та бази даних, об'єднані в локальну мережу для підвищення безпеки первинних даних. Описано концептуальну модель даних у вигляді реляційної бази даних, наведено її ключові фрагменти. Дана модель охоплює всі основні сутності процесу забезпечення групової анонімності. Розглянуто реалізацію технології на основі платформи Java Enterprise Edition 8, сервера застосунків Oracle GlassFish Server, сервера баз даних MySQL та системи інженерних розрахунків SciLab.

Інформаційна технологія дає змогу забезпечувати групову анонімність даних у випадку існування загрози її порушення за рахунок аналізу даних допоміжного мікрофайлу. У технології передбачені операції побудови нечітких моделей груп за допомогою генетичного алгоритму та модифікація мікрофайлу за допомогою міметичного алгоритму, що дає змогу ефективно забезпечувати анонімність, уносячи в дані незначні спотворення. Загалом, запропонована інформаційна технологія базується на використанні шести застосунків: починаючи зі створення цільового подання мікрофайлу та завершуючи розв'язанням, власне, задачі забезпечення групової анонімності даних у мікрофайлі.

Застосування технології проілюстровано розв'язанням задачі забезпечення анонімності групи військових на основі реальних даних Спостереження за американським суспільством 2013 р. (American Community Survey – 2013). Показано, що розв'язання задачі силами колективу з п'яти фахівців дає змогу, щонайменше, в два з половиною рази пришвидшити процес підготовки мікрофайлу порівняно з існуючою технологією

Ключові слова: інформаційна технологія, групова анонімність, мікрофайл, нечітка модель, еволюційний алгоритм

UDC 004.09:004.62:510.6
DOI: 10.15587/1729-4061.2018.150805

IMPROVING EFFICIENCY FOR ENSURING DATA GROUP ANONYMITY BY DEVELOPING AN INFORMATION TECHNOLOGY

O. Chertov
Doctor of Technical Sciences, Professor
Department of Applied Mathematics
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Peremohy ave., 37, Kyiv, Ukraine, 03056
E-mail: chertov@i.ua
D. Tavrov
PhD
Department of Economics
Private Institution "University "Kyiv School of Economics""
Dmytrivska str., 92-94,
Kyiv, Ukraine, 01135
E-mail: dtavrov@kse.org.ua

1. Introduction

In the present-day world, a steady increase in the volume of digital data is observed. A considerable portion of these data needs to be made public in order to ensure possibility of conducting studies of various kinds. At the same time, adequate protection of the data from breach of privacy should be ensured. Organizations that publish data while ensuring their privacy are called data managing organizations [1]. Data managing organizations include national statistical organizations (for example, the State Statistics Service of Ukraine), international statistical organizations (for example, the European Union Statistical Office), trade associations, medical institutions, libraries, archives, etc.

In their activities, data managing organizations implement the CSID process of data processing which consists of four subprocesses: capture, storage, integration, dissemination. Data are captured through observations, censuses or surveys and stored as databases or individual microdata files (microfiles). Dissemination of such data involves creation of original tables or microfiles of certain data samples. Problems that data managing organizations solve within the framework of the CSID process of data processing include entry of data to the database and their control, verification, depersonalization and aggregation, data disclosure control, i.e. their anonymization.

Anonymity of an object in a data set is the property of this object to be indistinguishable among other elements of this set [2]. Usually, there are two kinds of anonymity in literature: individual anonymity connected with information about some respondent and group anonymity that is connected with information about a group of persons. Providing of group anonymity as a component of the CSID process of data processing was discussed for the first time in [3].

The widespread introduction of methods that ensure anonymity of data about individual groups of respondents is constrained by the lack of relevant industrial information technologies and systems. Therefore, development of an information technology (IT) to provide group anonymity of data as an integral part of implementation of the CSID process of data processing which should improve effectiveness of the process of preparing microfiles for publication is a focal

problem. Such IT should make it possible to form a group of respondents in a microfile and ensure its anonymity with a minimum possible distortion of the microfile data.

Since group anonymization requires different levels of training from IT users, it is advisable to distribute different operations and actions among users having different roles. The following distribution of roles is proposed:

– a statistician whose duties include data capture and input to the database (DB), etc.;

– a data scientist whose duties include preparation of metadata about microfiles (description of their structure) and their modification;

– junior analyst whose duties include direct anonymization of data including choice of group parameters and methods for ensuring group anonymity and evaluating quality of the resulting solutions;

– senior analyst whose duties include making final decision on the fact of ensuring group anonymity;

– database administrator whose duties include providing support for the database and preparation of microfiles for publication.

The developed IT should provide high level of reliability and security of primary data. The development tools used at the stage of the IT creation should be distributed freely.

2. Literature review and problem statement

A microfile is a two-dimensional data array in which each line corresponds to a particular *respondent* (*record*) and each column to some *attribute* of this respondent (record). There are three classes of attributes:

– *parameter attribute* having values enabling microfile to be broken into *parameter subfiles*, that is, data arrays in which entries have the same value of the parameter attribute;

– *vital attributes* which make it possible to form criteria of belonging of microfile records to a particular group using their values as a basis. A set that consists of microfile records with certain values of vital and parameter attributes of the microfile that correspond to a group shall be called further a *group model*;

– *basic attributes:* they are neither parameter nor vital attributes.

To violate group anonymity based on the microfile data, a *target microfile representation* (TMR) relative to a given group is created. The TMR in a form of a *quantitative signal*, i. e. a vector of values corresponding to the number of respondents belonging to the group and having value of the parameter attribute corresponding to the group is the most common in practice. Threat of group anonymity violation is defined in literature [4] as a possibility to detect in the TMR *outliers*, i.e. values that significantly exceed the rest of the values. Outliers in the TMR indicate an abnormal number of respondents belonging to the group relative to some value of the parameter attribute (for example, an abnormal number of military personnel in a certain region).

Group anonymity can only be ensured by primary data modification which introduces distortion to the data (preferably insignificant). Removal of an vital attribute from a microfile is such a modification at first glance. However, as shown in [5], group anonymity can be violated in a number of cases by means of extraneous data. A method of constructing a *fuzzy model of a group of respondents* based on an *auxiliary microfile*, i. e. a microfile close by its structure to that anonymity of which should be ensured, is proposed in [6]. Such a model is a set of fuzzy rules with the help of which it is possible to construct an *auxiliary target representation* with respect to the optional microfile (ATMR), outliers of which may coincide with the TMR outliers. In a case of successful construction of such a fuzzy model, removal of the vital attribute from the microfile is not a guarantee of group anonymity, so additional methods of its providing should be applied.

Since automated construction of a fuzzy rule base is a complex problem, it is proposed in literature to use genetic algorithms for its solution [7]. For the first time, such algorithms have been proposed for constructing trainable systems of classifiers [8]. In literature, there are two main approaches to constructing rule bases:

– according to the Michigan approach [9], each individual in the genetic algorithm is a separate rule;

– according to the Pittsburgh approach [10], each individual in the genetic algorithm is a complete rule base.

Advantage of the Michigan [11] approach is that rules do not depend on each other and its computational complexity is significantly smaller [12].

To date, the μ-ARGUS free application package developed in Java language [13] is the most powerful software product for ensuring data anonymity. By means of μ-ARGUS, individual anonymity of microfiles can be ensured using algorithms such as re-encoding and data roughening [14], *k*-anonymization [15], data exchange [16], noise polluting [17].

At the same time, the μ-ARGUS package does not ensure group anonymity and microfile data must be stored as separate files and not be introduced to a database which would facilitate creation of a corresponding information technology.

Free sdcMicro application package written in the programming language *R* [18] implements methods of micro-aggregation [19], noise pollution, data exchange, generation of synthetic data, recoding and roughening. At the same time, the need to own an environment *R* and represent microfiles as separate files reduces the scope of application of this software product. The package also does not support methods for providing group anonymity.

An information technology [20–22] described in literature supports providing of group anonymity of data. Microfiles for this technology are stored not as separate files but in a database and at the same time:

– values of the basic attributes are not taken into account, that is, it does not protect against breach of group anonymity through analyzing auxiliary microfiles;

– users have to repeatedly apply the anonymization method with different parameters until a microfile of satisfactory quality is obtained;

– users of this technology should be additionally instructed because it does not provide distribution of roles among users of various professions.

Therefore, there are grounds to state that virtually no industrial information technologies are available currently to provide group anonymity of data that would take into account a combination of values of basic microfile attributes and satisfy the requirements set forth earlier in section 1. This determines necessity of development of such information technology.

3. The aim and objectives of the study

This study objective was to improve effectiveness of ensuring group anonymity of data at the stage of prepar-

ing microfiles for publication by means of development of a specialized information technology which would make it possible to analyze data of auxiliary microfiles.

To achieve this objective, the following tasks were solved:

– to develop an IT architecture that satisfies the above requirements;

– to develop a conceptual model for the IT database that contains all necessary essentials and reflects their interrelation;

– to implement the IT using modern software development tools meeting the above requirements;

– to assess in practice improvement of effectiveness of the process of preparing microfiles for publication with the help of the developed IT.

4. Materials and methods used in the study of impact of the developed information technology on effectiveness of microfile preparation

4. 1. The task of providing group anonymity and methods of its solution

Denote by **M** the microfile for which it is necessary to provide group data anonymity. Denote the microfile entries are by $\mathbf{r}^{(i)}$, i=1,..., ρ, attributes by $\mathbf{w}_j$, j=1,..., η. Denote the number of values of the parameter attribute $\mathbf{w}_p$ by l_p. Then the microfile **M** can be broken into parameter submicrofiles $\mathbf{M}_1$,..., $\mathbf{M}_{lp}$, containing ρ_i number of entries each. Denote the TMR by $\mathbf{q}=(q_1, q_2,..., q_{lp})$ where q_k is the number of vital records contained in $\mathbf{M}_k$.

Denote indices of TMR values that are outliers by $OUT(\mathbf{q})$. Outliers are determined in literature [4] with the help of modified τ Thompson method (MMTT):

1. Find median $M_{\mathbf{q}}$ and pseudo-quadratic deviation $s_{ps\mathbf{q}}$ TMR with its values arranged in an ascending order:

$$M_q = \begin{cases} q_{(l_p+1)/2}, & l_p - \text{odd}, \\ \dfrac{q_{l_p/2} + q_{l_p/2+1}}{2}, & l_p - \text{even}, \end{cases}$$

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

where $q_{0.75}$ and $q_{0.25}$ are the upper and lower quartiles, respectively.

2. Calculate absolute deviations from the median $\forall q_i$, i=1,..., l_p:

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

3. Calculate quantity

$$\tau = \frac{t_{\alpha/2} \cdot (l_p - 1)}{\sqrt{l_p}\sqrt{l_p - 2 + t^2_{\alpha/2}}},$$

where $t_{\alpha/2}$ is the critical value t of the Student distribution for the number of degrees of freedom (l_p–2) and the level of significance α.

4. If $d_i > \tau s_{ps\mathbf{q}}$, then the i-th value of the TMR is a outlier. Then remove it from the TMR and proceed to step 1. If this criterion is not satisfied for any i, the algorithm ends.

The *task of providing group anonymity* (TPGA) consists in selecting such data modification that in the TMR $\mathbf{q}^*$ constructed for the modified microfile $\mathbf{M}^*$ outliers calculated for MMTT are masked, that is, $OUT(\mathbf{q}) \cap OUT(\mathbf{q}^*) = \emptyset$. In this case, distortions introduced in the microfile data have to be insignificant. In practice, microfiles are usually modified by means of a pairwise exchange of respondents similar in the sense of the defining metric [22]:

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

where I_k (J_l) is the k-th categorical (l-th ordinal) defining attribute, i. e. the attribute that is of interest for potential microfile researchers; $\chi(v_1, v_2)$ is equal to χ_1 if values v_1 and v_2 of the attributes belong to the same category and χ_2 is different; γ_k and ω_l are nonnegative weight coefficients (the greater weight the more important attribute).

Since the choice of a particular $\mathbf{q}^*$ is associated with a specific amount of distortions calculated as the sum of values (1) for each pair of respondents, the TPGA is reduced to selection of $\mathbf{q}^*$ which will provide minimum amount of distortions [22]. It is impossible to determine optimal $\mathbf{q}^*$ in advance, therefore, when solving TPGA, *fuzzy constraints* are imposed to values of $\mathbf{q}$ [23] that are set by functions $\mu_i(x)$ for each i-th value of $\mathbf{q}$. Each such function is equal to 1 for $x \le \varepsilon_j$, monotonously decreases to 0 for $\varepsilon_j < x \le q_j$ and is equal to 0 for $x > q_j$ where ε_j is the threshold value below which the i-th value of TMR $\mathbf{q}$ should be reduced. The quantity $\mu_i(q_i)$ is called *compatibility of* q_i with the fuzzy limitation imposed on it. Compatibility of $\mu(\mathbf{q})$ of the whole signal $\mathbf{q}$ with a set of fuzzy constraints is defined as the product of all $\mu_i(q_i)$, i=1,..., l_p. Threshold values should be determined as follows:

$$\varepsilon_j = \mathbf{q}^{K\max} - (q_j - \mathbf{q}^{K\max}) \cdot 0.2, \quad (2)$$

where $\mathbf{q}^{K\max}$ is the K-th largest value of a subset of TMR values $\mathbf{q}'=(q_j)$ which consists of values with indices belonging to $OUT'(\mathbf{q})$, the complement of $OUT(\mathbf{q})$ to the set $\{1,..., l_p\}$.

Taking into account the above, TPGA can be formulated as a task of searching for a sequence of paired exchanges of records in the form

$$\mathbf{S} = \left(\left(\mathbf{r}^{(i_1)}, \mathbf{r}^{(j_1)} \right), ..., \left(\mathbf{r}^{(i_Q)}, \mathbf{r}^{(j_Q)} \right) \right), \quad (3)$$

where i_k, j_k, k=1,..., Q are indexes of microfile records that are exchanged between submicrofiles in the frame of modification. Each such sequence will be called the TPGA *solution* and have to satisfy the following conditions:

$$\mu\left(\mathbf{q}^*(\mathbf{S}) \right) \ge \alpha_{comp},$$

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

$$\sum_{k=1}^{Q} \text{InfM}\left(\mathbf{r}^{(i_k)}, \mathbf{r}^{(j_k)} \right) \le K_{dist} \cdot C_{\max}, \quad (4)$$

where $\mathbf{q}^*(\mathbf{S})$ is modified TMR; $\mu(\mathbf{q}^*(\mathbf{S}))$ is compatibility of $\mathbf{q}^*$ with fuzzy constraints; α_{comp} is the compatibility thresh-

old; K_{out} is the sensitivity threshold; K_{dist} is the threshold of distortions; $C_{\max}$ is the maximum total value (1) for solved TPGA.

In literature, search for sequences of the above format is performed using *mimetic algorithms* (MA), that is, evolutionary algorithms that combine stochasticity with the elements of local search [24]. The population in MA for solving TPGA is composed of matrices of order $Q\times 4$ which are denoted by U. Each row of the matrix defines records for exchange between submicrofiles in the following way:

– element u_{i1}, i=1,..., Q is the index of the submicrofile from which record should be deleted;

– element u_{i3}, i=1,..., Q is the index of the submicrofile in which entry should be added;

– element u_{i2}, i=1,..., Q is the index of the record within the submicrofile to be deleted;

– element u_{i4}, i=1,..., Q is the record index within the submicrofile to be exchanged for the entry defined by u_{i1} and u_{i2}.

The structure of U is subjected to certain constraints. In particular, the number of entries of the submicrofile index i, i=1,..., l_p in the first column cannot exceed q_i, and $(\rho_i - q_i)$ in the third column. Microfile records cannot occur in U more than once.

The function of adaptability of individuals in the population has the form:

$$f(U)=\Phi(U)\cdot\Upsilon(U)\cdot\Psi(U)=$$
$$=\prod_{j=i_1}^{i_k}\mu_{A_j}\left(q_j^*(U)\right)\times\frac{C_{\max}-\sum_{i=1}^{Q}\mathrm{InfM}\left(\mathbf{M}_{u_{i1}}(u_{i2}),\mathbf{M}_{u_{i3}}(u_{i4})\right)}{C_{\max}}\times$$
$$\times\frac{1}{1+e^{\frac{1}{2}(Q_U-L)}}, \quad (5)$$

where $\Phi(U)$ is compatibility $\mu(\mathbf{q})$ of the signal with the imposed fuzzy constraints (a measure of quality of masking outliers in the interval [0, 1]); $\Upsilon(U)$ is a measure of quality of distortion minimization in the interval [0, 1]; $\Psi(U)$ is the term of penalty (in the interval [0, 1]) against an unlimited growth of dimensionality of individuals.

Mimetic algorithm includes the following steps:

1. Generate a population $P=\{U_i\}$ of μ individuals, i=1,..., μ, apply to each of them operator of local search $LS(U_i)$, i=1,..., μ.

2. Calculate value of the adaptivity function (5) to $\forall U_i$ from P.

3. Check the condition of completion and stop the algorithm if the condition is met.

4. Select λ pairs of paternal individuals and put them in the set P'.

5. Apply the operator of crossing $REC(U_{i1}, U_{i2})$ to each pair U_{i1}, U_{i2} from P'. Place heirs in the set P''.

6. Apply mutation operator

$$MUT(U_j)=(MUT_4\circ MUT_3\circ MUT_2\circ MUT_1)(U_j)\ \forall U_j \text{ from } P''.$$

Each operator MUT_k, k=1, ..., 4, acts on the k-th column of the individual U_j separately.

7. Apply $LS(U_j)$ $\forall U_j$ from P''.

8. Calculate the value of the adaptivity function (5) $\forall U_i$ from P''.

9. Select μ individuals from $P\cup P''$ with the largest value of the adaptivity function and place them in P into the place of μ individuals with the lowest value of the adaptivity function.

10. Go to step 2.

The cutting operator [25] was used in MA as the crossing operator, mutations of exchange (MUT_1, MUT_3) [26] and mutations of random exchange (MUT_2, MUT_4) [27] were used as the operators of mutation, the operator of tournament selection was used as the operator of selection [28].

The operator of local search [22] provides execution of the following steps:

1. Execute steps 2–3 for each row with U.

2. Randomly generate quantity r evenly divided into [0, 1].

3. If $r\le p_{mem}$ ($r>p_{mem}$), assign to the element u_{i4} (u_{i2}) index of the record from $\mathbf{M}_{ui3}$ ($\mathbf{M}_{ui1}$) closest as to (1) to the record u_{i2} (u_{i4}) from $\mathbf{M}_{ui1}$ ($\mathbf{M}_{ui3}$).

The initial population is generated randomly and individuals must have different number of rows. As a rule, the number of executed generations of the algorithm is the criterion of completion.

As noted above, it is sometimes possible to violate group anonymity, even if solutions are found that meet requirements set forth above. Having access to the $\mathbf{M}^{\text{aux}}$ auxiliary microfile, it is possible to build a *fuzzy group model* that can be used to determine TMR outliers in $\mathbf{M}$. The corresponding process consists of the following steps [6]:

– with the help of such fuzzy model, each microfile record can be juxtaposed to the degree of its membership in the $\mu_G(\mathbf{r}^{(i)})$ group as a value from the [0, 1] interval. The degree of membership is a measure of validity of the record membership in the group in absence of vital attributes which clearly indicate this membership;

– based on the degrees of membership of all records, one can build a ATMR in the form

$$q_j^{aux}=\left|\mathbf{r}\in\mathbf{M}_j\,|\,\mu_G(\mathbf{r})\ge\alpha\right|,\quad j=1,...,l_p, \quad (6)$$

where α is the membership threshold which can be used to delete entries with a low degree of membership (as a rule, α=0.5 is taken);

– detect outliers in (6) with the help of MMTT.

To enable construction of a fuzzy model, the auxiliary $\mathbf{M}^{\text{aux}}$ microfile should be similar in structure to the main $\mathbf{M}$ microfile. In particular, two microfiles can be harmonized, that is reduced to a single structure with attributes having values of the same interpretation.

The fuzzy model of the group consists of conditional fuzzy statements (fuzzy rules) R_i, i=1,... , m of canonical form

$$R_i:\ \text{If}\ L_1\in A_{i1}, L_2\in A_{i2},...,L_t\in A_{it},\ \text{then}\ G, \quad (7)$$

where L_k, k=1,..., t are linguistic variables [29] whose basic variables are defined on the sets of values of basic attributes of the microfile, w_{bk}, k=1,..., t, respectively; A_{ij} is the fuzzy value of the L_j variable which occurs in the fuzzy rule R_i; G is the class of records belonging to the group. Logical connective "and" is modeled as a fuzzy cut in a form of a product.

Each linguistic variable L_k in the fuzzy group model corresponds to some basic attribute of the microfile, w_{bk}, k=1,..., t. At the same time, each variable, L_k, k=1,..., t, has several values, LL_k^j, j=1,..., l_{Lk}. Entries with values not belonging to the carrier of at least one base variable of the corresponding linguistic variable are removed from the microfile.

The task of forming a set of fuzzy rules of form (7) can be considered as a task of identifying subgroups [30] in a set of respondents whose distribution is of interest for studies. Each rule in a fuzzy model is associated with a quality mea-

sure [31] which indicates its ability to effectively identify an interesting subgroup. Quality measures proposed in [6] are used in this study:

– discrimination factor:

$$DF(R_i)=\frac{\sum\limits_{\mathbf{r}^{aux}\in G} APC^{\alpha}\left(\mathbf{r}^{aux},R_i\right)}{\rho_v^{aux}}-\frac{\sum\limits_{\mathbf{r}^{aux}\in \mathbf{M}^{aux}} APC^{\alpha}\left(\mathbf{r}^{aux},R_i\right)}{\rho^{aux}}, \quad (8)$$

where ρ^{aux} is the number of entries in $\mathbf{M}^{aux}$; ρ_v^{aux} is the number of vital records in $\mathbf{M}^{aux}$; $\mathbf{r}^{aux}$ is the record from $\mathbf{M}^{aux}$; APC^{α} is the record compatibility with a fuzzy rule antecedent:

$$APC^{\alpha}(\mathbf{r},R_i)=\begin{cases}\prod\limits_j \mu_{A_{ij}}\left(r_{b_j}\right), & \text{if } \prod\limits_j \mu_{A_{ij}}\left(r_{b_j}\right)\geq\alpha,\\ 0, & \text{otherwise,}\end{cases}$$

where μ_{Aij} is the function of membership of the value A_{ij} of the linguistic variable L_j contained in the rule R_i; Π is the fuzzy cut; r_{bj} is the value of the j-th base attribute of the record $\mathbf{r}$. Positive value of (8) indicates that the fuzzy rule refers to the group disproportionately more vital records from $\mathbf{M}^{aux}$ than records from $\mathbf{M}^{aux}$ in general;

– relative validity factor:

$$RCF(R_i)=\frac{\sum\limits_{\mathbf{r}^{aux}\in G} APC^{\alpha}\left(\mathbf{r}^{aux},R_i\right)}{\sum\limits_{\mathbf{r}^{aux}\notin G} APC^{\alpha}\left(\mathbf{r}^{aux},R_i\right)}. \quad (9)$$

The value of (9) which exceeds threshold of relative validity γ indicates that the fuzzy rule incorrectly classifies a small number of entries from $\mathbf{M}^{aux}$ as belonging to the group.

The minuend in (8) is called a fuzzy rule carrier and denoted by κ.

Rules for a fuzzy group model can be built automatically based on the *genetic algorithm* (GA) first proposed in [32] which corresponds to the Michigan approach described above. The population in GA for building fuzzy rules consists of individuals, each corresponding to a separate fuzzy rule. Each rule is represented as a vector $R_i=(R_{i1}, R_{i2},..., R_{it})$ whose values are indices of fuzzy values of linguistic variables, L_k, k=1,...,.

Function of adaptability of individuals in the population has the form:

$$f(R_i)=\begin{cases}DF(R_i)\cdot RCF(R_i), & DF(R_i)>0,\\ 0, & DF(R_i)\leq 0,\end{cases} \quad i=1,2,\ldots,\mu. \quad (10)$$

The genetic algorithm consists of the following steps:

1. Generate population $\mathbf{R}=\{R_i\}$ with μ rules, i=1,..., μ.

2. Calculate value of the function of adaptability (10) $\forall R_i$ from **R**.

3. Check the completion condition and stop the algorithm if it is met.

4. Select λ pairs of paternal individuals and place them in the set **R'**.

5. Apply the operator of crossing, $REC(R_{i1}, R_{i2})$, to each pair R_{i1}, R_{i2} from **R'**. Place heirs in the set **R''**.

6. Apply mutation operator $MUT(R_j)$ $\forall R_j$ from **R''**.

7. Calculate value of the function of adaptivity of (10) to $\forall Ri$ from **R''**.

8. Substitute individuals from **R''** for λ pairs of individuals from **R** with the smallest value of the adaptivity function.

9. Go to step 2.

The operator of uniform recombination [33] is used as an operator of crossing in GA, mutation of random substitution [27] is used as a mutation operator, and the operator of tournament selection [28] is used as the selection operator. The original population is generated in a random way. Usually, the number of performed generations of the algorithm serves as a criterion of completion.

The rules derived from the use of GA should:

– have a positive value of (8);

– have the value of (9) exceeding the preset threshold γ;

– have a carrier κ exceeding the preset value;

– be not partial cases of more general rules.

Adequacy of the constructed fuzzy group model for the problem of detecting outliers in the TMR on the basis of auxiliary microfile will be evaluated using the metric proposed in [34]:

$$MB=\min_{\substack{0\leq t_1\leq l_p\\ 0\leq t_2\leq l_p'}} \ln B(t_1,t_2), \quad (11)$$

where B is Baussian factor calculated as

$$B(t_1,t_2)=\left[\frac{TP+FP+FN+TN+1}{(TP+FP+t_1+1)(FN+TN+t_2+1)}\right]\times$$
$$\times\left[\frac{(t_1+1)(t_2+1)}{t_1+t_2+1}\right]\cdot\frac{(TP+FP+FN+TN)!}{(TP+FN)!(FP+TN)!}\times$$
$$\times\sum_{i=0}^{t_1}\sum_{j=0}^{t_2}\frac{\left(\frac{t_1!}{i!(t_1-i)!}\right)^2\times\left(\frac{t_2!}{j!(t_2-j)!}\right)^2}{\frac{(t_1+t_2)!}{(i+j)!(t_1+t_2-i-j)!}\times\frac{(TP+FP+t_1)!}{(TP+i)!(FP+t_1-TP-i)!}\times\frac{(FN+TN+t_2)!}{(FN+j)!(TN+t_2-FN-j)!}},$$

where t_1, t_2 are nonnegative integers; *TP, FP, TN, FN* are elements of the matrix of inconsistencies

$$\mathbf{Z}=\begin{pmatrix}TP & FP\\ FN & TN\end{pmatrix}=$$
$$=\begin{pmatrix}\left|OUT(\mathbf{q})\cap OUT\left(\mathbf{q}^{aux}\right)\right| & \left|OUT'(\mathbf{q})\cap OUT\left(\mathbf{q}^{aux}\right)\right|\\ \left|OUT(\mathbf{q})\cap OUT'\left(\mathbf{q}^{aux}\right)\right| & \left|OUT'(\mathbf{q})\cap OUT'\left(\mathbf{q}^{aux}\right)\right|\end{pmatrix}.$$

Interpretation of values (11) is performed according to Table 1.

Table 1

Interpretation of the values of the metric of adequacy of the fuzzy group model

Metric value	Adequacy
Less than 0	Very low
From 0 to 1	Low
From 1 to 3	Mediocre
From 3 to 5	Strong
More than 5	Very strong

Typically, five different values of such a metric are sufficient for a qualitative description of the model adequacy.

4. 2. Architecture of the information technology for providing group anonymity. Conceptual data model

This study proposes a three-level client-server IT architecture (Fig. 1) with workstations of the users having different roles, an application server and a database server. This approach to creation of an IT enables fulfillment of the requirements set forth in Section 1, in particular:

– support for users having different roles;

– ensuring a high level of reliability and security of primary data by means of their storage in a separate DB server with a limited access;

– ensuring high flexibility and efficiency of the system by distributing tasks among the application servers and the DB.

The IT components perform the following functions:

– the application server manages connections and transactions of clients, their authentication, simultaneous processing of data flows, balancing of the network load, etc.;

– the database server manages the database, provides data integrity, processes requests from clients, manages user accounts, etc. Clients do not have direct access to the database (all communications are performed through the application server) which improves level of data security;

– client workstations provide clients with opportunities to perform functions in accordance with task allocation. For example, statistician may view and edit the group parameters, TMR outliers, read metadata. Data scientist may view and edit metadata. Junior analyst may view and edit the TMR values, fuzzy model parameters, MMTT, GA, MA, TPGA solutions, review metadata, group parameters and TPGA. Senior analyst may review any information, edit TPGA parameters. The database administrator may review any information from the database, edit metadata;

Applications perform the following functions:

– the application of TMR creation builds corresponding TMR or auxiliary TMR;

– the application of harmonization of microfiles harmonizes main and auxiliary microfiles;

– the application of outlier detection detects outliers in the TMR with the help of MMTT;

– the application of creation of fuzzy rules starts GA for deriving rules of a fuzzy group model;

– the application of verification of model adequacy calcue lates the metric of adequacy (11) for the given model;

– the application of TPGA solution starts the MA with parameters set for solving the corresponding task.

The information technology proposed in this paper can be presented in three stages:

– the stage of constructing the group model (S1);

– the stage of constructing the fuzzy group model (S2);

– the stage of solving TPGA (S3).

The stage S1 consists of the following operations and actions:

– operation O1-1: Group Specifying (actions: A1-1-1 Downloading Metadata, A1-1-2 Attribute Specifying, A1-1-3 Task Type Specifying);

– operation O1-2: Detecting Outliers in the TMR (aca tions: A1-2-1 Construction of TMR, A1-2-2 Performing MMTT, A1-2-3 Screening Outliers).

Stage S2 consists of the following operations and actions:

– operation O2-1: Creating an Auxiliary Microfile (aca tions: A2-1-1 Downloading Auxiliary Metadata, A2-1-2 Harmonization of Microfiles);

– operation O2-2: Building a Fuzzy Model" (actions: A2-2-1 Specifying Basic Attributes", A2-2-2 Defining Fuzzy Values, A2-2-3 Specifying Parameters and Performing EA);

– operation O2-3: Verifying the Fuzzy Model Adequacy (actions: A2-3-1 Constructing the ATMR, A2-3-2 Performing the MMTT, A2-3-3 Outlier Screening, A2-3-4 Calculating Adequacy Metric).

Stage S3 consists of the following operations and actions:

– operation O3-1: Specifying TPGA parameters (aca tions: A3-1-1 Specifying Thresholds, A3-1-2 Specifying the *K*-th Largest TMR Value);

– operation O3-2: Performing MA (actions: A3-2-1 Specifying Functions of Membership of Fuzzy Constraints, A3-2-2 Specifying MA Parameters, A3-2-3 Launch of MA, A3-2- 4 Selecting TPGA Solution).

Sequence of performing all actions by clients of different roles is presented in the UML diagrams of activity (Fig. 2, 3).

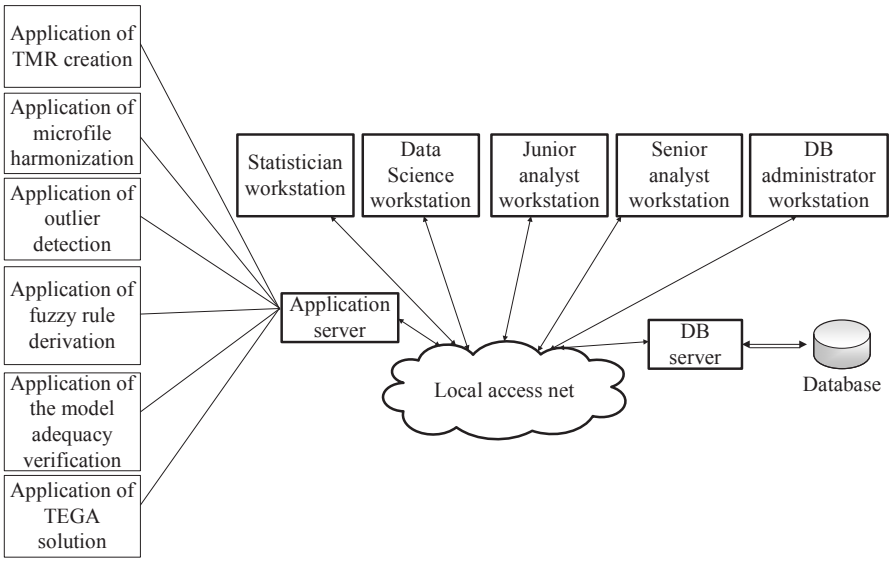

Fig. 1. Architecture of the information technology for providing data group anonymity

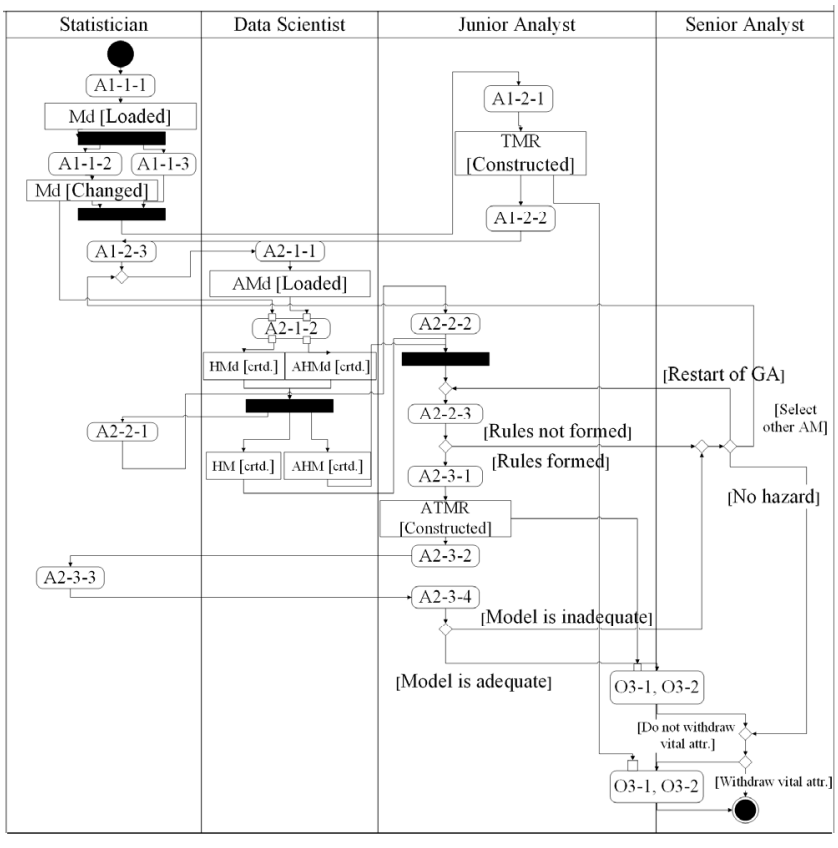

Fig. 2. Diagram of activity of clients of IT for providing group anonymity (stages S1, S2)

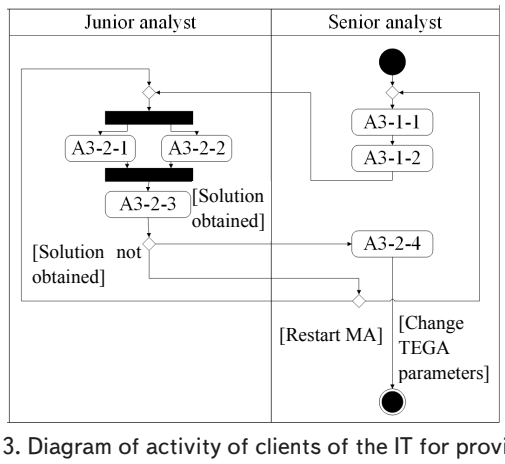

Fig. 3. Diagram of activity of clients of the IT for providing group anonymity (stage S3)

The conceptual data model developed within the framework of the proposed IT in a form of a relational database located on the database server can be presented in several fragments. Fig. 4 shows a fragment of the conceptual model of data which corresponds to essences of the microfile and its target representation. The essences contained in this fragment correspond to the following concepts related to the providing of group anonymity:

– the Microfile essence corresponds to data in the microfiles and contains, besides the primary key ID_Microfile (identifier) mandatory attributes: MI_Name (microfile name), MI_Desc (microfile description), MI_Data (microfile data in BLOB format);

– the Attribute essence corresponds to attributes of the microfiles and contains, besides the primary key, AT_Name (attribute name), mandatory attributes: AT_Desc (attribute description) and AT_Type (attribute type: nominal or categorical);

– the AttrValue essence corresponds to the values of attributes of the microfiles and contains mandatory attributes: AV_Date (date of value entry), AV_Value (value) and AV_Desc (semantic description of the value);

– the AttrCharacteristic essence corresponds to characteristics of the microfile attributes and contains, besides the primary AC_Date key, mandatory attributes: AC_Weight (the attribute weight used in the metric (1)) and AC_Xi (parameter χ from the metric (1));

– the GRM essence corresponds to the TMR and contains, besides the primary key GR_Date (creation date), mandatory attribute GR_Data (TMR value in a form of a text line having values separated with commas);

– the MMTT_Params essence corresponds to the parameters of the MMTT and contains, besides the primary key MMTT_Date (creation date), mandatory attribute MMTT_Alpha (level of significance α from the MMTT described in section 4.1);

– the MMTT_GRM essence corresponds to signal outliers: optional GR_Outliers attribute contains indexes of the TMR values that correspond to the outliers detected wih the help of MMTT in a format similar to the format of storage of values of the TMR itself;

– the Visual_Outlier essence corresponds to the TMR outliers selected by the statistician for masking and contains, besides the primary key VO_Date (creation date), the mandatory attribute VO_Outlier which contains indexes of the TMR values that correspond to those outliers in a format similar to the format for storing values of the TMR itself;

– the Problem essence corresponds to the TPGA and contains, besides the primary key ID_Problem (identifier),

mandatory PR_Date attributes (task creation date), PR_Stage (number of the stage at which the task is located) and the optional PR_RemoveVital attribute (a flag indicating whether there is a need to remove vital attributes from the microfile);

– the User essence corresponds to the IT users and contains, beside the primary key ID_User (identifier), mandatory attributes US_Login (user login), US_Password (user password), US_Role (user role), US_Name (username), US_IsActive (check box) besides the primary ID_User key (identifier), mandatory attributes: US_Login (the user login), US_Password (the user password), US_Role (the user role), US_Name (the user name), US_IsActive (a flag indicating whether the user is active in the system) and US_Date (date of user creation);

– the Role essence corresponds to roles of the IT users and contains, besides the primary RO_Date key (creation date), mandatory attributes: RO_Title (role name) and RO_IsActive (a flag indicating whether the role is active in the system);

– the UserRole essence is required for organizing "many-to-many" communication between the IT users and their roles.

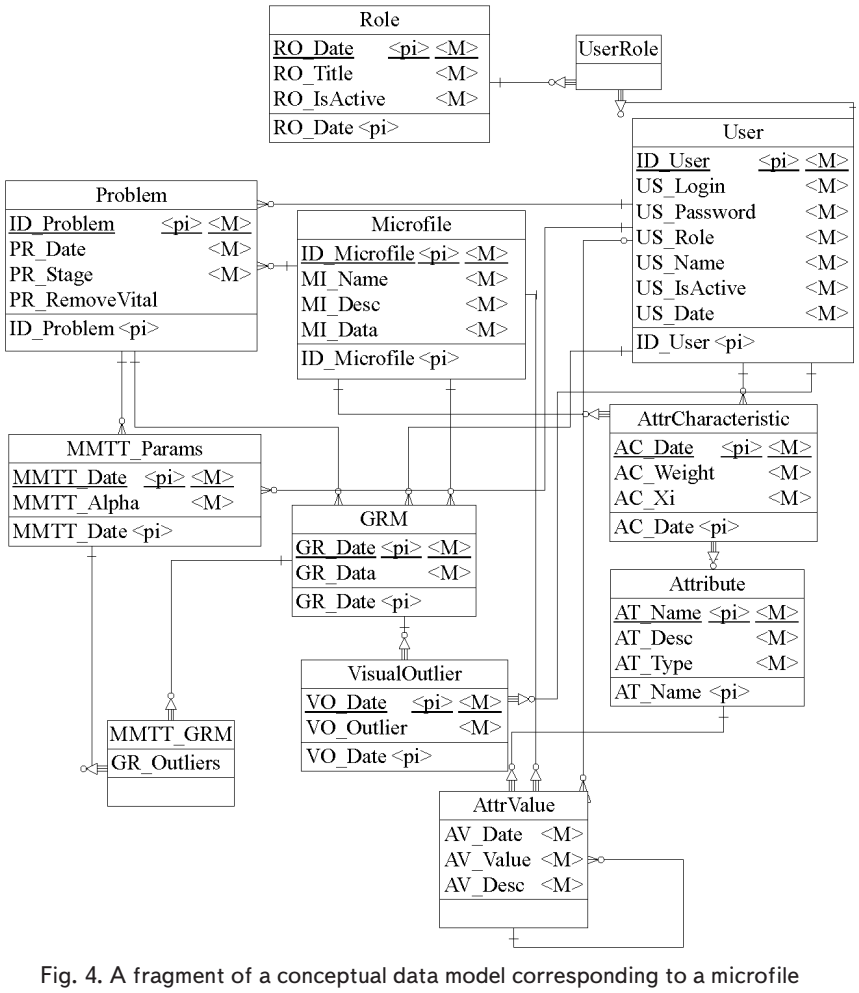

Fig. 4. A fragment of a conceptual data model corresponding to a microfile

Fig. 5 shows a fragment of a conceptual data model that corresponds to fuzzy group models in a microfile. The essences contained in this fragment correspond to the following concepts related to providing of group anonymity:

– the FuzzyRule essence corresponds to the rules of the fuzzy group model and contains, besides the primary key FRU_Date (date of rule creation), the following mandatory attributes:

– FRU_Rule (a rule in a form of a line with values separated by commas and each of them corresponds to the index of fuzzy value of the linguistic variable from the number of those associated with this rule);

– FRU_DF (factor of discrimination (8) of the rule);

– FRU_RCF (factor of relative adequacy (9) of the rule);

– FRU_Kappa (carrier of the κ rule);

– the FuzzyModelParameter essence corresponds to the fuzzy values of the linguistic variables included in the model and contains, besides the primary key FMP_Date (date of creation of parameters), mandatory attributes FMP_Name (name of the value) and FMP_Params (parameters of the membership function of the corresponding fuzzy value in a form of a line with values separated by commas);

– the LinguisticVariable essence corresponds to linguistic variables and contains, besides the LV_Date primary key (date of variable creation), mandatory attributes LV_Desc (variable description) and LV_Name (variable name);

– the ModelAdequacy essence corresponds to the metric of adequacy of the fuzzy group model and contains, besides the MAD_Date primary key (date of the metric values creation), mandatory attribute MAD_MB (the value of metric (11));

– the GA_Params essence corresponds to the EA parameters for creation of a fuzzy group model and contains, besides of the GA_Date primary key (date of creation of parameters), the following mandatory attributes:

– GA_Mu (population size) and GA_Lambda (number of parents);

– GA_ProbC (probability of crossing) and GA_ProbM (mutation probability);

– GA_N (number of generations) and GA_NG (number of EA launches);

– GA_TournSize (the size of the tournament in selection);

– GA_Gamma (threshold of relative adequacy) and GA_Kappa (carrier threshold);

– the AGRM essence corresponds to the auxiliary TMR and contains, besides the AGRM_Date (creation date) primary key, mandatory attributes of the value of the auxiliary AGR_Data (the value of the auxiliary TMR in a form of a text line with values separated by comma) and the AGR_CrispData (the value of the clear auxiliary TMR in a form of a text string with values separated by commas);

– the MMTT_AGRM essence corresponds to outliers of the auxiliary signal: optional AGR_Outliers attribute contains indexes of values of the auxiliary TMR corresponding to the outliers detected by using MMTT in a

format similar to the format of storage of the values of auxiliary TMR itself;

– the Aux_Visual_Outlier essence corresponds to outliers of the auxiliary TMR selected by the statistician for masking and contains, in addition to the AVO_Date (creation date) primary key, mandatory AVO_Outlier attribute which contains indexes of values of the auxiliary TMR corresponding to those outliers in a format similar to the format of storing values of the auxiliary TMR itself;

– the MicrofileProblem essence is needed to organize the "many-to-many" communication between microfiles and the TPGA.

Fig. 6 shows a fragment of the conceptual model of data which corresponds to the TPGA solution.

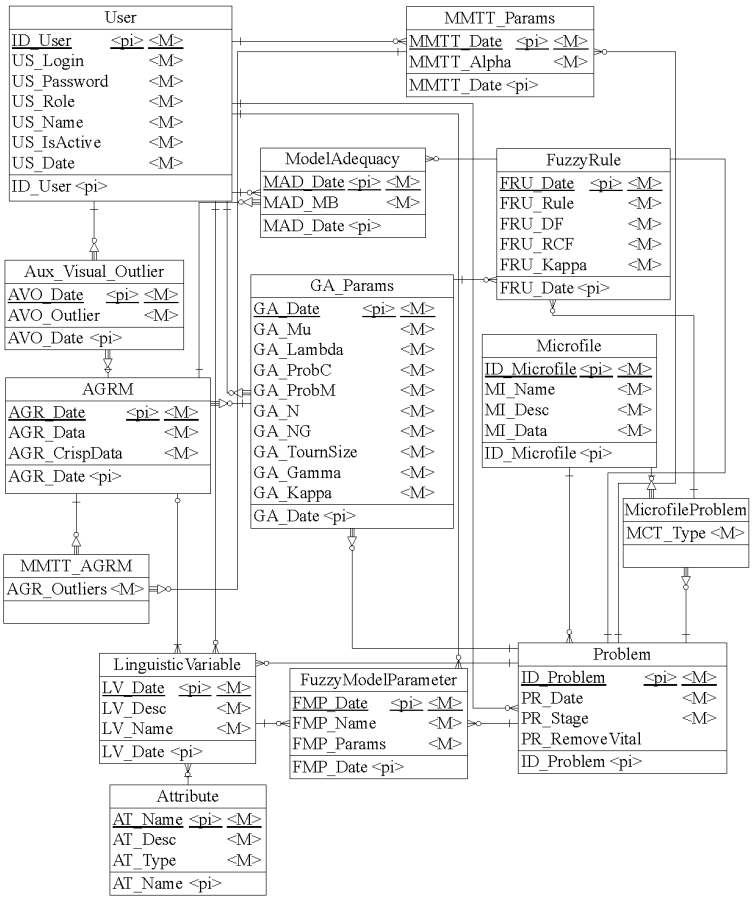

Fig. 5. A fragment of a conceptual model of data that corresponds to fuzzy group models in a microfile

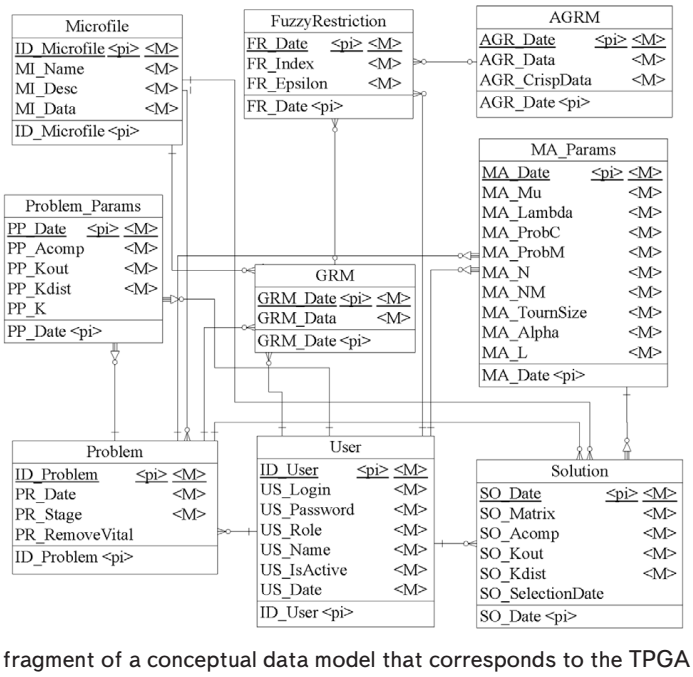

Fig. 6. A fragment of a conceptual data model that corresponds to the TPGA solution

The essences in this fragment correspond to the following concepts related to providing of group anonymity:

– the MA_Params essence corresponds to the MA parameters for solving the TPGA and contains, besides the MA_Date (date of creation of parameters) primary key, the following mandatory attributes:

– MA_Mu (population size) and MA_Lambda (number of parents);

– MA_ProbC (probability of crossing);

– MA_ProbM (probability of mutation);

– MA_N (number of generations);

– MA_NG (number of MA starts);

– MA_TournSize (the size of the selection tournament);

– MA_Alpha (degree of significance for MMTT);

– MA_L (maximum number of rows in an individual from the population);

– the FuzzyRestriction essence corresponds to fuzzy constraints on the TMR value and contains, besides the FR_Date (date of creation of parameters), primary key, mandatory attributes:

– FR_Index (the index of the TMR value to which restriction is imposed);

– FR_Epsilon (threshold value).

– The Problem_Params essence corresponds to the TPGA parameters and, besides the PP_Date (date of creation of parameters) primary key, contains the following mandatory attributes:

– PP_Acomp (threshold of compatibility of TPGA solution with fuzzy constraints);

– PP_Kout (sensitivity threshold of TPGA solution);

– PP_Kdist (threshold of distortion of TPGA solution);

– optional attribute PP_K (value of K for forming fuzzy constraints);

– The Solution essence corresponds to the TPGA solutions and, besides the SO_Date (date of solution creation) primary key, contains the following mandatory attributes:

– SO_Matrix (solution as an individual in MA in a form of a line with values separated by commas);

– SO_Acomp (compatibility of the TPGA solution with fuzzy constraints);

– SO_Kout (sensitivity of the TPGA solution);

– SO_Kdist (distortions introduced by TPGA solution);

– optional SO_SelectionDate attribute (date of selection of the TPGA solution).

4. 3. Implementation of the information technology providing group anonymity

To implement the IT, tools meeting the requirements put forward in section 1 were chosen:

– Oracle GlassFish Server has been selected as an application server because it provides interaction with clients using a small number of software tools, enables efficient organization of work in the database and is distributed freely. Client interaction with the server was organized using Java Message Service realized with the help of Apache ActiveMQ since it enables asynchronous message interchange between the server and clients which increases the IT flexibility;

– MySQL server has been chosen as the database server because it is easy to use and freely distributed. Interaction of clients with the database was organized using Java Database Connectivity interface;

– two different systems were used to realize IT applications:

– Java Platform, Enterprise Edition 8 which is relatively easy to use, portable, stable and distributed freely;

– Scilab engineering calculation system which supports matrix computation needed for effective implementation of the methods providing group anonymity and, unlike its counterparts, is distributed freely;

– Java Platform, Standard Edition 8, and the Swing library were used to support clients' workstations.

Individual IT applications perform the following functions:

– the TMR creation application is launched from the junior analyst's workstation. The application calls the Scilab buildGRM function. The microfile data read from the database and indexes of vital and parameter attributes with their values are sent to this function input. The function returns a one-dimensional array in which every element corresponds to the number of respondents belonging to the group and has a corresponding parameter value. The application records the corresponding array in the database and sends it to the junior analyst's workstation;

– the application of microfile harmonization is started by the data scientist. It calls the Scilab Harmonize function and data of the main and auxiliary microfiles read from the database as well as the harmonization parameters specified by the data scientist in his workstation are sent to its input. Harmonization parameters are represented as an object of the HarmonizationParams class. The harmonize function returns harmonized data of microfiles. The application records corresponding data to the database and sends the harmonized metadata of both microfiles to the statistician and data scientist workstations;

– the application for outlier detection is started by junior analyst. The application calls the Scilab detectOutliers function and parameter α MMTT and TMR read from the DB are sent to its input. The detectOutliers function returns the array of indexes of TMR values calculated by the MMTT method that are outliers. The application records the corresponding array to the database and sends it to the junior analyst's workstation;

– the application of construction of fuzzy rules is started by junior analyst. The application calls the Scilab ga function and data of the auxiliary microfile, parameters of the GA and values of the linguistic variables read from the database tare sent to its input. The function performs the GA to construct the fuzzy model and returns a matrix whose rows are fuzzy rules as well as characteristics of these rules. The application records the found rules to the database and sends them to the junior and senior analysts workstations;

– the application for verifying the model adequacy is started by junior analyst. The application calls Scilab getModelAdequacy function and data on outliers in the TMR and auxiliary TMR read from the database are sent to its input. The function calculates value of metric of adequacy (11). The application records the calculated value to the database and sends it to the junior analyst's workstation;

– the application for TPGA solution is launched by junior analyst. The application starts the Scilab ma function and the microfile data, TMR data, TPGA and EA parameters as well as fuzzy constraints on TMR values read from the DB are sent to the input of this function. The function executes MA of the TPGA solution and returns the resulting solutions. The application records solutions and their characteristics to the database and sends to the junior and senior analysts' workstations. At workstations of the relevant analysts, solutions are applied to the TMRs, and the modified TMRs are displayed on the screens.

4. 4. Description of the experimental study of the information technology providing group anonymity

To illustrate work of the information technology providing data group anonymity, consider the task of masking geographical distribution of the US servicemen in an **M** microfile of a one-percent sample from Observations on American Society (2013) [35]. The task was solved by a team of five specialists having roles of a statistician, a data scientist, junior and senior analysts, and a database administrator.

The **M** microfile contains 1,380,924 entries and, among other things, the following attributes:

– Occupation (SOC Classification) with values 551010, 552010, 553010 and 559830 (occupational codes according to the US Occupational Classification) which correspond to servicemen of various ranks. The Occupation attribute will be considered as vital attribute for the given TPGA;

– Place of Work: State, 1980 Onward and Place of Work: PUMA, 2000 onward), Domain of Microfiles of a Free Use which in combination determine a unique code of the geographical unit in which this or that respondent is working. Combination of these two attributes shall be considered as a parameter attribute for the given TPGA.

Let a user, senior analyst, decide to remove the Occupation attribute from the microfile. The main purpose of application of the developed IT is to check whether such removal is sufficient to ensure anonymity of the military group and, if insufficient, apply the corresponding MA.

A user, statistician, have chosen a microfile of a five-percent sample of the US population census (2000) [35] containing 6,309,848 records as an auxiliary $\mathbf{M}^{aux}$ microfile. This microfile is similar in structure to the main microfile, **M**. In particular, both microfiles have been harmonized by a user, data scientist, as follows:

– only the parameter attribute Place of Work, the vital attribute Occupation and 13 basic attributes given in the Table 2 were left in both microfiles. Each basic attribute for metric (1) calculation is considered categorical with a weight of 1: the metric (1) shows the number of attribute values that are distorted by one exchange of records;

– the Occupation attribute value in both microfiles was transformed as follows: entries that had values 551010, 552010, 553010 and 559830 acquired value "1" and the rest of the records got values "0".

Linguistic variables for the fuzzy group model, $L_j, j=1,...,13$, that correspond to the basic harmonized attributes in Table 1 have parameters given in Table 3. The following types of membership functions are used in the table:

$$PIMF(x;a;b;c;d)=\begin{cases}0, & x\le a,\\ 2\left(\dfrac{x-a}{b-a}\right)^2, & a\le x\le \dfrac{a+b}{2},\\ 1-2\left(\dfrac{x-b}{b-a}\right)^2, & \dfrac{a+b}{2}\le x\le b,\\ 1, & b\le x\le c,\\ 1-2\left(\dfrac{x-c}{d-c}\right)^2, & c\le x\le \dfrac{c+d}{2},\\ 2\left(\dfrac{x-d}{d-c}\right)^2, & \dfrac{c+d}{2}\le x\le d,\\ 0, & x\ge d,\end{cases}$$

$$TRAPMF(x;a;b;c;d)=\begin{cases}0, & x\le a,\\ \dfrac{x-a}{b-a}, & a\le x\le b,\\ 1, & b\le x\le c,\\ \dfrac{d-x}{d-c}, & c\le x\le d,\\ 0, & x\ge d,\end{cases}$$

$$GAUSSMF(x;a;b)=e^{\frac{(x-b)^2}{2a^2}},$$

$$SIGMF(x;a;b)=\frac{1}{1+e^{-a(x-b)}}.$$

Table 2

Basic attributes harmonized for TPGA

No.	Name	Value
1	Age	000: less than 1 year; 1 to 130: 1 to 130 years; 135: 135 years
2	Sex	1: male, 2: female
3	Educational attainment [general version]	00: does not apply or no education, 01: elementary school ,up to the 4th grade, 02: 5–8th grades, 03: 9th grade, 04: 10th grade, 05: 11th grade, 06: 12th grade, 07: 1st year of college, 08: 2nd year of college, 09: 3rd year of college, 10: 4th year of college, 11: 5th year of college and higher
4	Marital status	1: married, living in family, 2: married, family living separately, 3: divorced, 4: broken marriage, 5: widower, 6: never married
5	Total personal income	Respondent's income for the previous year, USD
6	Usual hours worked per week	00: does not apply, 01 to 98: 1 to 98 hrs/week, 99: 99 hrs/week and more
7	Weeks worked last year, intervalled	0: does not apply, 1: 1 to 13 weeks, 2: 14 to 26 weeks, 3: 27 to 39 weeks, 4: 40 to 47 weeks, 5: 48 to 49 weeks, 6: 50 to52 weeks
8	Race [general version]	1: Caucasoid, 2: Negroid, 3: Indian, 4: Chinese, 5: Japanese, 6: other Mongoloid, 7: other race, 8: two main races, 9: three and more main races
9	Hispanic origin [general version]	0: not of Latin American origin, 1: Mexican, 2: Puerto-Rican, 3: Cuban, 4: other, 9: nonindicated
10	Means of transportation to work	00: does not apply, 10: motor transport, 11: car, 12: driver, 13; passenger, 14: lorry, 15 van, 20: motorcycle, 30: public transport, 31: bus or trolleybus, 32: tram, 33: metro, 34: railway, 35: taxi, 36: ferry, 40: bicycle, 50: on foot, 60: other, 70: working at home
11	Time of departure for work	0000: does not apply, other values correspond to the time of leaving home for work last week (values 0001 to 2359 correspond to the instants of time from 00:01 to 23:59, respectively)
12	Travel time to work	000: does not apply, other values correspond to time spent on the way to work, minutes
13	Speaks English	0: does not apply, 1: not speaking, 2: speaking, 3: speaking just English, 4: speaking English very well, 5: speaking English well, 6: speaking English not very well, 7: unknown, 8: impossible to establish

Table 3

Parameters of linguistic variables for a fuzzy group model for TPGA

No.	Bearer of the base variable	Value
1	[18, 45]	"Very young": $\mu_{1.1}(x)$= =*PIMF*(x; 7.05; 15.40; 22,50; 27,20) "Young": $\mu_{1.2}(x)$=*GAUSSMF*(x; 2.0; 27.5) "Middle-aged": $\mu_{1.3}(x)$=*GAUSSMF*(x; 2.0; 32.5) "Not very old": $\mu_{1.4}(x)$=*GAUSSMF*(x; 2.0; 37.5) "Old": $\mu_{1.5}(x)$= =*PIMF*(x; 37.85; 42.50; 47.50; 54.85)
2	[1, 2]	"Male": $\mu_{2.1}(x)$=*TRAPMF*(x;1;1;1;1) "Female": $\mu_{2.2}(x)$=*TRAPMF*(x;2;2;2;2)
3	[1, 11]	"Short": $\mu_{3.1}(x)$=*TRAPMF*(x; 1; 1; 8; 10) "Tall": $\mu_{3.2}(x)$=*TRAPMF*(x; 8; 10; 11;11)
4	[1, 6]	"Married": $\mu_{4.1}(x)$=*TRAPMF*(x;1; 1; 2; 2) "Single": $\mu_{4.2}(x)$=*TRAPMF*(x; 3; 3; 6; 6)
5	[0, 200000]	"Short": $\mu_{5.1}(x)$=*PIMF*(x; 0; 0; 9000; 12000) "Medium height": $\mu_{5.2}(x)$= =*PIMF*(x; 9000; 12000; 70000; 90000) "Tall": $\mu_{5.3}(x)$= =*PIMF*(x; 70000; 90000; 200000; 200000)
6	[0, 100]	"Short": $\mu_{6.1}(x)$=*PIMF*(x;0.0;0.0;29.9;40.3) "Medium height": $\mu_{6.2}(x)$= =*GAUSSMF*(x;2.5;40.0) "Tall": $\mu_{6.3}(x)$=*PIMF*(x;40.2;50.1;100.0;100.0)
7	[1, 6]	"Nonstandard": $\mu_{7.1}(x)$=*TRAPMF*(x; 1; 1; 5; 6) "Standard": $\mu_{7.2}(x)$=*TRAPMF*(x; 5; 6; 6; 6)
8	[1, 2]	"Caucasoid": $\mu_{8.1}(x)$=*TRAPMF*(x; 1; 1; 1; 1) "Negroid": $\mu_{8.2}(x)$=*TRAPMF*(x; 2; 2; 2; 2)
9	[0, 9]	"No": $\mu_{9.1}(x)$=*TRAPMF*(x; 0; 0; 0; 0) "Yes": $\mu_{9.2}(x)$=*TRAPMF*(x; 1; 1; 9; 9)
10	[0, 70]	"Own transport": $\mu_{10.1}(x)$= =*TRAPMF*(x; 0; 0; 20; 20) "Public transport": $\mu_{10.2}(x)$= =*TRAPMF*(x; 30; 30; 36; 36) "On foot": $\mu_{10.3}(x)$=*TRAPMF*(x; 40; 40; 50; 50)
11	[1, 2359]	"Night": $\mu_{11.1}(x)$=*PIMF*(x; 1; 1; 530; 630) "Morning": $\mu_{11.2}(x)$= =*PIMF*(x; 530; 630; 800; 900) "Day": $\mu_{11.3}(x)$=*PIMF*(x; 800; 900; 2359; 2359)
12	[1, 119]	"Short duration": $\mu_{12.1}(x)$=*PIMF*(x; 1; 1; 10; 15) "Average duration": $\mu_{12..2}(x)$= =*PIMF*(x; 10; 15; 35; 45) "Prolonged": $\mu_{12.3}(x)$=*PIMF*(x; 35; 45; 120; 120)
13	[2, 5]	N.A. (the value is only used for deleting inadmissible records from microfiles)

Fuzzy model of the military group was built using GA with parameters given in Table 4.

Table 4

Parameters of the genetic algorithm for building a fuzzy model of the military group

Parameter	Value
Population size, μ	100
Number of parental couples, λ	20
Crossing probability, p_c	1.000
Mutation probability, p_m	0.050
Selection tournament size	10
Number of algorithm starts	10
Number of generations in each start	100
Relative reliability threshold, γ	0.750
Bearer threshold, κ	0.001

Thus, after extracting records with values not belonging to the bearers of the basic linguistic variables L_j, j=1,..., 13 from the microfiles **M** and $\mathbf{M}^{aux}$, the main microfile **M** began to contain 565,243 records of which 3,992 records were vital and the auxiliary microfile $\mathbf{M}^{aux}$ began to contain 3,205,478 records of which 14,263 records were vital.

5. Results of an experiment of testing the information technology providing group anonymity

As a result of the use of GA with parameters given in Table 4, a fuzzy model of a military group was constructed. It consisted of the rules given together with their characteristics in Table 5. It is worthwhile to note that in all the rules, the linguistic variable "Means of transportation to work" is represented by the meaning of "On foot" which, obviously, is a characteristic feature of soldiers residing in barracks.

Table 5

Rules of the fuzzy model of the military group

Rule	*DF*	*RCF*	κ
(1, 0, 0, 2, 2, 0, 0, 0, 0, 3, 1, 1)	0.032	0.755	0.032
(1, 0, 0, 0, 0, 3, 0, 0, 0, 3, 1, 0)	0,031	0.787	0.031
(1, 0, 0, 0, 1, 3, 2, 0, 1, 3, 0, 0)	0.012	0.801	0.012
(1, 0, 0, 0, 2, 0, 1, 1, 1, 3, 1, 1)	0.010	0.781	0.010
(1, 0, 0, 0, 1, 3, 2, 1, 0, 3, 0, 0)	0.012	0.851	0.012
(1, 1, 0, 0, 2, 0, 0, 0, 0, 3, 1, 1)	0.034	0.840	0.034
(1, 1, 0, 2, 0, 0, 2, 0, 0, 3, 1, 1)	0.025	0.765	0.025
(1, 1, 0, 2, 1, 3, 0, 0, 0, 3, 2, 0)	0.018	0.931	0.018
(1, 1, 0, 0, 1, 3, 0, 0, 1, 3, 2, 0)	0.017	0.915	0.018
(1, 1, 0, 0, 0, 0, 2, 1, 0, 3, 1, 1)	0.025	0.754	0.026
(1, 1, 0, 2, 2, 0, 0, 1, 0, 3, 1, 0)	0.032	0.751	0.032
(1, 0, 1, 2, 1, 3, 0, 0, 0, 3, 2, 1)	0.018	0.951	0.018
(1, 0, 1, 0, 1, 3, 0, 0, 1, 3, 2, 0)	0.019	0.767	0.019
(1, 1, 1, 0, 1, 3, 2, 0, 0, 3, 2, 0)	0.009	1.876	0.009
(1, 1, 1, 0, 1, 3, 1, 0, 0, 3, 2, 1)	0.008	0.761	0.009
(1, 1, 1, 2, 1, 3, 2, 0, 0, 3, 0, 1)	0.010	1.325	0.010
(1, 1, 1, 2, 0, 3, 2, 0, 0, 3, 2, 1)	0.026	0.767	0.026
(1, 1, 2, 0, 2, 0, 0, 0, 0, 3, 1, 0)	0.002	0.914	0.002

It is advisable to analyze use of the constructed fuzzy model to identify the TMR outliers in the main microfile **M** separately for each state. For example, for the State of New York, the target representation of the microfile **M** relative to the military group, $\mathbf{q}_{\text{NY}}$, and the auxiliary target representation of the $\mathbf{M}^{aux}$ microfile for the military group, $\mathbf{q}_{\text{NY}}^{aux}$, are shown in Fig. 7. The abscissa axis in the figure corresponds to the geographical units in which the military serve (values 1–33 correspond to the values of the parameter attribute 3600100–3603300, respectively, values 34–38 correspond to the values of the parameter attribute 3603700–3604100, respectively), and the ordinate axis corresponds to the number of the military that serve there.

Application of MMTT with parameter α=0.01 to the given TMR has allowed the user, junior analyst, to get a set of indices $OUT(\mathbf{q}_{NY})$={5, 7, 9, 11, 12, 17, 18, 20, 29, 31, 35, 36 , 37}. Only two indices in this set correspond to the real military bases [36], so the user, junior analyst, has left only them for further analysis: $OUT(\mathbf{q}_{\text{NY}})$={5, 29}. Index 5 corresponds to Fort Drum and Index 29 to Military Academy West Point.

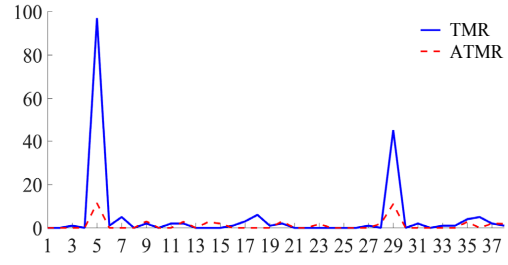

Fig. 7. TMR and auxiliary TMR in relation to the military group

Similar considerations after applying MMTT with parameter α=0.01 to the auxiliary TMR have allowed the user, junior analyst, to get a set $OUT(\mathbf{q}_{\mathrm{NY}}^{aux})=\{5, 29\}$. Equality of the two sets directly indicates the possibility of violating group anonymity of the military group even if the Occupation attribute is removed from the microfile. A similar analysis for the rest of the US states where the military number exceeds 0.5% of the total military number in **M** is given in Table 6.

Table 6

Results of applying the fuzzy model of the military group to the microfile M

State	Number of outliers in TMR	Number of outliers in TMR absent in ATMR	Number of outliers in ATMR	Number of outliers in ATMR absent in TMR
Alabama	2	2	1	1
Alaska	2	0	2	0
Arizona	4	1	4	1
Washington	4	1	3	0
Virginia	7	4	4	1
Hawaii	1	0	1	0
Georgia	7	3	4	0
Illinois	2	1	2	1
California	3	1	2	0
Kansas	2	2	0	0
Kentucky	2	1	1	0
Colorado	2	0	2	0
Connecticut	1	0	2	1
Louisiana	4	4	0	0
Maryland	3	2	1	0
Mississippi	1	0	1	0
Missouri	2	2	0	0
Nevada	1	0	1	0
New Jersey	2	2	0	0
New Mexico	2	2	0	0
New York	2	0	2	0
Ohio	2	1	3	2
Oklahoma	3	2	1	0
South Carolina	4	1	3	0
North Carolina	3	1	2	0
Texas	6	1	5	0
Florida	7	5	3	1
Total	81	39	50	8

The matrix of inconsistencies for the example is

$$\mathbf{Z}=\begin{pmatrix}48 & 39\\ 8 & 564\end{pmatrix}.$$

Based on this matrix, the user, junior analyst, has calculated the metric of adequacy of the fuzzy model (11) which was *MB*=55.067 which indicates high adequacy of the model.

Thus, to ensure reliable group anonymity, it is not enough just to remove the vital attribute Occupation from the microfile **M.** It is also necessary to apply MA to mask outliers of the auxiliary TMR $\mathbf{q}^{aux}$. Let us consider an example of the IT application to the state of New York.

The user, junior analyst, has imposed on the 5th and 29th count of TMR fuzzy constraints with membership function

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

and the function of adaptability (5) for the example has assumed this form:

$$f(U)=\frac{299-\sum_{i=1}^{Q}\sum_{k=1}^{13}\operatorname{sign}\left|\mathbf{M}_{u_{i1}}(u_{i2},w_{b_k})-\mathbf{M}_{u_{i3}}(u_{i4},w_{b_k})\right|}{299}\times$$
$$\times ZMF\left(q_{NY_5}^{aux^*}(U),2,12\right)\cdot ZMF\left(q_{NY_{29}}^{aux^*}(U),2,11\right)\times$$
$$\times\frac{1}{1+e^{\frac{1}{2}(Q-25)}},$$

where $C_{\max}$=299; w_{bk} is the k-th basic attribute, k=1,..., 13.

The user, junior analyst, has selected parameters of MA for masking outliers in the ATMR given in Table 7. In this case, mutation probability has increased 10-fold when the mean-square deviation of values of the adaptability function of individuals in some population became less than 0.03.

Table 7

Parameters of the mimetic algorithm for solving the TPGA

Parameter	Value
Population size, μ	100
Number of parental couples, λ	20
Crossing probability, p_c	1.000
Mutation probability, p_m	0.001
Local search parameter, p_{mem}	0.750
Selection tournament size	5
Compatibility threshold, α_{comp}	0.500
Sensibility threshold, K_{out}	0.000
Distortion threshold, K_{dist}	0.250
Algorithm start number	10
Number of generations in each start	1,000

The MA functioning with the specified parameters has resulted in 1000 individuals of the last generations of

each start of which 983 met requirements for thresholds of compatibility, sensitivity and distortions. Average value of the metric (1) for all 983 solutions was 62.518, that is, anonymity was ensured by distorting 62.518/(13·1380924)≈ ≈ 0.0003 % values of the microfile attributes.

Solution of $\mathbf{q}^{aux*}$ with the smallest value of 53 for the metric (1) is given in Fig. 8.

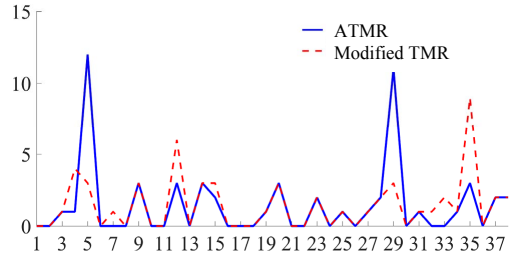

Fig. 8. Auxiliary TMR $\mathbf{q}^{aux}$ and modified auxiliary TMR $\mathbf{q}^{aux*}$

Outliers in $\mathbf{q}^{aux^*}$ obtained with the help of MMTT correspond to items with indices 12 and 35, i. e. OUT($\mathbf{q}^{aux}$) ∩OUT ($\mathbf{q}^{aux^*}$)=∅.

6. Discussion of the results of solving the TPGA with the help of the information technology

The process of applying the proposed information technology to provide group anonymity of data in an automated mode has been demonstrated in this experimental study. It was shown that the developed IT satisfies the above requirements as it:

– makes it possible to build models of respondent groups in a microfile by setting parameter and vital attributes and their values;

– makes it possible to build fuzzy models of respondent groups in a microfile based on genetic algorithms with flexibly specified parameters;

– makes it possible to provide group anonymity of data by means of a mimetic algorithm introducing small distortions into data.

At the same time, various operations and actions within the process of providing data group anonymity are performed by users with different roles which enables an increase in efficiency of preparing microfiles for publication due to division of labor and specialization of individual specialists. Such operations include harmonization of microfiles, selection of an auxiliary microfile, parameterization of algorithms and methods for providing group anonymity, making a decision on extraction of vital attributes and final completion of the process of data anonymization.

The high level of reliability and security of primary data is provided by combining all IT components into a local net with a limited access.

Solution of the experiment problem with a team of five specialists has taken 7 hrs 50 min while solution of a similar problem with the help of the IT described in [20] has taken 19 hours 45 minutes, that is 2.5 times longer. Such an increase in speed of preparation of microfiles to publication is due to organization of an effective interaction of users of various specialties and integration into the IT methods that provide a more effective solution of TPGA compared with those described in [20].

Writing of a manual of the most effective use of the IT, in particular, recommendations on selection of GA and MA parameters for the user, junior analyst, and criteria of completion of the anonymization process for the user, senior analyst, requires additional studies. Availability of such a manual will significantly extend the range of the IT users and enable preparation of data for publication by organizations that are not specialized in statistical processing.

7. Conclusions

1. A three-level client-server architecture of an information technology providing data group anonymity was proposed in which clients, application servers and databases are united into a local network. The technology takes into account the possibility of violating the group anonymity in conditions of accessibility to the auxiliary microfile which ensures an increase in the level of data security.

2. A conceptual model of a relational database for the proposed IT was developed. It contains all essences of the process of providing data group anonymity and reflects relations between them. Key fragments of the constructed data model were presented. They correspond to the essences of the microfile and its target representation, fuzzy models of the groups in the microfile and solution of the problem of providing group anonymity.

3. Implementation of the technology based on the Java Enterprise Edition 8 platform, Oracle GlassFish application server, MySQL database server and SciLab engineering calculation system was considered. Interaction of the technology applications with clients' workstations and functions written in the SciLab system was described. The proposed implementation satisfies the requirements put forward to the information technology.

4. Practical application of the information technology was illustrated by solution of a task of anonymizing a group of military personnel based on real data given in Observation of the American Society 2013. It has been established that application of the technology makes it possible to speed up 2.5 times the process of preparation of microfiles for publication by a team of five specialists.

References

1. Duncan G. T., Elliot M., Salazar-González J.-J. Statistical Confidentiality. Principles and Practice. Springer-Verlag, 2011. 212 p. doi: https://doi.org/10.1007/978-1-4419-7802-8
2. A Terminology for Talking About Privacy by Data Minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management, Version v0.34. URL: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml
3. Chertov O. R., Tavrov D. Y. Providing group anonymity as a part of CSID data process // Shtuchnyi intelekt. 2017. Issue 3-4. P. 127–138.

4. Chertov O., Tavrov D. Improving efficiency of providing data group anonymity by automating data modification quality evaluation // Eastern-European Journal of Enterprise Technologies. 2017. Vol. 5, Issue 4 (89). P. 31–39. doi: https://doi.org/10.15587/1729-4061.2017.113046
5. Chertov O., Tavrov D. Microfiles as a Potential Source of Confidential Information Leakage // Studies in Computational Intelligence. 2014. P. 87–114. doi: https://doi.org/10.1007/978-3-319-08624-8_4
6. Tavrov D., Chertov O. Evolutionary approach to violating group anonymity using third-party data // SpringerPlus. 2016. Vol. 5, Issue 1. doi: https://doi.org/10.1186/s40064-016-1692-9
7. Butz M. V. Learning Classifier Systems // Springer Handbook of Computational Intelligence. 2015. P. 961–981. doi: https://doi.org/10.1007/978-3-662-43505-2_47
8. Holland J. H. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975. 183 p.
9. Valenzuela-Rendón M. The Fuzzy Classifier System: Motivations and First Results // Proceedings of Parallel Solving from Nature (PPSN II). 1991. P. 330–334.
10. Smith S. F. A Learning System Based on Genetic Adaptive Algorithms. Pittsburgh: University of Pittsburgh, 1980. 214 p.
11. Overview on evolutionary subgroup discovery: analysis of the suitability and potential of the search performed by evolutionary algorithms / Carmona C. J., González P., del Jesus M. J., Herrera F // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014. Vol. 4, Issue 2. P. 87–103. doi: https://doi.org/10.1002/widm.1118
12. Ishibuchi H., Nakashima T., Murata T. Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems // IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics). 1999. Vol. 29, Issue 5. P. 601–618. doi: https://doi.org/10.1109/3477.790443
13. μ-ARGUS Version 5.1.3. User's Manual / Hundepool A., de Wolf P.-P., Bakker J., Reedijk A., Franconi L. et. al. 2018. URL: http://neon.vb.cbs.nl/casc/Software/MUmanual5.1.3.pdf
14. Angiuli O., Waldo J. Statistical Tradeoffs Between Generalization and Suppression in the De-Identification of Large-Scale Data Sets // 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC). 2016. doi: https://doi.org/10.1109/compsac.2016.198
15. Sweeney L. K-Anonymity: A Model for Protecting Privacy // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2002. Vol. 10, Issue 05. P. 557–570. doi: https://doi.org/10.1142/s0218488502001648
16. Fienberg S., McIntyre J. Data Swapping: Variations on a Theme by Dalenius and Reiss // Journal of Official Statistics. 2005. Vol. 21, Issue 2. P. 309–323.
17. Evfimievski A. Randomization in privacy preserving data mining // ACM SIGKDD Explorations Newsletter. 2002. Vol. 4, Issue 2. P. 43–48. doi: https://doi.org/10.1145/772862.772869
18. Templ M. Statistical Disclosure Control for Microdata Using the R-package sdcMicro // Transactions on Data Privacy. 2008. Vol. 1, Issue 2. P. 67–85.
19. Domingo-Ferrer J., Mateo-Sanz J. M. Practical data-oriented microaggregation for statistical disclosure control // IEEE Transactions on Knowledge and Data Engineering. 2002. Vol. 14, Issue 1. P. 189–201. doi: https://doi.org/10.1109/69.979982
20. Chertov O. R. Minimizatsiya spotvoren pry formuvanni mikrofailu z zamaskovanymy danymy // Visnyk Skhidnoukrainskoho natsionalnoho universytetu im. V. Dalia. 2012. Issue 8 (179). P. 240–246.
21. Chertov O., Tavrov D. Providing Group Anonymity Using Wavelet Transform // Lecture Notes in Computer Science. 2012. P. 25–36. doi: https://doi.org/10.1007/978-3-642-25704-9_5
22. Chertov O., Tavrov D. Two-Phase Memetic Modifying Transformation for Solving the Task of Providing Group Anonymity // Studies in Fuzziness and Soft Computing. 2016. P. 239–253. doi: https://doi.org/10.1007/978-3-319-32229-2_17
23. Zadeh L. A. Toward a restriction-centered theory of truth and meaning (RCT) // Information Sciences. 2013. Vol. 248. P. 1–14. doi: https://doi.org/10.1016/j.ins.2013.06.003
24. Neri F., Cotta C. A Primer on Memetic Algorithms // Studies in Computational Intelligence. 2012. P. 43–52. doi: https://doi.org/10.1007/978-3-642-23247-3_4
25. Goldberg D. E., Korb B., Deb K. Messy Genetic Algorithms: Motivation, Analysis, and First Results // Complex Systems. 1989. Vol. 3. P. 493–530.
26. Syswerda G. Schedule Optimization Using Genetic Algorithms // Handbook of Genetic Algorithms. New York: Van Nostrand Reinhold, 1991. P. 332–349.
27. Eiben A. E., Smith J. E. Introduction to Evolutionary Computing. Springer-Verlag, 2015. 287 p. doi: https://doi.org/10.1007/978-3-662-44874-8
28. Brindle A. Genetic Algorithms for Function Optimization. Edmonton: University of Alberta, 1981. 193 p.
29. Zadeh L. A. The concept of a linguistic variable and its application to approximate reasoning – I // Information Sciences. 1975. Vol. 8, Issue 3. P. 199–249. doi: https://doi.org/10.1016/0020-0255(75)90036-5

30. Wrobel S. An algorithm for multi-relational discovery of subgroups // Lecture Notes in Computer Science. 1997. P. 78–87. doi: https://doi.org/10.1007/3-540-63223-9_108
31. Lavrač N., Flach P., Zupan B. Rule Evaluation Measures: A Unifying View // Lecture Notes in Computer Science. 1999. P. 174–185. doi: https://doi.org/10.1007/3-540-48751-4_17
32. Selecting fuzzy if-then rules for classification problems using genetic algorithms / Ishibuchi H., Nozaki K., Yamamoto N., Tanaka H. // IEEE Transactions on Fuzzy Systems. 1995. Vol. 3, Issue 3. P. 260–270. doi: https://doi.org/10.1109/91.413232
33. Syswerda G. Uniform Crossover in Genetic Algorithms // Proceedings of the 3rd International Conference on Genetic Algorithms. Morgan Kaufmann Publishers Inc., 1989. P. 2–9.
34. Olivetti E., Greiner S., Avesani P. Statistical independence for the evaluation of classifier-based diagnosis // Brain Informatics. 2014. Vol. 2, Issue 1. P. 13–19. doi: https://doi.org/10.1007/s40708-014-0007-6
35. Integrated Public Use Microdata Series, Version 8.0 [Dataset] / Ruggles S., Flood S., Goeken R., Grover J., Meyer E., Pacas J., Sobek M. Minneapolis: University of Minnesota, 2018. URL: https://usa.ipums.org/usa/
36. 2011 Demographics. Profile of the Military Community. 2012. URL: http://www.militaryonesource.mil/12038/MOS/Reports/2011_Demographics_Report.pdf