Data Group Anonymity: General Approach

Oleg Chertov
Applied Mathematics Department
NTUU "Kyiv Polytechnic Institute"
Kyiv, Ukraine
chertov@i.ua

Dan Tavrov
Applied Mathematics Department
NTUU "Kyiv Polytechnic Institute"
Kyiv, Ukraine
dan.tavrov@i.ua

***Abstract*—In the recent time, the problem of protecting privacy in statistical data before they are published has become a pressing one. Many reliable studies have been accomplished, and loads of solutions have been proposed.**

Though, all these researches take into consideration only the problem of protecting individual privacy, i.e., privacy of a single person, household, etc. In our previous articles, we addressed a completely new type of anonymity problems. We introduced a novel kind of anonymity to achieve in statistical data and called it group anonymity.

In this paper, we aim at summarizing and generalizing our previous results, propose a complete mathematical description of how to provide group anonymity, and illustrate it with a couple of real-life examples.

Keywords-group anonymity; microfiles; wavelet transform

I. INTRODUCTION

Throughout mankind's history, people always collected large amounts of demographical data. Though, until the very recent time, such huge data sets used to be inaccessible for publicity. And what is more, even if some potential intruder got an access to such paper-written data, it would be way too hard for him to analyze them properly!

But, as information technologies develop more, a greater number of specialists (to wide extent) gain access to large statistical datasets to perform various kinds of analysis. For that matter, different data mining systems help to determine data features, patterns, and properties.

As a matter of fact, in today world, in many cases population census datasets (usually referred to as *microfiles*) contain this or that kind of sensitive information about respondents. Disclosing such information can violate a person's privacy, so convenient precautions should be taken beforehand.

For many years now, mostly every paper in major of providing data anonymity deals with a problem of protecting an individual's privacy within a statistical dataset. As opposed to it, we have previously introduced a totally new kind of anonymity in a microfile which we called *group anonymity*. In this paper, we aim at gathering and systematizing all our works published in the previous years. Also, we would like to generalize our previous approaches and propose an integrated survey of group anonymity problem.

II. RELATED WORK

A. Individual Anonymity

We understand by *individual data anonymity* a property of information about an individual to be unidentifiable within a dataset.

There exist two basic ways to protect information about a single person. The first one is actually protecting the data in its formal sense, using data encryption, or simply restricting access to them. Of course, this technique is of no interest to statistics and affiliated fields.

The other approach lies in modifying initial microfile data such way that it is still useful for the majority of statistical researches, but is protected enough to conceal any sensitive information about a particular respondent. Methods and algorithms for achieving this are commonly known as *privacy preserving data publishing (PPDP)* techniques. The Free Haven Project [1] provides a very well prepared anonymity bibliography concerning these topics.

In [2], the authors investigated all main methods used in PPDP, and introduced a systematic view of them. In this subsection, we will only slightly characterize the most popular PPDP methods of providing individual data anonymity. These methods are also widely known as *statistical disclosure control (SDC)* techniques.

All SDC methods fall into two categories. They can be either *perturbative* or *non-perturbative*. The first ones achieve data anonymity by introducing some data distortion, whereas the other ones anonymize the data without altering them.

Possibly the simplest perturbative proposition is to add some noise to initial dataset [3]. This is called *data randomization.* If this noise is independent of the values in a microfile, and is relatively small, then it is possible to perform statistical analysis which yields rather close results compared to those ones obtained using initial dataset. Though, this solution is not quite efficient. As it was shown in [4], if there are other sources available aside from our microfile with intersecting information, it will be very possible to violate privacy.

Another option is to reach data *k-anonymity*. The core of this approach is to somehow ensure that all combinations of microfile attribute values are associated with at least *k* respondents. This result can be obtained using various methods [5, 6].

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Yet another technique is to *swap* confidential microfile attribute values between different individuals [7].

Non-perturbative SDC methods are mainly represented by *data recoding* (data enlargement) and *data suppression* (removing the data from the original microfile) [6].

In previous years, novel methods evolved, e.g., matrix decomposition [8], or factorization [9]. But, all of them aim at preserving individual privacy only.

B. Group Anonymity

Despite the fact that PPDP field is developing rather rapidly, there exists another, completely different privacy issue which hasn't been studied well enough yet. Speaking more precisely, it is another kind of anonymity to be achieved in a microfile.

We called this kind of anonymity *group anonymity*. The formal definition will be given further on in this paper, but in a way this kind of anonymity aims at protecting such data features and patterns which cannot be determined by analyzing standalone respondents.

The problem of providing group anonymity was initially addressed in [10]. Though, there has not been proposed any feasible solution to it then.

In [11, 12], we presented a rather effective method for solving some particular group anonymity tasks. We showed its main features, and discussed several real-life practical examples.

The most complete survey of group anonymity tasks and their solutions as of time this paper is being written is [13]. There, we tried to gather up all existing works of ours in one place, and also added new examples that reflect interesting peculiarities of our method. Still, [13] lacks a systematized view and reminds more of a collection of separate articles rather than of an integrated study.

That is why in this paper we set a task of embedding all known approaches to solving group anonymity problem into complete and consistent group anonymity theory.

III. Formal Definitions

To start with, let us propose some necessary definitions.

Definition 1. By *microdata* we will understand various data about respondents (which might equally be persons, households, enterprises, and so on).

Definition 2. Respectively, we will consider *a microfile* to be microdata reduced to one file of attributive records concerning each single respondent.

A microfile can be without any complications presented in a matrix form. In such a matrix **M**, each row corresponds to a particular respondent, and each column stands for a specific attribute. The matrix itself is shown in Table I.

TABLE I. Microfile Data in a Matrix Form

Respondents	Attributes: u_1	u_2	…	u_η
r_1	ω_{11}	ω_{12}	…	$\omega_{1\eta}$
r_2	ω_{21}	ω_{22}	…	$\omega_{2\eta}$
…	…	…	…	…
r_μ	$\omega_{\mu 1}$	$\omega_{\mu 2}$	…	$\omega_{\mu\eta}$

In such a matrix, we can define different classes of attributes.

Definition 3. An identifier is a microfile attribute which unambiguously determines a certain respondent in a microfile.

From a privacy protection point of view, identifiers are the most security-intensive attributes. The only possible way to prevent privacy violation is to completely eliminate them from a microfile. That is why, we will further on presume that a microfile is always *de-personalized*, i.e., it does not contain any identifiers.

In terms of group anonymity problem, we need to define such attributes whose distribution is of a big privacy concern and has to be thoroughly considered.

Definition 4. We will call an element $s_k^{(v)} \in S_v$, $k = \overline{1, l_v}$, $l_v \le \mu$, where S_v is a subset of a Cartesian product $u_{v_1} \times u_{v_2} \times ... \times u_{v_t}$ (see Table I), *a vital value combination*. Each element of $s_k^{(v)}$ is called *a vital value*. Each u_{v_j}, $j = \overline{1, t}$ is called *a vital attribute*.

In other words, vital attributes reflect characteristic properties needed to define a subset of respondents to be protected.

But, it is always convenient to present multidimensional data in a one-dimensional form to simplify its modification. To be able to accomplish that, we have to define yet another class of attributes.

Definition 5. We will call an element $s_k^{(p)} \in S_p$, $k = \overline{1, l_p}$, $l_p \le \mu$, where S_p is a subset of microfile data elements corresponding to the p^{th} attribute, *a parameter value*. The attribute itself is called *a parameter attribute*.

Parameter values are usually used to somehow arrange microfile data in a particular order. In most cases, resultant data representation contains some sensitive information which is highly recommended to be protected. (We will delve into this problem in the next section.)

Definition 6. A group $G(V, P)$ is a set of attributes consisting of several vital attributes $V = \{V_1, V_2, ..., V_l\}$ and a parameter attribute P, $P \ne V_j$, $j = 1, ..., l$.

Now, we can formally define a group anonymity task.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

Group Anonymity Definition. The task of *providing data group anonymity* lies in modifying initial dataset for each group $G_i(V_i, P_i)$, $i = 1,...,k$ such way that sensitive data features become totally confided.

In the next section, we will propose a generic algorithm for providing group anonymity in some most common practical cases.

IV. General Approach to Providing Group Anonymity

According to the Group Anonymity Definition, initial dataset **M** should be perturbed separately for each group to ensure protecting specific features for each of them.

Before performing any data modifications, it is always necessary to preliminarily define what features of a particular group need to be hidden. So, we need to somehow transform initial matrix into another representation useful for such identification. Besides, this representation should also provide more explicit view of how to modify the microfile to achieve needed group features.

All this leads to the following definitions.

Definition 7. We will understand by *a goal representation* Ω (**M**, *G*) of a dataset **M** with respect to a group *G* such a dataset (which could be of any dimension) that represents particular features of a group within initial microfile in a way appropriate for providing group anonymity.

We will discuss different forms of goal representations a bit later on in this section.

Having obtained goal representation of a microfile dataset, it is almost always possible to *modify* it such way that security-intensive peculiarities of a dataset become concealed. In this case, it is said we obtain *a modified goal representation* Ω' (**M**, *G*) of initial dataset **M**.

After that, we need to somehow map our modified goal representation to initial dataset resulting in *a modified microdata* $\mathbf{M}^*$. Of course, it is not necessary that such data modifications lead to any feasible solution. But, as we will discuss it in the next subsections, if to pick specific mappings and data representations, it is possible to provide group anonymity in any microfile.

So, a generic scheme of providing group anonymity is as follows:

1) Construct a (depersonalized) microfile **M** *representing statistical data to be processed.*

2) Define one or several groups $G_i(V_i, P_i)$, $i = 1,...,k$ *representing categories of respondents to be protected.*

3) For each i from 1 *to k*:

a) Choosing data representation: Pick a goal representation Ω_i (**M**, G_i) for a group $G_i(V_i, P_i)$.

b) Performing data mapping: Define a mapping function Υ: **M** $\rightarrow$ Ω_i (**M**, G_i) (called *goal mapping function*) and obtain needed goal representation of a dataset.

c) Performing goal representation's modification: Define a functional Ξ: Ω_i (**M**, G_i) $\rightarrow$ Ω'_i (**M**, G_i) (also called *modifying functional*) and obtain a modified goal representation.

d) Obtaining the modified microfile. Define an *inverse goal mapping function* Υ^{-1}: Ω'_i (**M**, G_i) $\rightarrow$ $\mathbf{M}^*$ and obtain a modified microfile.

4) Prepare the modified microfile for publishing.

Now, let us discuss some of these algorithm steps a bit in detail.

A. Different Ways to Construct a Goal Representation

In general, each particular case demands developing certain data representation models to suit the stated requirements the best way. Although, there are loads of real-life examples where some common models might be applied with a reasonable effect.

In our previous works, we drew a particular attention to one special data goal representation, namely, *a goal signal*. The goal signal is a one-dimensional numerical array $\theta = (\theta_1, \theta_2, ..., \theta_m)$ representing statistical features of a group. It can consist of values obtained in different ways, but we will defer this discussion for some paragraphs.

In the meantime, let us try to figure out what particular features of a goal signal might turn out to be security-intensive. To be able to do that, we need to consider its graphical representation which we will call *a goal chart*. In [13], we summarized the most important goal chart features and proposed some approaches to modifying them. In order not to repeat ourselves, we will only outline some of them:

1) Extremums. In most cases, it is the most sensitive information; we need to transit such extremums from one signal position to another (or, which is also completely convenient, create some new extremums, so that initial ones just "dissolve").

2) Statistical features. Such features as signal mean value and standard deviation might be of a big importance, unless a corresponding parameter attribute is nominal (it will become clear why in a short time).

3) Frequency spectrum. This feature might be rather interesting if a goal signal contains some parts repeated cyclically.

Coming from a particular aim to be achieved, one can choose the most suitable modifying functional Ξ to redistribute the goal signal.

Let us understand how a goal signal can be constructed in some widely spread real-life group anonymity problems.

In many cases, we can count up all the respondents in a group with a certain pair of vital value combination and a parameter value, and arrange them in any order proper for a parameter attribute. For instance, if parameter values stand for a person's age, and vital value combinations reflect his or her yearly income, then we will obtain a goal signal representing quantities of people with a certain income distributed by their

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

age. In some situations, this distribution could lead to unveiling some restricted information, so, a group anonymity problem would evidently arise.

Such a goal signal is called *a quantity signal* $q = (q_1, q_2, ..., q_m)$. It provides a quantitative statistical distribution of group members from initial microfile.

Though, as it was shown in [12], sometimes absolute quantities do not reflect real situations, because they do not take into account all the information given in a microfile. A much better solution for such cases is to build up *a concentration signal*:

$$c = (c_1, c_2, ..., c_m) \equiv \left(\frac{q_1}{\rho_1}, \frac{q_2}{\rho_2}, ..., \frac{q_m}{\rho_m} \right). \quad (1)$$

In (1), ρ_i, $i = 1, ..., m$ stand for the quantities of respondents in a microfile from a group defined by a superset for our vital value combinations. This can be explained on a simple example. Information about people with AIDS distributed by regions of a state can be valid only if it is represented in a relative form. In this case, q_i would stand for a number of ill people in the i^{th} region, whereas ρ_i could possibly stand for the whole number of people in the i^{th} region.

And yet another form of a goal signal comes to light when processing comparative data. A representative example is as follows: if we know concentration signals built separately for young males of military age and young females of the same age, then, maximums in their difference might point at some restricted military bases.

In such cases, we deal with two concentration signals $c^{(1)} = (c_1^{(1)}, c_2^{(1)}, ..., c_m^{(1)})$ (also called *a main concentration signal*) and $c^{(2)} = (c_1^{(2)}, c_2^{(2)}, ..., c_m^{(2)})$ (*a subordinate concentration signal*). Then, the goal signal takes a form of *a concentration difference signal* $\delta = (c_1^{(1)} - c_1^{(2)}, c_2^{(1)} - c_2^{(2)}, ..., c_m^{(1)} - c_m^{(2)})$.

In the next subsection, we will address the problem of picking a suitable modifying functional, and also consider one of its possible forms already successfully applied in our previous papers.

B. Picking Appropriate Modifying Functional

Once again, there can be created way too many unlike modifying functionals, each of them taking into consideration these or those requirements set by a concrete group anonymity problem definition. In this subsection, we will look a bit in detail at two such functionals.

So, let us pay attention to the first goal chart feature stated previously, which is in most cases the feature we would like to protect. Let us discuss the problem of altering extremums in an initial goal chart.

In general, we might perform this operation quite arbitrarily. The particular scheme of such extremums redistribution would generally depend on the quantity signal nature, sense of parameter values, and correct data interpreting. But, as things usually happen in statistics, we might as well want to guarantee that data utility wouldn't reduce much. By *data utility preserving* we will understand the situation when the modified goal signal yields similar, or even the same, results when performing particular types of statistical (but not exclusively) analysis.

Obviously, altering the goal signal completely off-hand without any additional precautions taken wouldn't be very convenient from the data utility preserving point of view. Hopefully, there exist two quite dissimilar, thought powerful techniques for preserving some goal chart features.

The first one was proposed in [14]. Its main idea is to *normalize* the output signal using such transformation that both mean value and standard deviation of a signal remain stable. Surely, this is not ideal utility preserving. But, the signal obtained this way at least yields the same results when performing basic statistical analysis. So, the formula goes as follows:

$$\theta^* = (\theta + \frac{\sigma^*}{\sigma} \cdot \varepsilon - \varepsilon^*) \cdot \frac{\sigma}{\sigma^*}. \quad (2)$$

In (2), $\varepsilon = \frac{1}{m} \sum_{i=1}^{m} \theta_i$, $\varepsilon^* = \frac{1}{m} \sum_{i=1}^{m} \theta_i^*$, $\sigma = \sqrt{\frac{\sum_{i=1}^{m} (\theta_i - \varepsilon)^2}{m-1}}$, $\sigma^* = \sqrt{\frac{\sum_{i=1}^{m} (\theta_i^* - \varepsilon^*)^2}{m-1}}$.

The second method of modifying the signal was initially proposed in [11], and was later on developed in [12, 13]. Its basic idea lies in applying *wavelet transform* to perturbing the signal, with some slight restrictions necessary for preserving data utility:

$$\theta(t) = \sum_i a_{k,i} \cdot \varphi_{k,i}(t) + \sum_{j=k}^{1} \sum_i d_{j,i} \cdot \psi_{j,i}(t). \quad (3)$$

In (3), $\varphi_{k,i}$ stands for shifted and sampled *scaling functions*, and $\psi_{j,i}$ represents shifted and sampled *wavelet functions*. As we showed in our previous researches, we can gain group anonymity by modifying *approximation coefficients* $a_{k,i}$. At the same time, if we don't modify *detail coefficients* $d_{j,i}$ we can preserve signal's frequency characteristics necessary for different kinds of statistical analysis.

More than that, we can always preserve the signal's mean value without any influence on its extremums:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

$$\theta^*_{fin} = \theta^*_{mod} \cdot \left(\sum_{i=1}^{m} \theta_i \Big/ \sum_{i=1}^{m} \theta^*_{mod\ i} \right). \quad (4)$$

In the next section, we will study several real-life practical examples, and will try to provide group anonymity for appropriate datasets. Until then, we won't delve deeper into wavelet transforms theory.

C. The Problem of Minimum Distortion when Applying Inverse Goal Mapping Function

Having obtained modified goal signal θ^*_{fin}, we have no other option but to modify our initial dataset **M**, so that its contents correspond to θ^*_{fin}.

It is obvious that, since group anonymity has been provided with respect to only a single respondent group, modifying the dataset **M** almost inevitably will lead to introducing some level of data distortion to it. In this subsection, we will try to minimize such distortion by picking sufficient inverse goal mapping functions.

At first, we need some more definitions.

Definition 8. We will call microfile **M** attributes *influential ones* if their distribution plays a great role for researchers.

Obviously, vital attributes are influential by definition.

Keeping in mind this definition, let us think over a particular procedure of mapping the modified goal signal θ^*_{fin} to a modified microfile $\mathbf{M}^*$. The most adequate solution, in our opinion, implies swapping parameter values between pairs of somewhat close respondents. We might interpret this operation as "transiting" respondents between two different groups (which is in fact the case).

But, an evident problem arises. We need to know how to define whether two respondents are "close" or not. This could be done if to measure such closeness using *influential metric* [13]:

$$InfM(r, r^*) = \sum_{p=1}^{n_{ord}} \zeta_p \left(\frac{r(I_p) - r^*(I_p)}{r(I_p) + r^*(I_p)} \right)^2 + \\ + \sum_{k=1}^{n_{nom}} \gamma_k \left(\chi\left(r(J_k), r^*(J_k) \right) \right)^2. \quad (5)$$

In (5), I_p stands for the p^{th} ordinal influential attribute (making a total of n_{ord}). Respectively, J_k stands for the k^{th} nominal influential attribute (making a total of n_{nom}). Functional $r(\cdot)$ stands for a record's r specified attribute value. Operator $\chi(v_1, v_2)$ is equal to χ_1 if values v_1 and v_2 represent one category, and χ_2, if it is not so. Coefficients ζ_p and γ_k should be taken coming from importance of a certain attribute (for those ones not to be changed at all they ought to be as big as possible, and for those ones that are not important they could be zero).

With the help of this metric, it is not too hard to outline the generic strategy of performing inverse data mapping. One needs to search for every pair of respondents yielding minimum influential metric value, and swap corresponding parameter values. This procedure should be carried out until the modified goal signal θ^*_{fin} is completely mapped to $\mathbf{M}^*$.

This strategy seems to be NP-hard, so, the problem of developing more computationally effective inverse goal mapping functions remains open.

V. Some Practical Examples of Providing Group Anonymity

In this subsection, we will discuss two practical examples built upon real data to show the proposed group anonymity providing technique in action.

According to the scheme introduced in Section IV, the first thing to accomplish is to compile a microfile representing the data we would like to work with. For both of our examples, we decided to take 5-Percent Public Use Microdata Sample Files provided by the U.S. Census Bureau [15] concerning the 2000 U.S. census of population and housing microfile data. But, since this dataset is huge, we decided to limit ourselves with analyzing the data on the state of California only.

The next step (once again, we will carry it out the same way for both examples) is to define group(s) to be protected. In this paper, we will follow [11], i.e. we will set a task of protecting military personnel distribution by the places they work at. Such a task has a very important practical meaning. The thing is that extremums in goal signals (both quantity and concentration ones) with a very high probability mark out the sites of military cantonments. In some cases, these cantonments aren't likely to become widely known (especially to some potential adversaries).

So, to complete the second step of our algorithm, we take "Military service" attribute as a vital one. This is a categorical attribute, with integer values ranging from 0 to 4. For our task definition, we decided to take one vital value, namely, "1" which stands for "Active duty".

But, we also need to pick an appropriate parameter attribute. Since we aim at redistributing military servicemen by different territories, we took "Place of Work Super-PUMA" as a parameter attribute. The values of this categorical attribute represent codes for Californian statistical areas. In order to simplify our problem a bit, we narrowed the set of this attribute's values down to the following ones: 06010, 06020, 06030, 06040, 06060, 06070, 06080, 06090, 06130, 06170, 06200, 06220, 06230, 06409, 06600, and 06700. All these area codes correspond to border, island, and coastal statistical areas.

From this point, we need to make a decision about the goal representation of our microdata. To show peculiarities of different kinds of such representations, we will discuss at least two of them in this section. The first one would be the quantity signal, and the other one would be its concentration analogue.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

A. Quantity Group Anonymity Problem

So, having all necessary attributes defined, it is not too hard to count up all the military men in each statistical area, and gather them up in a numerical array sorted in an ascending order by parameter values. In our case, this quantity signal looks as follows:

q=(19, 12, 153, 71, 13, 79, 7, 33, 16, 270, 812, 135, 241, 14, 60, 4337).

The graphical representation of this signal is presented in Fig. 1a.

As we can clearly see, there is a very huge extremum at the last signal position. So, we need to somehow eliminate it, but simultaneously preserve important signal features. In this example, we will use wavelet transforms to transit extremums to another region, so, according to the previous section, we will be able to preserve high-frequency signal spectrum.

As it was shown in [11], we need to change signal approximation coefficients in order to modify its distribution. To obtain approximation coefficients of any signal, we need to *decompose* it using appropriate *wavelet filters* (both high- and low-frequency ones). We won't explain in details here how to perform all the wavelet transform steps (refer to [12] for details), though, we will consider only those steps which are necessary for completing our task.

So, to decompose the quantity signal q by two levels using Daubechies second-order low-pass wavelet decomposition filter $l \equiv \left(\frac{1-\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{1+\sqrt{3}}{4\sqrt{2}}\right)$, we need to perform the following operations:

$a_2 = (q *_{\downarrow 2} l) *_{\downarrow 2} l$ = (2272.128, 136.352, 158.422, 569.098).

By $*_{\downarrow 2}$ we denote the operation of convolution of two vectors followed by dyadic downsampling of the output. Also, we present the numerical values with three decimal numbers only due to the limited space of this paper.

By analogue, we can use the flipped version of l (which would be a high-pass wavelet decomposition filter) denoted by $h = \left(\frac{1+\sqrt{3}}{4\sqrt{2}}, \frac{3+\sqrt{3}}{4\sqrt{2}}, \frac{3-\sqrt{3}}{4\sqrt{2}}, \frac{1-\sqrt{3}}{4\sqrt{2}}\right)$ to obtain detail coefficients at level 2:

$d_2 = (q *_{\downarrow 2} l) *_{\downarrow 2} h$ = (–508.185, 15.587, 546.921, –315.680).

According to the wavelet theory, every numerical array can be presented as the sum of its low-frequency component (at the last decomposition level) and a set of several high-frequency ones at each decomposition level (called *approximation* and *details* respectively). In general, the signal approximation and details can be obtained the following way (we will also substitute the values from our example):

$A_2 = (a_2 *_{\uparrow 2} l) *_{\uparrow 2} l$ = (1369.821, 687.286, 244.677, 41.992, –224.980, 11.373, 112.860, 79.481, 82.240, 175.643, 244.757, 289.584, 340.918, 693.698, 965.706, 1156.942);

$D_1 + D_2 = d_1 *_{\uparrow 2} h + (d_2 *_{\uparrow 2} h) *_{\uparrow 2} l$ = (–1350.821, –675.286, –91.677, 29.008, 237.980, 67.627, –105.860, –46.481, –66.240, 94.357, 567.243, –154.584, –99.918, –679.698, –905.706, 3180.058).

To provide group anonymity (or, redistribute signal extremums, which is the same), we need to replace A_2 with another approximation, such that the resultant signal (obtained when being summed up with our details $D_1 + D_2$) becomes different. Moreover, the only values we can try to alter are approximation coefficients.

So, in general, we need to solve a corresponding optimization problem. Knowing the dependence between A_2 and a_2 (which is pretty easy to obtain in our model example), we can set appropriate constraints, and obtain a solution $\hat{a}_2$ which completely meets our requirements.

For instance, we can set the following constraints:

$$\begin{cases} 0.637 \cdot \hat{a}_2(1) - 0.137 \cdot \hat{a}_2(4) \le 1369.821; \\ 0.296 \cdot \hat{a}_2(1) + 0.233 \cdot \hat{a}_2(2) - 0.029 \cdot \hat{a}_2(4) \le 687.286; \\ 0.079 \cdot \hat{a}_2(1) + 0.404 \cdot \hat{a}_2(2) + 0.017 \cdot \hat{a}_2(4) \le 244.677; \\ -0.137 \cdot \hat{a}_2(1) + 0.637 \cdot \hat{a}_2(2) \ge -224.980; \\ -0.029 \cdot \hat{a}_2(1) + 0.296 \cdot \hat{a}_2(2) + 0.233 \cdot \hat{a}_2(3) \ge 11.373; \\ 0.017 \cdot \hat{a}_2(1) + 0.079 \cdot \hat{a}_2(2) + 0.404 \cdot \hat{a}_2(3) \ge 112.860; \\ -0.012 \cdot \hat{a}_2(2) + 0.512 \cdot \hat{a}_2(3) \ge 79.481; \\ -0.137 \cdot \hat{a}_2(2) + 0.637 \cdot \hat{a}_2(3) \ge 82.240; \\ -0.029 \cdot \hat{a}_2(2) + 0.296 \cdot \hat{a}_2(3) + 0.233 \cdot \hat{a}_2(4) \ge 175.643; \\ 0.233 \cdot \hat{a}_2(1) - 0.029 \cdot \hat{a}_2(3) + 0.296 \cdot \hat{a}_2(4) \ge 693.698; \\ 0.404 \cdot \hat{a}_2(1) + 0.017 \cdot \hat{a}_2(3) + 0.079 \cdot \hat{a}_2(4) \le 965.706; \\ 0.512 \cdot \hat{a}_2(1) - 0.012 \cdot \hat{a}_2(4) \le 1156.942. \end{cases}$$

The solution might be as follows: $\hat{a}_2$ = (0, 379.097, 31805.084, 5464.854).

Now, let us obtain our new approximation $\hat{A}_2$, and a new quantity signal $\hat{q}$:

$\hat{A}_2 = (\hat{a}_2 *_{\uparrow 2} l) *_{\uparrow 2} l$ = (–750.103, –70.090, 244.677, 194.196, 241.583, 345.372, 434.049, 507.612, 585.225, 1559.452, 2293.431, 2787.164, 3345.271, 1587.242, 449.819, –66.997);

$\hat{q} = \hat{A}_2 + D_1 + D_2$ = (–2100.924, –745.376, 153.000, 223.204, 479.563, 413.000, 328.189, 461.131, 518.985, 1653.809, 2860.674, 2632.580, 3245.352, 907.543, –455.887, 3113.061).

Two main problems almost always arise at this stage. As we can see, there are some negative elements in the modified

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

goal signal. This is completely awkward. A very simple though quite adequate way to overcome this backfire is to add a reasonably big number (2150 in our case) to all signal elements. Obviously, the mean value of the signal will change. After all, these two issues can be solved using the following formula: $q^*_{mod} = (\hat{q} + 2150) \cdot \left(\sum_{i=1}^{16} q_i \right) \Big/ \left(\sum_{i=1}^{16} (\hat{q}_i + 2150) \right)$.

If to round q^*_{mod} (since quantities have to be integers), we obtain the modified goal signal as follows:

q^*_{fin} = (6, 183, 300, 310, 343, 334, 323, 341, 348, 496, 654, 624, 704, 399, 221, 686).

The graphical representation is available in Fig. 1b.

As we can see, the group anonymity problem at this point has been completely solved: all initial extremums persisted, and some new ones emerged.

The last step of our algorithm (i.e., obtaining new microfile $\mathbf{M}^*$) cannot be shown in this paper due to evident space limitations.

B. Concentration Group Anonymity Problem

Now, let us take the same dataset we processed before. But, this time we will pick another goal mapping function. We will try to build up a concentration signal.

According to (1), what we need to do first is to define what ρ_i to choose. In our opinion, the whole quantity of males 18 to 70 years of age would suffice.

By completing necessary arithmetic operations, we finally obtain the concentration signal:

c = (0.004, 0.002, 0.033, 0.009, 0.002, 0.012, 0.002, 0.007, 0.001, 0.035, 0.058, 0.017, 0.030, 0.003, 0.004, 0.128).

The graphical representation can be found in Fig. 2a.

Let us perform all the operations we've accomplished earlier, without any additional explanations (we will reuse notations from the previous subsection):

$a_2 = (c *_{\downarrow 2} l) *_{\downarrow 2} l$ = (0.073, 0.023, 0.018, 0.059);

$d_2 = (c *_{\downarrow 2} l) *_{\downarrow 2} h$ = (0.003, –0.001, 0.036, –0.018);

$A_2 = (a_2 *_{\uparrow 2} l) *_{\uparrow 2} l$ = (0.038, 0.025, 0.016, 0.011, 0.004, 0.009, 0.010, 0.009, 0.008, 0.019, 0.026, 0.030, 0.035, 0.034, 0.034, 0.037);

$D_1 + D_2 = d_1 *_{\uparrow 2} h + (d_2 *_{\uparrow 2} h) *_{\uparrow 2} l$ = (–0.034, –0.023, 0.017, –0.002, –0.002, 0.003, –0.009, –0.002, –0.007, 0.016, 0.032, –0.013, –0.005, –0.031, –0.030, 0.091).

The constraints for this example might look the following way:

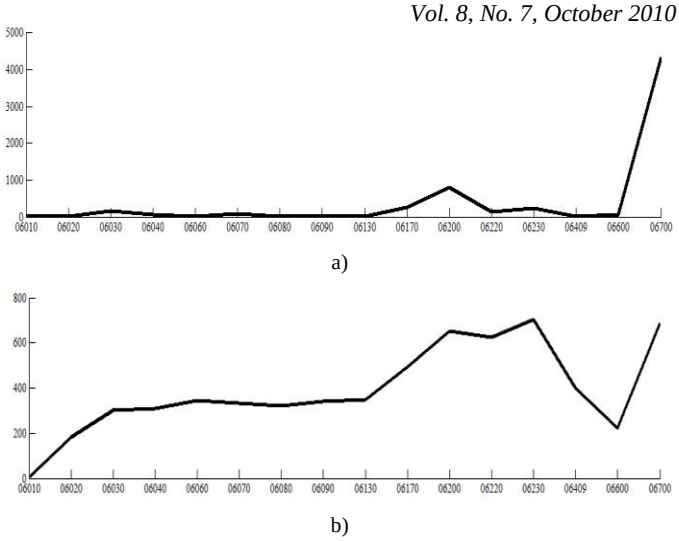

Figure 1. Initial (a) and modified (b) quantity signals.

$$\begin{cases} 0.637 \cdot \hat{a}_2(1) - 0.137 \cdot \hat{a}_2(4) \le 0.038; \\ 0.296 \cdot \hat{a}_2(1) + 0.233 \cdot \hat{a}_2(2) - 0.029 \cdot \hat{a}_2(4) \le 0.025; \\ 0.079 \cdot \hat{a}_2(1) + 0.404 \cdot \hat{a}_2(2) + 0.017 \cdot \hat{a}_2(4) \le 0.016; \\ -0.012 \cdot \hat{a}_2(1) + 0.512 \cdot \hat{a}_2(2) \le 0.011; \\ -0.137 \cdot \hat{a}_2(1) + 0.637 \cdot \hat{a}_2(2) \ge 0.005; \\ -0.029 \cdot \hat{a}_2(1) + 0.296 \cdot \hat{a}_2(2) + 0.233 \cdot \hat{a}_2(3) \ge 0.009; \\ 0.017 \cdot \hat{a}_2(1) + 0.079 \cdot \hat{a}_2(2) + 0.404 \cdot \hat{a}_2(3) \ge 0.010; \\ -0.012 \cdot \hat{a}_2(2) + 0.512 \cdot \hat{a}_2(3) \ge 0.009; \\ -0.137 \cdot \hat{a}_2(2) + 0.637 \cdot \hat{a}_2(3) \ge 0.009; \\ -0.029 \cdot \hat{a}_2(2) + 0.296 \cdot \hat{a}_2(3) + 0.233 \cdot \hat{a}_2(4) \ge 0.019; \\ 0.233 \cdot \hat{a}_2(1) - 0.029 \cdot \hat{a}_2(3) + 0.296 \cdot \hat{a}_2(4) \le 0.034; \\ 0.404 \cdot \hat{a}_2(1) + 0.017 \cdot \hat{a}_2(3) + 0.079 \cdot \hat{a}_2(4) \le 0.034; \\ 0.512 \cdot \hat{a}_2(1) - 0.012 \cdot \hat{a}_2(4) \le 0.037. \end{cases}$$

One possible solution to this system is as follows: $\hat{a}_2$ = = (0, 0.002, 0.147, 0.025).

We can obtain new approximation and concentration signal:

$\hat{A}_2 = (\hat{a}_2 *_{\uparrow 2} l) *_{\uparrow 2} l$ = (–0.003, –0.000, 0.001, 0.001, 0.001, 0.035, 0.059, 0.075, 0.093, 0.049, 0.022, 0.011, –0.004, 0.003, 0.005, 0.000);

$\hat{c} = \hat{A}_2 + D_1 + D_2$ = (–0.037, –0.023, 0.018, –0.001, –0.002, 0.038, 0.051, 0.073, 0.086, 0.066, 0.054, –0.002, –0.009, –0.028, –0.026, 0.092).

Once again, we need to make our signal non-negative, and fix its mean value. But, it is obvious that the corresponding quantity signal q^*_{mod} will also have a different mean value. Therefore, fixing the mean value can be done in "the quantity domain" (which we won't present here).

Nevertheless, it is possible to make the signal non-negative after all:

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

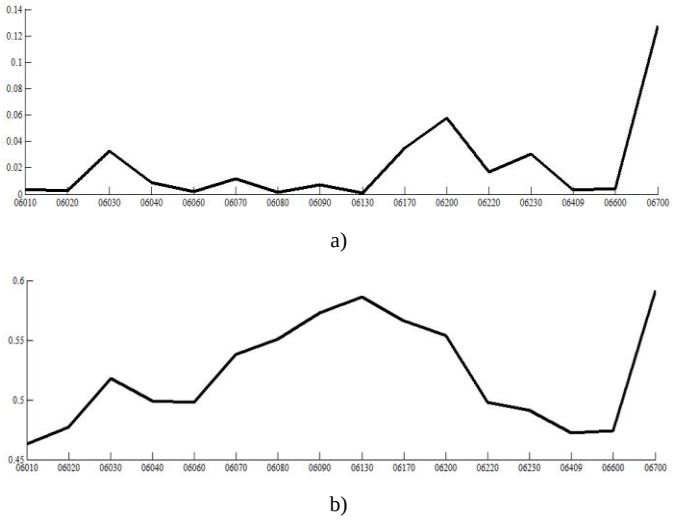

Figure 2. Initial (a) and modified (b) concentration signals.

$c^*_{mod} = \hat{c} + 0.5$ = (0.463, 0.477, 0.518, 0.499, 0.498, 0.538, 0.551, 0.573, 0.586, 0.566, 0.554, 0.498, 0.491, 0.472, 0.474, 0.592).

The graphical representation can be found in Fig. 2b. Once again, the group anonymity has been achieved.

The last step to complete is to construct the modified $\mathbf{M}^*$, which we will omit in this paper.

VI. SUMMARY

In this paper, it is the first time that group anonymity problem has been thoroughly analyzed and formalized. We presented a generic mathematical model for group anonymity in microfiles, outlined the scheme for providing it in practice, and showed several real-life examples.

As we think, there still remain some unresolved issues, some of them are as follows:

1) Choosing data representation: There are still many more ways to pick convenient goal representation of initial data not covered in this paper. They might depend on some problem task definition peculiarities.

2) Performing goal representation's modification: It is obvious that the method discussed in Section V is not an exclusive one. There could be as well proposed other sufficient techniques to perform data modifications. For instance, choosing different wavelet bases could lead to yielding different outputs.

3) Obtaining the modified microfile: There has to be developed computationally effective heuristics to perform inverse goal mapping.

REFERENCES

[1] The Free Haven Project [Online]. Available: http://freehaven.net/anonbib/full/date.html.

[2] B. Fung, K. Wang, R. Chen, P. Yu, "Privacy-preserving data publishing: a survey on recent developments," ACM Computing Surveys, vol. 42(4), 2010.

[3] A. Evfimievski, "Randomization in privacy preserving data mining," ACM SIGKDD Explorations Newsletter, 4(2), pp. 43-48, 2002.

[4] H. Kargupta, S. Datta, Q. Wang, K. Sivakumar, "Random data perturbation techniques and privacy preserving data mining", Knowledge and Information Systems, 7(4), pp. 387-414, 2005.

[5] J. Domingo-Ferrer, J. M. Mateo-Sanz, "Practical data-oriented microaggregation for statistical disclosure control," IEEE Transactions on Knowledge and Data Engineering, 14(1), pp. 189-201, 2002.

[6] J. Domingo-Ferrer, "A survey of inference control methods for privacy-preserving data mining," in Privacy-Preserving Data Mining: Models and Algorithms, C. C. Aggarwal and P. S. Yu, Eds. New York: Springer, 2008, pp. 53-80.

[7] S. E. Fienberg, J. McIntyre, Data Swapping: Variations on a Theme by Dalenius and Reiss, Technical Report, National Institute of Statistical Sciences, 2003.

[8] S. Xu, J. Zhang, D. Han, J. Wang, "Singular value decomposition based data distortion strategy for privacy protection," Knowledge and Information Systems, 10(3), pp. 383-397, 2006.

[9] J. Wang, W. J. Zhong, J. Zhang, "NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets," in The 6th IEEE Conference on Data Mining, International Workshop on Privacy Aspects of Data Mining. Washington: IEEE Computer Society, 2006, pp. 513-517.

[10] O. Chertov, A. Pilipyuk, "Statistical disclosure control methods for microdata," in International Symposium on Computing, Communication and Control. Singapore: IACSIT, 2009, pp. 338-342.

[11] O. Chertov, D. Tavrov, "Group anonymity," in IPMU-2010, CCSI, vol. 81, E. Hüllermeier and R. Kruse, Eds. Heidelberg: Springer, 2010, pp. 592-601.

[12] O. Chertov, D. Tavrov, "Providing group anonymity using wavelet transform," in BNCOD 2010, LNCS, vol. 6121, L. MacKinnon, Ed. Heidelberg: Springer, 2010, in press.

[13] O. Chertov, Group Methods of Data Processing. Raleigh: Lulu.com, 2010.

[14] L. Liu, J. Wang, J. Zhang, "Wavelet-based data perturbation for simultaneous privacy-preserving and statistics-preserving", in 2008 IEEE International Conference on Data Mining Workshops. Washington: IEEE Computer Society, 2008, pp. 27-35.

[15] U.S. Census 2000. 5-Percent Public Use Microdata Sample Files [Online]. Available: http://www.census.gov/Press-Release/www/2003/PUMS5.html.

http://sites.google.com/site/ijcsis/
ISSN 1947-5500

УДК 621.391: 517.518.3

О.Р. Чертов

НЕДІАДНІ ОДНОВИМІРНІ ВЕЙВЛЕТ-ПЕРЕТВОРЕННЯ

Вступ

Застосування апарату класичного гармонічного аналізу для розвинення функцій часу чи простору, які визначені на обмеженому інтервалі з локальними особливостями, є малоефективним, оскільки базисна функція рядів Фур'є (синусоїда) існує на всій числовій прямій і за своєю природою є гладкою та періодичною.

Для дослідження нестаціонарних сигналів у 80–90-х рр. XX ст. було запропоновано принципово інший клас (солітоноподібних) функцій – *вейвлети* [1, 2].

На практиці найчастіше застосовуються одно- та двовимірні дискретні вейвлет-перетворення (ВП) з коефіцієнтом масштабування, що дорівнює 2, бо вони мають найбільш ефективну програмну реалізацію. Такі ВП також називають діадними (dyadic). Проте існує низка проблемних галузей (геофізика, статистика, економіка, обробка зображень, включаючи рентгенографічні знімки), де часто більш прийнятними в силу різних міркувань є коефіцієнти масштабування, які не дорівнюють 2, що дають недіадні ВП.

Для побудови недіадних дискретних ВП відомо кілька підходів. Перший із них запропоновано Парком і Вудбьорном в [3], але він спирається на побудову базисів Ґрьобнера, яка "має досить високу складність і фактично не є прийнятною для практичного використання, за винятком рідкісних випадків" [4, с. 126]. Підхід, що розвивається Боссаром, дає можливість будувати вейвлети з коефіцієнтом масштабування, який визначений будь-яким додатним раціональним числом, але відповідне ВП обмежене ортогональним випадком [5]. Підхід Брателі та Йоргенсена [6], що базується на представленнях алгебри Кунца [7], знайшов, зокрема, практичне застосування для побудови недіадних вейвлетів, які мають аналітичне визначення, наприклад, для недіадних вейвлетів Хаара, Котельникова–Шеннона і Мейєра [8, 9]. Даний підхід дає змогу повністю відновити недіадне дискретне біортогональне ВП за низькочастотним фільтром декомпозиції. У той же час питання, чи є можливість збереження в недіадному (зокрема, тріадному) дискретному ВП не тільки вказаного фільтра, але й одного чи кількох інших фільтрів існуючого діадного ВП, є відкритим.

Постановка задачі

Об'єктом дослідження є дискретне ВП, що застосовується для математичної обробки різноманітних сигналів природного походження. Мета статті – виявлення випадків, коли можлива побудова недіадного дискретного ВП, що має спільний (з точністю до коефіцієнтів зведення) фільтр з відповідним діадним ВП, яке визначене не аналітично, а за допомогою біортогональних числових низькочастотних та високочастотних фільтрів декомпозиції і відновлення, та демонстрація практичних аспектів розрахунку такого недіадного ВП. Необхідність у проведенні вказаних розрахунків виникає, наприклад, коли потрібно вдало підібраний вейвлет при коефіцієнті масштабування, що дорівнює 2, перенести на інші кратності масштабування.

Всі побудови ВП обмежуються одновимірним випадком.

Загальні властивості вейвлет-перетворень

Вейвлетний базис простору $L_2(\mathbb{R})$ будується з фінітних функцій, які повинні прямувати до нуля на нескінченності. Чим швидше ці функції прямують до нуля, тим зручніше їх використовувати як базис перетворень під час аналізу реальних сигналів.

Найбільш поширеним методом теорії вейвлетів є кратномасштабний (мультироздільний) аналіз, який базується на зображенні даних з різним ступенем деталізації. Це дає можливість вивчати глобальні особливості даних при великомасштабному зображенні і виділяти локальні особливості на менших масштабах [10, 11]. Кратність кратномасштабного аналізу є коефіцієнтом масштабування відповідного ВП. При цьому масштабуюча функція $\varphi(x)$, яка інколи в російській та вітчизняній літературі називається скейлінг-функцією, визначається за допомогою такого рівняння:

$$\varphi(x) = \sqrt{N}\sum_n h_n \varphi(Nx - n),$$

де n – цілі числа; N – коефіцієнт масштабування; h_n – низькочастотний фільтр, причому $\sum_n |h_n|^2 < \infty$.

Доведено [2, с. 333–335], що якщо одні й ті самі дійсні ортонормовані фільтри, які мають компактний носій, використовуються і для розкладення, і для відновлення сигналів, то точне відновлення симетричними вейвлетами неможливе (за винятком функцій Хаара). У той же час на практиці часто трапляються ситуації, коли симетричність є принциповою. Наприклад, симетричні похибки квантування зображення внаслідок специфіки зорової системи людини менш помітні, ніж асиметричні, природні геофізичні сигнали, як правило, симетричні тощо.

Отже, для забезпечення симетричності замість ортонормованих потрібно використовувати біортогональні фільтри, тобто дві ортогональні одна одній пари {масштабуюча функція, вейвлет}, одну пару – для декомпозиції (розкладення) сигналу, а іншу – для його відновлення (реконструкції).

Масштабуючій функції відповідають низькочастотні фільтри (фільтри апроксимації), а вейвлету – високочастотні фільтри (фільтри деталізації).

Кожен фільтр визначається набором коефіцієнтів, тобто своїми значеннями в точках скінченного носія.

Біортогональні сплайн-вейвлети

Біортогональні сплайн-вейвлети, запропоновані в праці [12], широко застосовуються під час стиснення зображень, бо вони мають добру чисельну стійкість та дають невеликі вейвлет-коефіцієнти на гладких зображеннях [1, с. 293–294]. Використаємо ці сплайн-вейвлети для ілюстрації методу побудови недіадних одновимірних біортогональних вейвлетів, що застосовується в даній статті. У табл. 1 зведено коефіцієнти фільтрів біортогональних сплайн-вейвлетів, які було взято безпосередньо з функції rbiowavf('rbio4.4') системи комп'ютерної математики MATLAB (менш точні значення цих коефіцієнтів наведено в [1, с. 295] і [2, с. 368]).

На рис. 1 наведено графіки масштабуючих функцій і вейвлет-функцій, відновлених за вказаними фільтрами за 10 ітерацій каскадного алгоритму Добеші [2, с. 270–279].

Вибір коефіцієнта масштабування

Найвищу швидкість вейвлет-розкладання забезпечує кратномасштабний аналіз із коефіцієнтом масштабування (масштабним множником) N, що дорівнює 2. Використання степенів двійки є зручним і з огляду на деякі інші міркування [2, с. 410–420]. Проте відомо [13], що замість двійки можна брати будь-яке інше раціональне число, більше одиниці. У ряді випадків це дає змогу краще враховувати особливості сигналів, що аналізуються. Але відкритим залишається питання оптимального вибору коефіцієнта масштабування вейвлетів.

Скажімо, під час аналізу даних економічного походження на одному рівні розкладання коефіцієнт може дорівнювати 5 (оскільки в тижні п'ять робочих днів), на наступному рівні –

***Таблиця* 1.** Коефіцієнти фільтрів біортогональних сплайн-вейвлетів (із системи MATLAB)

Точка носія фільтра	Коефіцієнти фільтрів				
	Низькочастотний фільтр декомпозиції, розділений на $\sqrt{2}$	Низькочастотний фільтр декомпозиції	Низькочастотний фільтр реконструкції	Високочастотний фільтр декомпозиції	Високочастотний фільтр реконструкції
–5	0	0	0	–0,037828455507264	0
–4	0	0	0,037828455507264	–0,023849465019557	0
–3	–0,045635881556954	–0,064538882628697	–0,023849465019557	0,110624404418437	0
–2	–0,028771763113971	–0,040689417609164	–0,110624404418437	0,377402855612831	–0,064538882628697
–1	0,295635881556704	0,418092273221617	0,377402855612831	–0,852698679008894	0,040689417609164
0	0,557543526228443	0,788485616405583	0,852698679008894	0,377402855612831	0,418092273221617
1	0,295635881556704	0,418092273221617	0,377402855612831	0,110624404418437	–0,788485616405583
2	–0,028771763113971	–0,040689417609164	–0,110624404418437	–0,023849465019557	0,418092273221617
3	–0,045635881556954	–0,064538882628697	–0,023849465019557	–0,037828455507264	0,040689417609164
4	0	0	0,037828455507264	0	–0,064538882628697

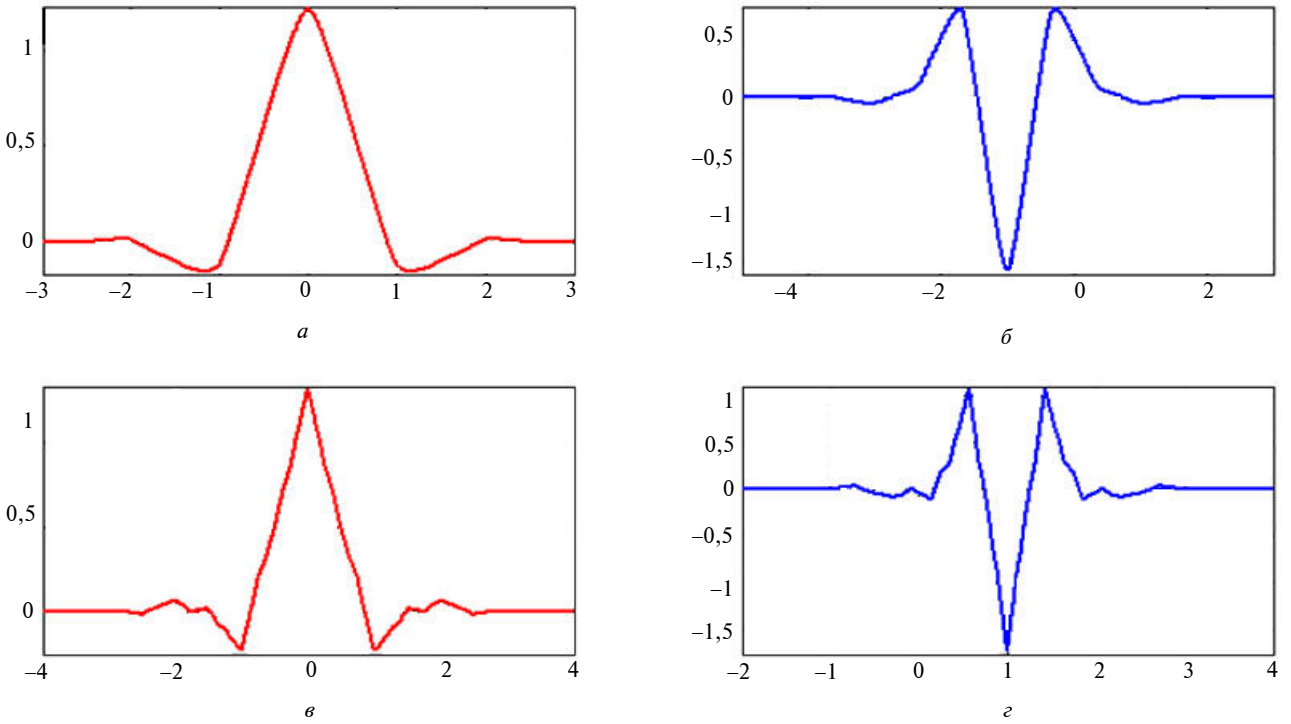

Рис. 1. Діадні сплайн-функції, відновлені за 10 ітерацій: *а* – масштабуюча функція декомпозиції; *б* – вейвлет декомпозиції; *в* – масштабуюча функція реконструкції; *г* – вейвлет реконструкції

21/5 (оскільки в середньому із врахуванням свят в місяці 21 робочий день), на подальшому рівні – 3 (квартал утворюють три місяці) тощо [11]. Для аналізу і обробки зображень у системах технічного зору [14] та комп'ютерній томографії [15] знаходять застосування методи адаптивного підбору коефіцієнта масштабування залежно від частотних особливостей зображення, що препарується.

При аналізі геофізичних даних вибір коефіцієнта масштабування повинен спиратися насамперед на геологічні фактори, що впливають на досліджуване випромінювання. Так, академік М.А. Садовський, вивчаючи реакцію гірських порід на вибухові впливи, був першим, хто звернув увагу на фундаментальні властивості геофізичного середовища, що проявляються в його ієрархічній побудові та відповідній поведінці в полі механічних геодинамічних напружень. Група, що працювала під його керівництвом, виявила дискретний розподіл твердотільностей (від порошинок до планет) з характерним коефіцієнтом пропорційності, що дорівнює 3,4 [16, 17].

На базі цих досліджень В.І. Уломовим було розроблено феноменологічну ґратову модель сейсмогенезу [18], яка відображала самоподібність у просторово-часовому та енергетичному розвитку геодинамічних структур і сейсмічних процесів. Зокрема, було показано, що ієрархічна множина геоблоків та осередків землетрусів має одну й ту саму чи дуже близьку фрактальну розмірність. Вказану ґратову модель у подальшому було використано для конструювання лінеаментно-доменно-фокальної моделі зон утворення осередків землетрусів, на основі якої в 1991–1997 рр. було створено комплект імовірнісних нормативних карт загального сейсмічного районування, що ввійшов до Міжнародних будівельних норм СНД [19].

Таким чином, вибір коефіцієнта масштабування істотно залежить від проблемної області та задачі, що розв'язується. Виходячи з того, що найбільш близьким до фрактальної розмірності геологічних структур, яка дорівнює 3,4, є коефіцієнт масштабування N, рівний 3, та враховуючи, що тріадні вейвлети найбільш прості за побудовою серед цілочислових недіадних, конкретні підрахунки в подальшому будемо проводити лише для тріадних вейвлетів, в той же час не обмежуючись цим випадком у загальних міркуваннях.

Побудова недіадного вейвлет-перетворення за спільним з діадним вейвлет-перетворенням низькочастотним фільтром декомпозиції

Щоб зафіксувати систему позначень, що буде надалі використовуватися, запишемо за її допомогою умову точного відновлення вихідного дискретного сигналу. При цьому вейвлет-перетворення здійснюватимемо симетричними біортогональними фільтрами з коефіцієнтом масштабування N.

Множину цілих чисел позначимо $\mathbb{Z}$.

Нехай маємо скінченний дискретний набір відліків (вимірювань) вихідного сигналу, в яких він набуває значень $x_0, x_1, x_2, ..., x_{L-1}$.

Припустимо, що здійснено вейвлет-аналіз (з коефіцієнтом масштабування N) вихідного сигналу за допомогою певних (у загальному випадку неортогональних) фільтрів $\{h_n^i\}$, тобто

$$d_{1,k}^i = \sum_{n=l_i}^{r_i} h_n^i x_{Nk-n}, \qquad (1)$$

де $[l_i, r_i]$ – носій фільтра $\{h_n^i\}$, $i = 0, 2, ..., N-1$, тобто $n = l_i, l_i+1, l_i+2, ..., r_i-1, r_i$; $n \in \mathbb{Z}$.

Індекс "1" в позначеннях коефіцієнтів $d_{1,k}^i$ у формулі (1) вказує на те, що вони були отримані на першому рівні (кроці) вейвлет-аналізу. В загальному випадку процедура вейвлет-аналізу є ітеративною (каскадною). Якщо значення x_k вихідного сигналу розглядати як $d_{0,k}^0$, то формулу знаходження коефіцієнтів на $(m+1)$-му рівні можна записати таким чином:

$$d_{m+1,k}^i = \sum_{n=l_i}^{r_i} h_n^i d_{m,Nk-n}^0, \qquad (2)$$

де $m = 0, 1, 2, ...$ (значення $\log_N L$, округлене до найближчого більшого цілого числа); інші індекси набувають таких самих значень, що й у формулі (1).

Графічно процедуру знаходження за формулою (2) коефіцієнтів $d_{m+1}^i \equiv \{d_{m+1,k}^i\}$ вейвлет-аналізу можна зобразити схемою, що наведена на рис. 2.

Формально вихідний сигнал $x_0, x_1, x_2, ..., x_{L-1}$, де L – кількість його відліків, можна записати як степеневий ряд $X(z) \equiv \sum_{n=0}^{L-1} x_n z^n$, де $z \equiv e^{-i\omega}$.

Аналогічно, фільтрам $\{h_n^i\}$ можна співставити передатні функції (розкладення за степенями z):

$$H_i(z) \equiv \sum_{n=l_i}^{r_i} h_n^i z^n, \; i = 0, \; 1, \; 2, \; \ldots, \; N-1.$$

На етапі декомпозиції (аналізу) спочатку один низькочастотний фільтр H_0 і $(N-1)$ високочастотних фільтрів $H_1, ..., H_{N-1}$ здійснюють фільтрацію вихідного сигналу X, а потім – розрідження (децимацію) N раз отриманого результату, тобто залишається лише кожний N-й відлік, починаючи з першого. У всьому іншому алгоритм розрахунку недіадного ВП з коефіцієнтом масштабування N не відрізняється від пірамідального алгоритму для діадного перетворення [1, 2].

Операції, що здійснюються на етапі відновлення (синтезу), є оберненими до операцій етапу аналізу, а саме: спочатку N раз застосовується інтерполяція, тобто після кожного відліку додається $(N-1)$ нулів, потім виконується відновлення вихідного сигналу X за допомогою передатних функцій $G_i(z)$ фільтрів відновлення $\{g_n^i\}$, $i = 0, 1, 2, ..., N-1$.

Відновлений сигнал позначатимемо $\hat{X}(z) \equiv \sum_{n=0}^{L-1} \hat{x}_n z^n$.

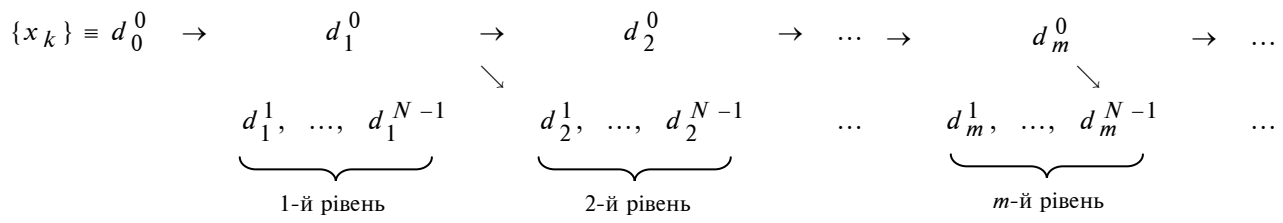

Рис. 2. Схема каскадної процедури вейвлет-аналізу з N фільтрами

Відновлення вихідного сигналу буде точним, якщо $X(z)=\hat{X}(z)$.

Розглянемо матрицю фільтрів розкладення:

$$\mathbf{H}(z) = \frac{1}{\sqrt{N}}\begin{pmatrix} H_0(z) & H_0(\rho z) & \dots & H_0(\rho^{N-1}z) \\ H_1(z) & H_1(\rho z) & & H_1(\rho^{N-1}z) \\ \dots & \dots & \dots & \dots \\ H_{N-1}(z) & H_{N-1}(\rho z) & & H_{N-1}(\rho^{N-1}z) \end{pmatrix}. \quad (3)$$

Як відомо [9, с. 133–134, теорема 2], якщо матриця невироджена, то можливе точне відновлення сигналу фільтрами, спряженими до фільтрів $G_i(z)$, $i=0,\ 1,\ 2,\ ...,\ N-1$, матриця яких

$$\mathbf{G}(z) = \frac{1}{\sqrt{N}}\begin{pmatrix} G_0(z) & G_0(\rho z) & \dots & G_0(\rho^{N-1}z) \\ G_1(z) & G_1(\rho z) & & G_1(\rho^{N-1}z) \\ \dots & \dots & \dots & \dots \\ G_{N-1}(z) & G_{N-1}(\rho z) & & G_{N-1}(\rho^{N-1}z) \end{pmatrix} \quad (4)$$

є транспонованою до оберненої матриці $\mathbf{H}(z)$ фільтрів декомпозиції.

У матрицях (3) і (4) для спрощення запису використане позначення $\rho \equiv e^{i2\pi/N}$.

Перетворенням Фур'є по циклічній групі $\mathbb{Z}/N\mathbb{Z}$, яка дорівнює $\{1,\rho,\rho^2,\dots,\rho^{N-1}\}$, матриця $\mathbf{H}(z)$ зводиться [6, с. 39] до такого вигляду:

$$\mathbf{H}(z) = \mathbf{A}(z^N)\begin{pmatrix} 1 & 1 & \dots & 1 \\ z & \rho z & \dots & \rho^{N-1}z \\ \dots & \dots & \dots & \dots \\ z^{N-1} & \rho^{N-1}z^{N-1} & \dots & \rho^{(N-1)^2}z^{N-1} \end{pmatrix}, \quad (5)$$

де елементи унітарної матриці $\mathbf{A}(w)$ визначаються формулою

$$A_{i,j}(w) = \frac{1}{N}\sum_{z^N=w} z^{-j}\sum_{k=0}^{N-1} H_{i,k}(z). \quad (6)$$

Таким чином, початкова задача підбору фільтрів недіадного вейвлет-розкладення зведена до побудови унітарної матриці $\mathbf{A}(w)$.

У наступних викладках вважатимемо, що $N=3$.

Для того щоб забезпечити максимальний збіг можливостей діадного і недіадного вейвлет-розкладень з розпізнавання однакових аномалій у вихідному сигналі, низькочастотний фільтр декомпозиції недіадного ВП візьмемо таким, що дорівнює низькочастотному фільтру декомпозиції діадного ВП з точністю до коефіцієнтів зведення. Це означає, що коефіцієнти низькочастотного фільтра декомпозиції недіадного ВП отримаємо, помноживши на $\sqrt{3}$ відповідні коефіцієнти низькочастотного фільтра декомпозиції діадного ВП, розділені на $\sqrt{2}$. Носій цього фільтра (біортогональних сплайн-вейвлетів) діадного розкладення дорівнює семи точкам, тому вказана рівність буде забезпечена на носієві від -3 до 3, тобто

$$H_0(z) = \sum_{n=-3}^{3} h_n^0 z^n. \quad (7)$$

Тоді під час підрахунку значень $A_{0,j}(w)$, $j=0,1,2$, у формулі (6) сума буде братися по значеннях $w=z^{-3},\ 1,\ z^3$. Елементарні розрахунки дають, що перший рядок $A_{0,0-2}(w)$ матриці $\mathbf{A}(w)$ дорівнює

$$(h_{-3}^0 w^{-1} + h_0^0 + h_3^0 w,\ h_{-2}^0 w^{-1} + h_1^0,\ h_{-1}^0 w^{-1} + h_2^0). \quad (8)$$

Дійсно, наприклад, маємо

$$A_{0,0}(w) = \frac{1}{3}\sum_{z^3=w}\sum_{k=0}^{2} H_{0,k}(z) =$$

$$= \frac{1}{3}\sum_{z^3=w}(H_0(z) + H_0(\rho z) + H_0(\rho^2 z)) =$$

$$= \frac{1}{3}\sum_{z^3=w}\left(\sum_{n=-3}^{3} h_n^0 z^n + \sum_{n=-3}^{3} h_n^0(\rho z)^n + \sum_{n=-3}^{3} h_n^0(\rho^2 z)^n\right) =$$

$$= \frac{1}{3}\sum_{m=-1}^{1}(1+\rho^{3m}+\rho^{6m})h_{3m}^0 z^{3m} = \sum_{m=-1}^{1} h_{3m}^0 z^{3m} =$$

$$= h_{-3}^0 z^{-3} + h_0^0 + h_3^0 z^3 = h_{-3}^0 w^{-1} + h_0^0 + h_3^0 w.$$

З одного боку, рівність (8) можна розглядати як розкладення першого рядка матриці $\mathbf{A}(w)$ по стовпцях $\{\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3\}$ цієї матриці в просторі $\mathbb{C}^3$, тобто

$$A_{0,0-2}(w) = (h_{-3}^0 w^{-1} + h_0^0 + h_3^0 w)\,\mathbf{e}_1 +$$

$$+ (h_{-2}^0 w^{-1} + h_1^0)\,\mathbf{e}_2 + (h_{-1}^0 w^{-1} + h_2^0)\,\mathbf{e}_3.$$

З іншого боку, ввівши позначення ковекторів

$$\tilde{\boldsymbol{\alpha}}_{-1} \equiv (h^0_{-3}, h^0_{-2}, h^0_{-1}),\ \tilde{\boldsymbol{\alpha}}_0 \equiv (h^0_0, h^0_1, h^0_2),$$

$$\tilde{\boldsymbol{\alpha}}_1 \equiv (h^0_3, 0, 0), \qquad (9)$$

рівність (8) можна переписати таким чином:

$$A_{0,0-2}(w) = \tilde{\boldsymbol{\alpha}}_{-1} w^{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1 w, \qquad (10)$$

тобто фактично розклавши перший рядок матриці $\mathbf{A}(w)$ по ковекторах $\{\tilde{\boldsymbol{\alpha}}_{-1}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1\}$ в просторі $^3\mathbb{C}$

Знайдемо матрицю $\mathbf{T}$ переходу від базису $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ до базису $\{\tilde{\boldsymbol{\alpha}}_{-1}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1\}$. Для цього потрібно розв'язати таке рівняння:

$$(\tilde{\boldsymbol{\alpha}}_{-1}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1) = \mathbf{T} \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \mathbf{e}_3 \end{pmatrix}.$$

Враховуючи (9), знаходимо

$$\mathbf{T} = \begin{pmatrix} h^0_{-3} & h^0_{-2} & h^0_{-1} \\ h^0_0 & h^0_1 & h^0_2 \\ h^0_3 & 0 & 0 \end{pmatrix}.$$

Аналогічно, легко розрахувати матрицю зворотного переходу від базису $\{\tilde{\boldsymbol{\alpha}}_{-1}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1\}$ до базису $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$:

$$\mathbf{T}^{-1} = \frac{1}{h^0_3 (h^0_{-2} h^0_2 - h^0_{-1} h^0_1)} \times$$

$$\times \begin{pmatrix} 0 & 0 & h^0_{-2} h^0_2 - h^0_{-1} h^0_1 \\ h^0_2 h^0_3 & -h^0_{-1} h^0_3 & h^0_{-1} h^0_0 - h^0_{-3} h^0_2 \\ -h^0_1 h^0_3 & h^0_{-2} h^0_3 & h^0_{-3} h^0_1 - h^0_{-2} h^0_0 \end{pmatrix}.$$

Враховуючи, що перший рядок матриці $\mathbf{A}(w)$ має вигляд, визначений формулою (10), всю матрицю $\mathbf{A}(w)$ можна факторизувати таким чином:

$$\mathbf{A}(w) = \mathbf{B} \cdot \mathbf{T}^{-1} \cdot \mathbf{D}(w) \cdot \mathbf{T} \equiv$$

$$\equiv \begin{pmatrix} \tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1 \\ b_{21} \quad b_{22} \quad b_{23} \\ b_{31} \quad b_{32} \quad b_{33} \end{pmatrix} \mathbf{T}^{-1} \begin{pmatrix} w^{-1} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & w \end{pmatrix} \mathbf{T}, \qquad (11)$$

де другий і третій рядки матриці $\mathbf{B}$ заповнюються такими дійсними числами, щоб вона була невиродженою; разом із діагональністю матриці $\mathbf{D}(w)$ і ортогональністю (за побудовою) матриць $\mathbf{T}^{-1}$ і $\mathbf{T}$ це буде гарантувати унітарність матриці $\mathbf{A}(w)$.

Для формули (11) матрицю $\mathbf{D}(w)$ було відновлено за допомогою побудови її проекцій на ковектори базису $\{\tilde{\boldsymbol{\alpha}}_{-1}, \tilde{\boldsymbol{\alpha}}_0, \tilde{\boldsymbol{\alpha}}_1\}$.

Покажемо, як це можна зробити, попередньо ввівши такі позначення:

одиничний оператор:

$$\mathbf{1} \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

проекції на ковектор $\tilde{\boldsymbol{\alpha}}_{-1}$:

$$\mathbf{P_{-1}} \equiv \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

і на ковектор $\tilde{\boldsymbol{\alpha}}_1$:

$$\mathbf{P_1} \equiv \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Тоді, враховуючи такі властивості операторів проекції, як

$$\tilde{\boldsymbol{\alpha}}_{-1}\mathbf{P_{-1}} = \tilde{\boldsymbol{\alpha}}_{-1},\ \tilde{\boldsymbol{\alpha}}_1\mathbf{P_{-1}} = 0,\ \tilde{\boldsymbol{\alpha}}_0\mathbf{P_{-1}} = 0$$

і

$$\tilde{\boldsymbol{\alpha}}_1\mathbf{P_1} = \tilde{\boldsymbol{\alpha}}_1,\ \tilde{\boldsymbol{\alpha}}_{-1}\mathbf{P_1} = 0,\ \tilde{\boldsymbol{\alpha}}_0\mathbf{P_1} = 0,$$

виконаємо необхідні розрахунки, базуючись на [6, с. 44–45, теорема 4.1]:

$$\beta_{-1}(w) \equiv A_{0,0-2}(w)(\mathbf{1} - \mathbf{P_{-1}} + w\mathbf{P_{-1}}) =$$

$$= (\tilde{\boldsymbol{\alpha}}_{-1} w^{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1 w)(\mathbf{1} - \mathbf{P_{-1}} + w\mathbf{P_{-1}}) =$$

$$= \tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1 w,$$

$$\beta_1(w) \equiv \beta_{-1}(w)(\mathbf{1} - \mathbf{P_1} + w^{-1}\mathbf{P_1}) =$$

$$= (\tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1 w)(\mathbf{1} - \mathbf{P_1} + w^{-1}\mathbf{P_1}) = \tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1,$$

звідки

$$A_{0,0-2}(w) = \beta_{-1}(w)\,(\mathbf{1} - \mathbf{P}_{-\mathbf{1}} + w^{-1}\mathbf{P}_{-\mathbf{1}}) =$$

$$= \beta_1(w)\,(\mathbf{1} - \mathbf{P}_{\mathbf{1}} + w\mathbf{P}_{\mathbf{1}})\,(\mathbf{1} - \mathbf{P}_{-\mathbf{1}} + w^{-1}\mathbf{P}_{-\mathbf{1}}) =$$

$$= (\tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1)\begin{pmatrix}1&0&0\\0&1&0\\0&0&w\end{pmatrix}\begin{pmatrix}w^{-1}&0&0\\0&1&0\\0&0&1\end{pmatrix} =$$

$$= (\tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1)\begin{pmatrix}w^{-1}&0&0\\0&1&0\\0&0&w\end{pmatrix},$$

тобто

$$\mathbf{D}(w) = \begin{pmatrix}w^{-1}&0&0\\0&1&0\\0&0&w\end{pmatrix}.$$

Громіздкі, але тривіальні розрахунки із врахуванням того, що

$$(\tilde{\boldsymbol{\alpha}}_{-1} + \tilde{\boldsymbol{\alpha}}_0 + \tilde{\boldsymbol{\alpha}}_1) = (h^0_{-3} + h^0_0 + h^0_3, h^0_{-2} + h^0_1, h^0_{-1} + h^0_2),$$

дають можливість формулу (11) записати як

$$\mathbf{A}(w) = \frac{1}{h^0_{-2}h^0_2 - h^0_{-1}h^0_1} \times$$

$$\times\begin{pmatrix}h^0_{-3} + h^0_0 + h^0_3 & h^0_{-2} + h^0_1 & h^0_{-1} + h^0_2\\ b_{21} & b_{22} & b_{23}\\ b_{31} & b_{32} & b_{33}\end{pmatrix}\times$$

$$\times\begin{pmatrix}\gamma_{11} & 0 & 0\\ \gamma_{21} & h^0_{-2}h^0_2 w^{-1} - h^0_{-1}h^0_1 & h^0_{-1}h^0_2 w^{-1} - h^0_{-1}h^0_2\\ \gamma_{31} & -h^0_{-2}h^0_1 w^{-1} + h^0_{-2}h^0_1 & -h^0_{-1}h^0_1 w^{-1} + h^0_{-2}h^0_2\end{pmatrix}, \quad (12)$$

де $\gamma_{11} = (h^0_{-2}h^0_2 - h^0_{-1}h^0_1)w$; $\gamma_{21} = h^0_{-3}h^0_2 w^{-1} - h^0_{-1}h^0_0 + (h^0_{-1}h^0_0 - h^0_{-3}h^0_2)\,w$; $\gamma_{31} = -h^0_{-3}h^0_1 w^{-1} + h^0_{-2}h^0_0 + (h^0_{-3}h^0_1 - h^0_{-2}h^0_0)w$.

Із формул (5) і (12) знаходимо передатні функції високочастотних фільтрів декомпозиції, що залишилися:

$$H_1(z) = \frac{1}{h^0_{-2}h^0_2 - h^0_{-1}h^0_1} \times$$

$$\times((b_{22}h^0_2 - b_{23}h^0_1)(h^0_{-3}z^{-3} + h^0_{-2}z^{-2} + h^0_{-1}z^{-1}) +$$

$$+ (b_{23}h^0_{-2} - b_{22}h^0_{-1})(h^0_0 + h^0_1 z + h^0_2 z^2) +$$

$$+ (b_{21}(h^0_{-2}h^0_2 - h^0_{-1}h^0_1) + h^0_0(b_{22}h^0_{-1} - b_{23}h^0_{-2}) +$$

$$+ h^0_{-3}(b_{23}h^0_1 - b_{22}h^0_2))z^3), \quad (13)$$

$$H_2(z) = \frac{1}{h^0_{-2}h^0_2 - h^0_{-1}h^0_1} \times$$

$$\times((b_{32}h^0_2 - b_{33}h^0_1)(h^0_{-3}z^{-3} + h^0_{-2}z^{-2} + h^0_{-1}z^{-1}) +$$

$$+ (b_{33}h^0_{-2} - b_{32}h^0_{-1})(h^0_0 + h^0_1 z + h^0_2 z^2) +$$

$$+ (b_{31}(h^0_{-2}h^0_2 - h^0_{-1}h^0_1) + h^0_0(b_{32}h^0_{-1} - b_{33}h^0_{-2}) +$$

$$+ h^0_{-3}(b_{33}h^0_1 - b_{32}h^0_2))z^3). \quad (14)$$

Передатні функції фільтрів реконструкції $G_i(z)$, $i = 0,1,2$, отримаємо, розрахувавши за формулою (4) матрицю, транспоновану до оберненої матриці $\mathbf{H}(z)$ фільтрів декомпозиції, та взявши всі її елементи комплексно спряженими:

$$G_0(z) = ((b_{22}b_{33} - b_{23}b_{32})((h^0_{-2}h^0_0 - h^0_{-3}h^0_1)z^5 +$$

$$+ (h^0_{-3}h^0_2 - h^0_{-1}h^0_0)z^4 + (h^0_{-1}h^0_1 - h^0_{-2}h^0_2)z^3) +$$

$$+ (b_{21}b_{32}h^0_2 - b_{22}h^0_2 b_{31} + b_{22}h^0_0 b_{33} - b_{23}b_{32}h^0_0 -$$

$$- b_{21}h^0_1 b_{33} + b_{23}h^0_1 b_{31})(-h^0_{-2}z^2 + h^0_{-1}z) +$$

$$+ (b_{21}b_{33}h^0_{-2} - b_{22}h^0_{-3}b_{33} + b_{22}h^0_{-1}b_{31} - b_{21}b_{32}h^0_{-1} -$$

$$- b_{23}h^0_{-2}b_{31} + b_{23}h^0_{-3}b_{32})(-h^0_1 z^{-1} + h^0_2 z^{-2}))\,/\,K, \quad (15)$$

$$G_1(z) = ((b_{32}h^0_2 + b_{32}h^0_{-1} - b_{33}h^0_{-2} - b_{33}h^0_1) \times$$

$$\times((h^0_{-2}h^0_0 - h^0_{-3}h^0_1)z^5 + (h^0_{-3}h^0_2 - h^0_{-1}h^0_0)z^4 +$$

$$+ (h^0_{-1}h^0_1 - h^0_{-2}h^0_2)z^3) + (-b_{31}h^0_{-1}h^0_1 + b_{31}h^0_{-2}h^0_2 -$$

$$- b_{32}h^0_{-3}h^0_2 + b_{32}h^0_{-1}h^0_0 - b_{32}h^0_2 h^0_3 + b_{33}h^0_{-3}h^0_1 -$$

$$- b_{33}h^0_{-2}h^0_0 + b_{33}h^0_1 h^0_3)(-h^0_{-2}z^2 + h^0_{-1}z) +$$

$$+ (-b_{31}h^0_{-1}h^0_1 + b_{31}h^0_{-2}h^0_2 - b_{32}h^0_{-3}h^0_2 + b_{32}h^0_{-1}h^0_0 +$$

$$+ b_{32}h^0_{-1}h^0_3 + b_{33}h^0_{-3}h^0_1 - b_{33}h^0_{-2}h^0_0 - b_{33}h^0_{-2}h^0_3) \times$$

$$\times(-h^0_1 z^{-1} + h^0_2 z^{-2}))\,/\,K, \quad (16)$$

$$G_2(z) = ((-b_{22}h^0_2 - b_{22}h^0_{-1} + b_{23}h^0_{-2} + b_{23}h^0_1) \times$$

$$\times((h^0_{-2}h^0_0 - h^0_{-3}h^0_1)z^5 + (h^0_{-3}h^0_2 - h^0_{-1}h^0_0)z^4 +$$

$$+ (h^0_{-1}h^0_1 - h^0_{-2}h^0_2)z^3) + (-b_{21}h^0_{-1}h^0_1 + b_{21}h^0_{-2}h^0_2 -$$

$$- b_{22}h^0_{-3}h^0_2 + b_{22}h^0_{-1}h^0_0 - b_{22}h^0_2 h^0_3 + b_{23}h^0_{-3}h^0_1 -$$

$$- b_{23}h^0_{-2}h^0_0 + b_{23}h^0_1h^0_3)(h^0_{-2}z^2 - h^0_{-1}z) +$$
$$+ (-b_{21}h^0_{-1}h^0_1 + b_{21}h^0_{-2}h^0_2 - b_{22}h^0_{-3}h^0_2 +$$
$$+ b_{22}h^0_{-1}h^0_0 + b_{22}h^0_{-1}h^0_3 + b_{23}h^0_{-3}h^0_1 -$$
$$- b_{23}h^0_{-2}h^0_0 - b_{23}h^0_{-2}h^0_3)(h^0_1z^{-1} - h^0_2z^{-2})) / K, \quad (17)$$

де K обчислюється за формулою

$$K \equiv 3(h^0_{-2}h^0_2 - h^0_{-1}h^0_1) \times$$
$$\times (K_1(h^0_{-3} + h^0_0 + h^0_3) + K_2(h^0_{-2} + h^0_1) + K_3(h^0_{-1} + h^0_2)),$$
$$K_1 \equiv b_{23}b_{32} - b_{33}b_{22},\ K_2 \equiv b_{21}b_{33} - b_{23}b_{31},$$
$$K_3 \equiv b_{22}b_{31} - b_{32}b_{21}.$$

Отримані формули (13)–(17) дають можливість обчислити високочастотні фільтри декомпозиції і всі фільтри реконструкції для недіадного вейвлет-розкладення з коефіцієнтом масштабування 3 за таких умов:

1) недіадний низькочастотний фільтр декомпозиції береться таким, що дорівнює відповідному діадному з носієм від –3 до 3 з точністю до коефіцієнтів зведення, тобто його передатна функція обчислюється за формулою (7);

2) другий і третій рядок матриці

$$\mathbf{B} = \begin{pmatrix} h^0_{-3} + h^0_0 + h^0_3 & h^0_{-2} + h^0_1 & h^0_{-1} + h^0_2 \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

заповнюються такими дійсними числами, щоб дана матриця була невиродженою.

Для наступних обчислень другий і третій рядок матриці **B** було взято (0 1 0) і (0 0 1), відповідно.

Таким чином, виходячи із вказаних формул, було розраховано всі фільтри, необхідні для одновимірного недіадного вейвлет-розкладення з коефіцієнтом масштабування 3 на базі сплайн-вейвлетів. Коефіцієнти відповідних фільтрів декомпозиції зведені в табл. 2, а фільтрів відновлення – в табл. 3.

Очевидно, що в табл. 1 і 2 стовпчики, які містять зведені коефіцієнти низькочастотного фільтра декомпозиції, збігаються.

На рис. 3 зображено графіки знайдених фільтрів недіадного ВП. Графіки відповідних

***Таблиця* 2.** Коефіцієнти фільтрів декомпозиції біортогональних сплайн-вейвлетів з коефіцієнтом масштабування 3

Точка носія фільтра	Коефіцієнти фільтрів декомпозиції			
	Низькочастотний фільтр, розділений на $\sqrt{3}$	Низькочастотний фільтр	Високочастотний фільтр (перший)	Високочастотний фільтр (другий)
–3	–0,045635881556954	–0,079043665504840	–0,026269528852333	–0,269924901354673
–2	–0,028771763113971	–0,049834155536734	–0,01656198227072	–0,170177830588116
–1	0,295635881556704	0,512056367396626	0,170177830588116	1,748612789839597
0	0,557543526228443	0,965693714858774	3,297731438151997	0,320940568013582
1	0,295635881556704	0,512056367396626	1,748612789839597	0,170177830588116
2	–0,028771763113971	–0,049834155536734	–0,170177830588116	–0,01656198227072
3	–0,045635881556954	–0,079043665504840	–3,271461909299664	–0,051015666658909

***Таблиця* 3.** Коефіцієнти фільтрів реконструкції біортогональних сплайн-вейвлетів з коефіцієнтом масштабування 3

Точка носія фільтра	Коефіцієнти фільтрів реконструкції		
	Низькочастотний фільтр	Високочастотний фільтр (перший)	Високочастотний фільтр (другий)
–2	–0,018779841093000	–0,000508995019128	0,061737609267150
–1	–0,192966793693924	–0,005230030241522	0,634366843234382
0	0	0	0
1	2,357517438703052	–0,046265883046327	–0,685862756522232
2	0,229437417833905	–0,00450267071865	–0,066749275004929
3	1,238226963033587	–0,330438346921981	–0,330438346921981
4	–2,338737597610051	0,624125147255082	0,624125147255082
5	–0,036470624139981	0,009732700960172	0,009732700960172

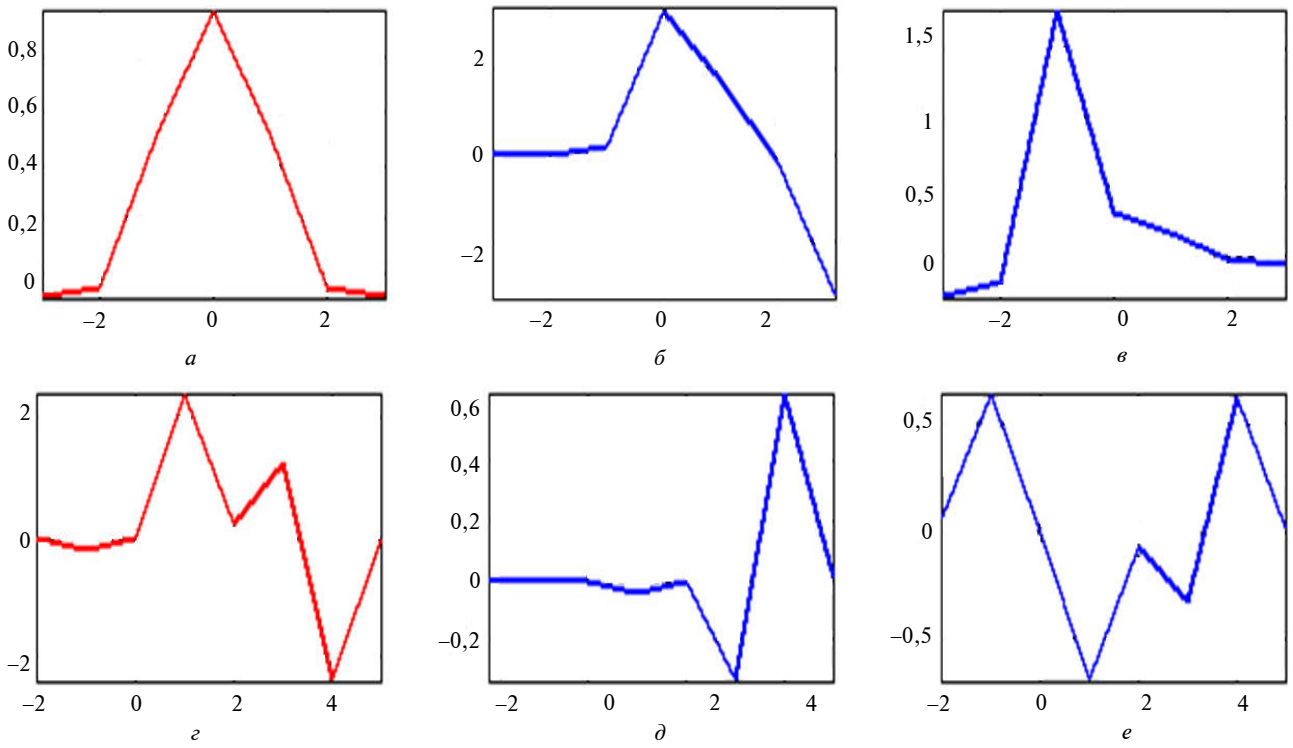

Рис. 3. Фільтри недіадних біортогональних сплайн-вейвлетів: *а* – низькочастотний фільтр декомпозиції; *б*, *в* – перший і другий високочастотні фільтри декомпозиції, відповідно; *г* – низькочастотний фільтр реконструкції; *д*, *е* – перший і другий високочастотні фільтри реконструкції, відповідно

недіадних масштабуючих та вейвлет-функцій не наводяться, оскільки каскадний процес їх побудови за методикою Добеші [2, с. 270–279] виявився розбіжним.

Неможливість збереження в недіадному дискретному вейвлет-перетворенні двох чи більше фільтрів існуючого діадного вейвлет-перетворення

Зберігати в недіадному дискретному ВП з існуючого діадного ВП лише фільтр чи фільтри реконструкції на практиці немає сенсу, бо в цьому випадку під час відновлення з відповідним коефіцієнтом масштабування будемо отримувати візуально схожими сигнали, різні за своєю структурою. Серед фільтрів декомпозиції зберігати потрібно насамперед низькочастотний фільтр, оскільки саме він відповідає за загальну поведінку сигналу і виділяє його згладжену (апроксимативну) складову. Наведеного вище конкретного прикладу відновлення тріадного ВП на основі діадних біортогональних сплайн-вейвлетів достатньо, щоб довести неможливість у загальному випадку збереження в недіадному дискретному ВП навіть двох фільтрів існуючого діадного ВП.

Для цього в табл. 4 зведено носії всіх фільтрів діадного та тріадного ВП із згаданого вище прикладу. Відповідно до табл. 1 всі коефіцієнти лівих та правих меж фільтрів діадних біортогональних сплайн-вейвлетів на носіях, вказаних у табл. 4, є ненульовими. Тому очевидно, що одночасно з низькочастотним фільтром декомпозиції в тріадному ВП може бути збережений лише високочастотний фільтр реконструкції за умови, що його коефіцієнт при 5-й степені полінома відповідної передатної функції дорівнює нулю, наприклад $G_1(z)$.

***Таблиця* 4.** Носії фільтрів біортогональних сплайн-вейвлетів

Тип ВП	Носії фільтрів			
	декомпозиції		реконструкції	
	низькочастотного	високочастотних	низькочастотного	високочастотних
Діадне	[–3,3]	[–5,3]	[–4,4]	[–2,4]
Тріадне	[–3,3]	[–3,3]	[–2,5]	[–2,5]

Але відповідно до формул (16), (17), якщо носій низькочастотного фільтра декомпозиції лежить в інтервалі від –3 до 3, то коефіцієнт при нульовій степені передатних функцій високочастотних фільтрів реконструкції тріадних вейвлетів завжди дорівнює нулю. Тому збереження значення цього коефіцієнта для діадного вейвлета в силу відомого [1, с. 284] зв'язку між фільтрами останнього дасть, що $h_{-1} = 0$. Тоді у формулах (16) і (17) занулиться доданок з $1/z$, що, в свою чергу, аналогічно призведе до рівності $h_{-2} = 0$. Проте одночасно коефіцієнти h_{-1} і h_{-2} не можуть дорівнювати нулю, бо тоді у формулах (13)–(17) отримаємо ділення на нуль через наявність у знаменнику множника, який дорівнює $h_{-2} h_2 - h_{-1} h_1$.

Отримане протиріччя доводить неможливість збереження в недіадному дискретному ВП хоча б ще одного фільтра існуючого діадного ВП, крім низькочастотного фільтра декомпозиції (розкладення).

Висновки

В результаті виконаних досліджень було встановлено, що неможливо зберегти в загальному випадку в недіадному дискретному ВП ще один фільтр відповідного діадного ВП, крім обов'язкового (з практичних міркувань) низькочастотного фільтра декомпозиції. Наведені викладки дають можливість сформувати для застосування на практиці найбільш загальний вигляд низькочастотних і високочастотних фільтрів декомпозиції та відновлення тріадного ВП за наявним діадним дискретним ВП.

У даній статті розглянутий лише одновимірний випадок, бо в просторах більшої розмірності можлива побудова несепарабельних (нерозділимих) ВП, тобто ВП, що не є тензорними добутками одновимірних фільтрів. Ця ситуація потребує окремого дослідження, бо відрізняється від проаналізованої в статті не стільки розмірністю, як більш складною структурою простору розкладення сигналів.

О.Р. Чертов

НЕДИАДНЫЕ ОДНОМЕРНЫЕ ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ

Доказана невозможность сохранения в общем случае в недиадном одномерном дискретном вейвлет-преобразовании еще одного фильтра соответствующего диадного вейвлет-преобразования, кроме низкочастотного фильтра декомпозиции.

O.R. Chertov

NON-DYADIC ONE-DIMENSIONAL WAVELET-TRANSFORMS

In this paper, we prove that a non-dyadic one-dimensional discrete wavelet-transform doesn't save any additional filter of the dyadic wavelet-transform, except the low-frequency filter of decomposition.

1. *Малла С.* Вейвлеты в обработке сигналов. – М.: Мир, 2005. – 672 с.
2. *Добеши И.* Десять лекций по вейвлетам. – Ижевск: НИЦ "Регулярная и хаотическая динамика", 2001. – 464 с.
3. *Park H., Woodburn C.* An algorithmic proof of Suslin's stability theorem for polynomial rings // J. of Algebra. – 1995. – **178**. – P. 277–298.
4. *Новиков И.Я., Протасов В.Ю., Скопина М.А.* Теория всплесков. – М.: Физматлит, 2006. – 616 с.
5. *Baussard A., Nicolier F., Truchetet F.* Rational multiresolution analysis and fast wavelet transform: application to wavelet shrinkage denoising // Signal Processing. – 2004. – **84**, N 10. – P. 1735–1747.
6. *Bratteli O., Jorgensen P.E.T.* Wavelet filters and infinite-dimensional unitary groups // Wavelet analysis and applications. – Providence (RI): Amer. Math. Soc., 2002. – P. 35–65 (AMS/IP Stud. Adv. Math.; V. 25).
7. *Cuntz J.* Simple C*-algebras generated by isometries // Comm. Math. Phys. – 1977. – **57**. – P. 173–185.
8. *Подкур П.Н.* О построении некоторых типов вейвлетов с коэффициентом масштабирования N // Электрон. науч. журн. "Исследовано в России". – 2007. – **93**. – С. 965–974. – http://zhurnal.ape.relarn.ru/articles/2007/093.pdf.
9. *Подкур П.Н.* О построении вейвлетов с коэффициентом масштабирования N на основе B-сплайнов // Там же. – **14**. – С. 128–138. – http:// zhurnal.ape.relarn.ru/articles/2007/014.pdf.

10. *Чертов О.Р., Мальчиков В.В.* Застосування неперервного вейвлет-перетворення для вибору коефіцієнта масштабування вейвлетів // Матер. XII Міжнар. наук. конф. ім. акад. М. Кравчука. I част., 15–17 травня 2008 р., Київ. – К.: ТОВ "Задруга", 2008. – С. 854.

11. *Чертов О.Р., Мальчиков В.В.* Выбор коэффициента масштабирования вейвлетов при кратномасштабном анализе сигналов // Матер. X Міжнар. наук.-техн. конф. "Системний аналіз і інформаційні технології" (САИТ-2008), 20–24 травня 2008 р. – К.: НТУУ "КПІ", 2008. – С. 151.

12. *Cohen A., Daubechies I., Feauveau J.-C.* Biorthogonal bases of compactly supported wavelets // Commun. on Pure and Appl. Math. – 1992. – **45**. – P. 485–560.

13. *Auscher P.* Ondelettes fractales et applications. Ph.D. Thesis. Universite IX Paris, Dauphine. – Paris, France, 1989.

14. *Жизняков А.Л.* Построение пирамид изображений с адаптивным выбором масштабного коэффициента // Искусственный интеллект. – 2006. – № 4. – С. 743–748.

15. *Алгоритмы* восстановления томографических изображений / А.Л. Жизняков, С.И. Семенов, Л.Т. Сушкова и др. // Информационно-измерительные и управляющие системы. – 2007. – **5**, № 9. – С. 29–38.

16. *Садовский М.А.* Естественная кусковатость горной породы // Докл. АН СССР. – 1979. – **247**, № 4. – С. 829–831.

17. *Садовский М.А., Писаренко В.Ф.* Сейсмический процесс в блоковой среде. – М.: Наука, 1991. – 96 с.

18. *Уломов В.И.* Решеточная модель очаговой сейсмичности и прогноз сейсмической опасности // Узбек. геолог. журн. – 1987. – № 6. – С. 20–25.

19. *Айзенберг Я.М., Хачиян Э.Е., Габричидзе Г.К. и др.* Международные строительные нормы СНГ. Строительство в сейсмических районах (Проект) 2002 г. // Сейсмостойкое строительство. Безопасность сооружений. – М.: Госстрой, 2002. – № 2. – С. 27–54.

Рекомендована Радою
факультету прикладної математики
НТУУ "КПІ"

Надійшла до редакції
17 вересня 2009 року

УДК 004.9

Чертов О. Р.

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ЛОКАЛІЗАЦІЇ І ПОШУКУ ПОМИЛОК У ПЕРВИННИХ ТА ЗВЕДЕНИХ ДЕМОГРАФІЧНИХ ДАНИХ

В статті описані інформаційні технології локалізації і пошуку помилок, що містяться в первинних та зведених демографічних даних. Викладення базується на успішному досвіді впровадження зазначених технологій в Україні та закордоном.

Ключові слова: *інформаційна технологія; демографічні дані; перепис.*

В статье описаны информационные технологии локализации и поиска ошибок, которые встречаются в первичных и сводных демографических данных. Изложение базируется на успешном опыте внедрения указанных технологий в Украине и за рубежом.

Ключевые слова: *информационная технология; демографические данные; перепись.*

The article describes information technologies of localization and search for errors that occur in primary and summary demographic data. The presentation is based on the successful experience of implementing these technologies in Ukraine and abroad.

Key words: *information technology, demographic data, census.*

Вступ

Сучасні системи обробки демографічної інформації характеризуються дуже великими обсягами даних, що оброблюються, та значною кількістю різнорідних правил контролю первинних та зведених даних [1, с. 76–113; 2]. Тому швидкість отримання результатів перепису населення чи іншого відповідного статистичного спостереження багато в чому залежить від застосовуємих інформаційних технологій локалізації та пошуку помилок в статистичних даних.

Загалом, поняття «інформаційна технологія» визначається в низці законодавчих та нормативних документів. Найбільш точне та лаконічне визначення міститься в ДСТУ 2226-93 [3]: «інформаційна технологія – технологічний процес, предметом перероблення й результатом якого є інформація». Більш розлого це поняття визначено в [4]: «інформаційна технологія – цілеспрямована організована сукупність інформаційних процесів з використанням засобів обчислювальної техніки, що забезпечують високу швидкість обробки даних, швидкий пошук інформації, розосередження даних, доступ до джерел інформації незалежно від місця їх розташування».

По аналогії з технологіями матеріального виробництва кожну інформаційну технологію можна розглядати як процес, що використовує сукупність засобів і методів збору, оброблення та передачі даних для отримання інформаційного продукту – інформації нової якості про стан об’єкту, процесу чи явища (див. рис.1) [5, с. 87]. Як зазначено в [6, с. 6], під інформаційними технологіями «починають розуміти будь-які засоби трансформації даних в корисний для досягнення мети системи управління інформаційний продукт».

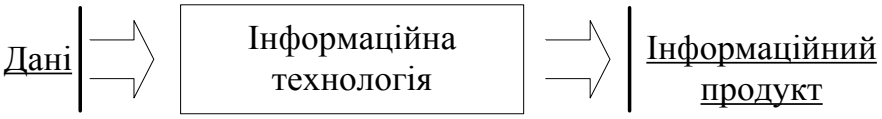

Рис. 1. Інформаційна технологія як процес перетворення даних в інформаційний продукт

Для ефективної організації всіх процесів інформаційної технології вимагається «визначення необхідних робіт (їх відповідного структурування, ув’язування за входом, виходом, термінами реалізації, виконавцями)» [6, с. 90]. Інформаційну технологію прийнято представляти у вигляді ієрархічної структури за різними рівнями [5, с. 91]: від більш загальних до більш детальних, наприклад, за рівнями етапів, операцій та дій.

З точки зору загального менеджменту виділяють три послідовні етапи процесу контролю [7, с. 440 – 449; 8, с. 83–85]:

1) встановлення стандартів;
2) зіставлення досягнутих результатів з встановленими стандартами;
3) прийняття необхідних коригуючих мір.

Враховуючи специфіку обробки демографічних первинних та зведених даних, зазначені етапи можна конкретизувати так, як це показано в табл. 1.

Етапи, виділені курсивом в табл. 1, є ключовими для пришвидшення виявлення та виправлення помилок в даних, оскільки ці етапи не можна повністю автоматизувати.

Таблиця 1

Етапи та особливості контролю демографічних даних

Загальні етапи процесу контролю	**Етапи контролю демографічних даних**	**Особливості етапів контролю демографічних даних**	
		первинних	**зведених**
встановлення стандартів	визначення правил контролю	визначення внутрішньобланкових та міжбланкових контролів	визначення внутрішньотабличних, міжрозрізних та міжтабличних контролів
зіставлення досягнутих результатів з встановленими стандартами	встановлення наявності помилок чи їх відсутності	в основному, арифметичний та логічний контроль	як правило, арифметичний контроль
	локалізація помилок, якщо вони виявлені	природно обмежується одним чи декількома (на рівні домогосподарства) бланками	як правило, дуже складно локалізувати помилку до рівня відповідного переписного документу
прийняття необхідних коригуючих заходів	*пошук причин виявлених помилок та їх усунення*	протокол контролю є основним джерелом інформації про помилку	помилки заборонено виправляти на зведених (агрегованих) даних

Класична технологія організації пошуку помилок в статистичних даних описана в [9, с. 91] в п. 3.4.3 «Організація автоматизованого розв'язання задач з використанням АРМ економіста-статистика»:

«… виконується перевірка правильності введеної інформації з використанням арифметичного і логічного контролю за формулами, які задаються економістами. Внаслідок контролю утворюється файл протоколу, який містить інформацію про виявлені помилки. Його можна переглянути на екрані або роздрукувати. Після аналізу помилок звіти можуть бути відкориговані.».

Дана технологія до початку XXI століття фактично була єдино можливою, але застосування для введення статистичних даних швидкодіючих промислових сканерів дозволяє організувати процес локалізації та пошуку помилок в первинних даних принципово більш зручним та ефективним для користувачів способом, який описано в наступному розділі.

Локалізація причини помилки в зведених даних, наприклад, коли загальна кількість респондентів певної групи в різних вихідних таблицях не співпадає, є нетривіальною справою, яка вимагає певних навичок та досвіду в побудові відповідних нерегламентних запитів до бази даних. Тому організація обміну досвідом між користувачами, що займаються пошуком помилок в зведених даних, може значно прискорити весь процес обробки наявних даних.

Постановка задачі

Об'єктом дослідження є процеси локалізації та пошуку помилок в демографічних даних – первинних (переписних бланках чи інших статистичних анкетах) та зведених (вихідних таблицях), що породжують проблему необхідності розширення функціональних можливостей інформаційних систем, які реалізують зазначені процеси. Предметом дослідження є інформаційні технології, які забезпечують організацію відповідної обробки демографічної інформації. Мета статті полягає в узагальнені та формалізації опису зазначених інформаційних технологій, впроваджених чи впроваджуємих за безпосередньої участі автора в різноманітних системах обробки даних статистики населення.

Інформаційна технологія локалізації і пошуку помилок в первинних даних. Розглянемо, як в автоматизованих системах «Перепис-2001» та «Перепис-Молдова 2004», які забезпечували обробку даних відповідно першого Всеукраїнського перепису населення 2001 р. та перепису населення Республіка Молдова в 2004 р., здійснювалися локалізація і пошук помилок під час вхідного контролю, тобто під час контролю первинних даних переписних документів [1, с. 95 – 102].

На рис. 2 представлена загальна схема технологічного процесу локалізації та пошуку помилок в первинних даних портфеля з переписними бланками.

Якщо в підсистемі проведення вхідного контролю в процесі контролю завантаженого без помилок портфеля в хоча б одному з переписних документів виявлено хоча б одну помилку, то відповідне поле в цьому документі, цей документ, а також усі документи, до яких цей документ належить, помічаються відповідними кольоровими позначками як помилкові. Ознайомившись з протоколом результатів вхідного контролю, користувач редагує помилкові переписні документи, які зберігаються в базі переписних, розрахункових та агрегованих даних. Під час редагування переписного документа користувач може бачити (за допомогою модуля інтерактивного і контекстного доступу до графічного образу просканованого документу) зображення графічного образу цього документа для порівняння змісту першоджерела з електронною копією, а також за допомогою модулів інтерактивного і контекстного доступу до довідників та класифікаторів і до протоколу помилок має контекстний доступ до відповідної інформації.

Фатальні помилки (червоний колір підсвічення помилкових переписних документів та показників на бланках вхідних форм в інтерфейсі користувача) необхідно обов'язково виправити. Для цього потрібно проаналізувати помилковий показник і поточну ситуацію (тобто проаналізувати залежність між показниками), при необхідності, звіривши введені переписні документи з їх графічними образами з архіву графічних образів. В деяких випадках фатальна помилка може вказувати на ситуацію, яка існує в реальному житті. В такому випадку необхідно зняти помилку.

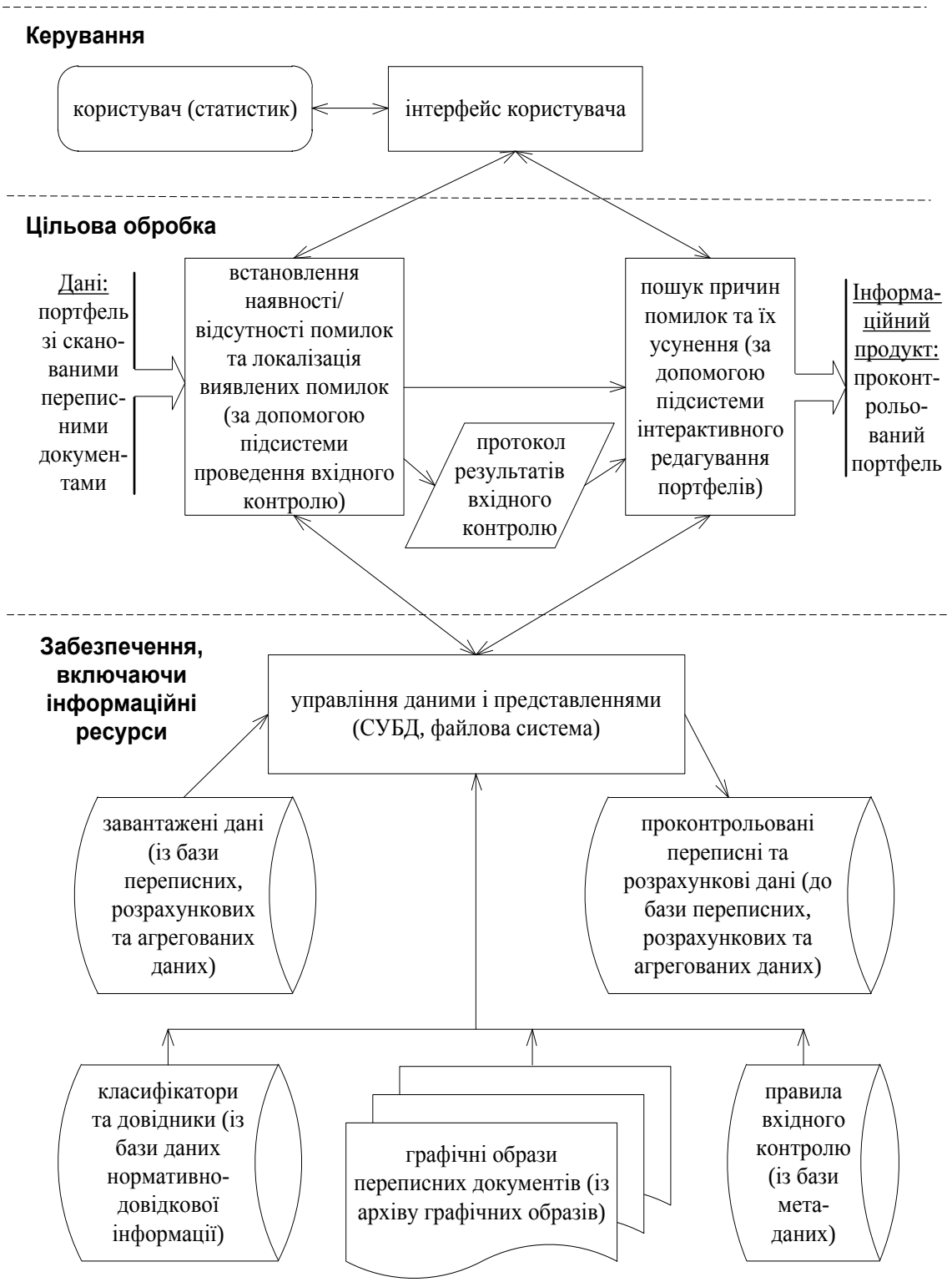

Рис. 2. Схема технологічного процесу локалізації та пошуку помилок в первинних даних портфелю

Нефатальна помилка (блакитний колір підсвічення помилкових переписних документів та показників на бланках вхідних форм в інтерфейсі користувача) вказує на таку ситуацію, що в загальному випадку може і не бути помилкою, і після звірення введених документів з їх графічними образами та аналізу показників не потребує коригування. Після виправлення помилок користувач знову проводить контроль обраних переписних документів.

В табл. 2 наведено формалізоване представлення інформаційної технології локалізації і пошуку помилок в переписних документах по первинним даним у вигляді ієрархічної структури за рівнями етапів, операцій та дій.

Таблиця 2

Етапи, операції та основні дії інформаційної технології локалізації і пошуку помилок в переписних документах по первинним даним

Етапи					
1. Встановлення наявності помилок чи їх відсутності		2. Локалізація помилок, якщо вони виявлені		3. Пошук причин виявлених помилок та їх усунення	
Операції	**Дії**	**Операції**	**Дії**	**Операції**	**Дії**
Визначення порядку застосування правил перевірки	Отримання із бази метаданих загального порядку застосування правил перевірки та рівня їх критичності	Зіставлення виявлених помилок і показників переписних документів	Отримання із бази метаданих відповідності між помилками і показниками, які можуть їх викликати	Перегляд переписних документів портфелю та результатів вхідного контролю	Перегляд певного переписного документу, значення показників котрого відмарковані відповідно до виявлених помилок
			Фіксація для виявленої помилки відповідного переписного документу		
	Вибір поточного правила перевірки в залежності від результатів попередніх перевірок		Фіксація для виявленої помилки відповідного аркушу переписного документу		Проглядання дерева перегляду переписних документів портфелю, відмаркованих до виявлених помилок
Застосування правил перевірки	Контроль повноти даних (по окремих переписних формах)	Маркування відсутності помилок чи виявлених помилок	Фіксація для виявленої помилки відповідного графічного образу переписного документу	Редагування переписного документу з метою усунення помилок	Контекстний доступ до графічного образу переписного документу
	Контроль форматів даних		Маркування в дереві перегляду позначками синього кольору переписних документів, значення показників котрих викликали нефатальні помилки		Контекстний доступ до протоколу результатів вхідного контролю
	Арифметичний контроль значень даних				Контекстний доступ до нормативно-довідкової інформації
	Логічний контроль значень та наявності даних				Зміна значення конкретного показника переписного документу
	Контроль відповідності даних класифікаторам та довідникам		Маркування в дереві перегляду позначками червоного кольору переписних документів, значення показників котрих викликали фатальні помилки		Маркування в дереві перегляду жовтим кольором переписного документу, в якому було змінено значення показника
Фіксація результатів перевірки з урахуванням критичності помилки	Фіксація результату перевірки — має місце відповідна помилка чи ні	Формування протоколу результатів вхідного контролю	Маркування синім кольором значень показників, що викликали нефатальні помилки	Зняття фатальної помилки контролю	Маркування жовтим кольором зміненого значення показника
	Визначення рівня критичності помилки, якщо він не є фіксованим		Маркування червоним кольором значень показників, що викликали фатальні помилки		Автокоригування значень показників переписного документа
			Запис в протокол контролю повідомлення про виявлену помилку та показники, значення котрих її викликало		Збереження відредагованих (чи всіх поточних) значень показників
			Запис в протокол контролю загальної інформації про виявлені помилки чи їх відсутність		Вибір в протоколі вхідного контролю фатальної помилки, яка по факту наявних даних не є помилкою, та зняття її
					Фіксація інформації про зняту фатальну помилку

Інформаційна технологія локалізації і пошуку помилок в зведених даних (вихідних таблицях)

Пошук причини помилки, виявленої на етапі проведення контролів зведених даних, зокрема, вихідних таблиць (ВТ), на практиці може бути виконаний лише за допомогою підсистеми підтримки нерегламентних запитів, яка дозволяє за певним розрізом отримати визначений розподіл даних з БД.

З часом, з накопиченням досвіду роботи ряд користувачів придумують певні шаблонні нерегламентні запити, котрі дозволяють швидко знаходити типові причини, які приводять до появи помилок у вихідних таблицях під час їх перевірки за допомогою внутрішньотабличних, міжрозрізних чи міжтабличних контролів. Тому сильно зростає значення підсистеми підтримки адаптивності, котра як би бере на себе роль по передачі досвіду з побудови нерегламентних запитів.

Ключовою особливістю застосування адаптивних технологій є можливість проведення аналізу діяльності великої кількості користувачів одночасно. Результати цього аналізу можуть використовуватися при налагодженні функціональності для конкретного

користувача. Адаптація на основі колективної взаємодії значно привабливіша за адаптацію на основі тільки особистого досвіду, оскільки в будь-який момент часу можна знайти готове рішення, прийняте на базі вже накопиченого досвіду [10, с. 99].

Розглянемо, як в автоматизованих системах з обробки демографічних даних, розроблених під керівництвом автора, здійснювалися локалізація і пошук помилок в зведених даних.

На рис. 3 представлена загальна схема технологічного процесу локалізації та пошуку помилок в зведених даних.

Якщо в підсистемі проведення контролю матриць вихідних таблиць в процесі контролю зведених даних по якомусь об'єкту адміністративно-територіальному устрою виявлено хоча б одну помилку, то відповідна інформація записується до протоколу результатів контролю матриць вихідних таблиць.

Але на відміну від аналогічної ситуації під час обробки первинних даних протокол результатів контролю зведених даних не дозволяє локалізувати помилку безпосередньо в переписних документах. Тому як для локалізації зазначених помилок, так і для пошуку причин їх виникнення потрібно застосовувати підсистему підтримки нерегламентних запитів.

Дана підсистема повинна реалізовувати наступні функції: формування та зберігання нового запиту, редагування, копіювання, вилучення чи виконання існуючого запиту.

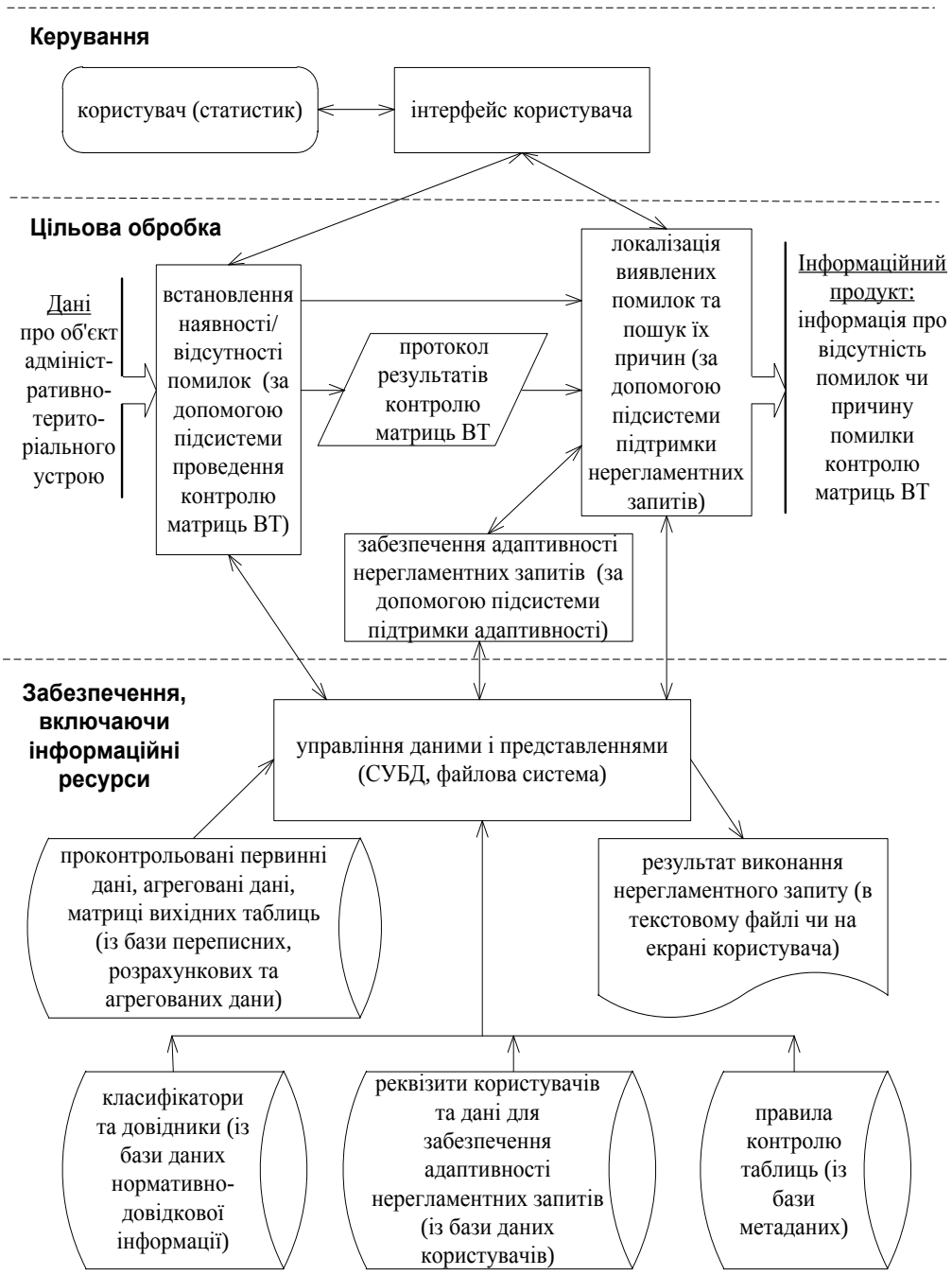

Рис. 3. Схема технологічного процесу локалізації та пошуку помилок в зведених даних

Під час формування нового нерегламентного запиту відбувається автоматичне визначення автора запиту та дати і часу формування запиту. Окрім того, вручну вводиться опис запиту (шифр запиту та анотація дій, що будуть виконуватися).

Кожен нерегламентний запит повинен містити, як мінімум, одне поле із хоча б однієї таблиці БД, що зберігають наступні дані: класифікатори і локальні довідники, первинні дані, тобто дані переписних документів, та розрахункові показники, агреговані дані.

Приклад сформованого в системі «Перепис-2001» нерегламентного запиту для пошуку респондентів, що не вказали свою національність в переписній формі (бланку) 2С, наведено на рис. 4. Фактично підсистема підтримки нерегламентних запитів дозволяє нефахівцеві в області програмування візуально побудувати SQL-запит до БД.

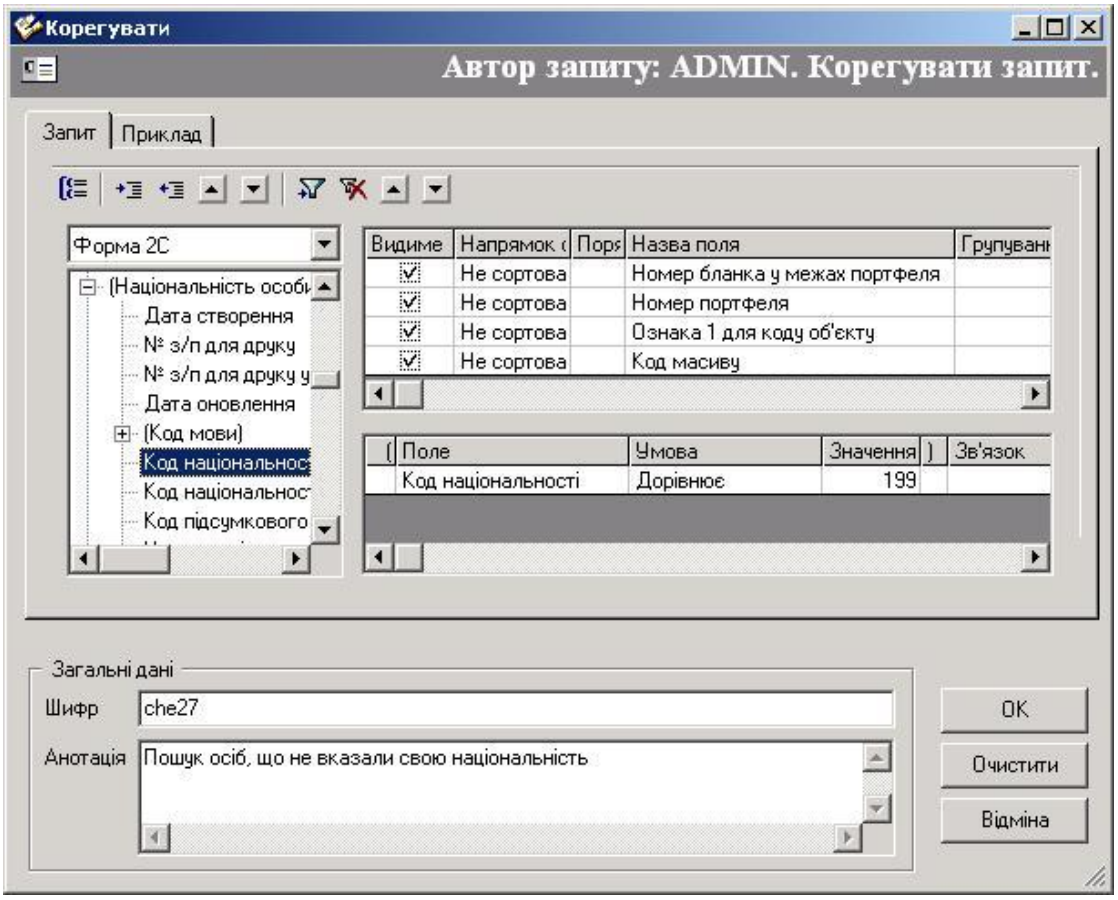

Рис. 4. Приклад нерегламентного запиту в системі «Перепис-2001»

Результат виконання нерегламентного запиту записується до текстового файлу чи видається на екран користувача.

Оскільки заборонено коригувати зведені дані, то за наявності помилок в вихідних таблицях пошук можливих причин цих помилок в первинних даних за допомогою підсистеми підтримки нерегламентних запитів ітеративно чергується з внесенням відповідних змін до первинних документів за допомогою підсистеми інтерактивного редагування портфелів, поки всі помилки в вихідних таблицях не будуть виправлені.

В загальному випадку причиною спрацьовування внутрішньотабличного, міжтабличного чи міжрозрізного контролю, можуть бути наступні помилки:

– помилка, викликана наявністю знятої фатальної помилки;

– помилка, пропущена під час верифікації (на етапі сканування переписних документів) та/чи під час вхідних контролів через похибку в постановці задачі;

– помилка, викликана реальним існуванням певної комбінації значень показників переписних документів, не передбаченої в постановці задачі;

– помилка в програмній реалізації (інформаційної системи обробки переписних даних, СКБД тощо).

Хоча потенційно помилок в зведених даних набагато менше, ніж в первинних даних, але їх локалізація та пошук причин виникнення – значно складніші. Застосування адаптивної підтримки співробітництва користувачів [10] під час побудови нерегламентних запитів за допомогою підсистеми підтримки адаптивності дозволяє як полегшити новачкам початок їх фахової роботи з пошуку причин помилок у вихідних таблицях, так і покращити взаємодію та обмін досвідом інших фахівців, зокрема, за рахунок надання позитивних відгуків на певні нерегламентні запити.

В табл. 3 наведено формалізоване представлення інформаційної технології локалізації і пошуку помилок в зведених даних із забезпеченням адаптивної підтримки співробітництва користувачів у вигляді ієрархічної структури за рівнями етапів, операцій та дій.

Таблиця 3.

Етапи, операції та основні дії інформаційної технології локалізації і пошуку помилок в зведених даних із забезпеченням адаптивної підтримки співробітництва користувачів

<table>
<tr><td colspan="6">Етапи</td></tr>
<tr><td colspan="2">1. Встановлення наявності (відсутності) помилок, локалізація виявлених помилок на рівні зведених даних</td><td colspan="2">2. Локалізація виявлених помилок на рівні переписних документів, пошук причин виявлених помилок</td><td colspan="2">3. Забезпечення адаптивності нерегламентних запитів</td></tr>
<tr><td>Операції</td><td>Дії</td><td>Операції</td><td>Дії</td><td>Операції</td><td>Дії</td></tr>
<tr><td rowspan="2">Визначення порядку застосування правил перевірки</td><td>Отримання із бази метаданих загального порядку застосування правил перевірки</td><td rowspan="3">Перегляд нерегламентних запитів та протоколу результатів контролю ВТ</td><td>Перегляд певного нерегламентного запиту</td><td rowspan="4">Призначення групи для користувача</td><td rowspan="2">Автоматичний перерозподіл користувачів по групам</td></tr>
<tr><td>Вибір поточного правила перевірки в залежності від результатів попередніх перевірок</td><td>Проглядання дерева перегляду нерегламентних запитів</td></tr>
<tr><td rowspan="2">Застосування правил перевірки</td><td>Внутрішньотабличний контроль</td><td>Контекстний доступ до протоколу результатів контролю ВТ</td><td rowspan="2">Ручне призначення групи (наприклад, для новачків)</td></tr>
<tr><td>Міжрозрізний контроль</td><td rowspan="3">Формування/ редагування нерегламентного запиту</td><td>Копіювання існуючого нерегламентного запиту</td></tr>
<tr><td>Міжтабличний контроль</td><td>Визначення службових реквізитів запиту (автор запиту, дата і час формування чи зміни запиту тощо)</td><td rowspan="3">Формування даних для забезпечення адаптивності нерегламентних запитів</td><td>Автоматичний підрахунок рангу запита</td></tr>
<tr><td>Фіксація результатів перевірки</td><td>Фіксація результату перевірки — чи має місце відповідна помилка</td><td>Визначення для запиту поля з таблиці БД, що зберігає класифікатори і локальні довідники, дані переписних документів та розрахункові показники чи агреговані дані</td><td>Ручне визначення користувачем рангу поточного нерегламентного запиту</td></tr>
<tr><td rowspan="3"></td><td>Фіксація комірки, (або рядка чи графи), розрізу і таблиці, в яких була виявлена помилка</td><td rowspan="5"></td><td>Редагування атрибутів поля запиту («Видиме», «Напрямок сортування», «Порядок сортування»)</td><td></td><td>Підрахунок середнього часу на пошук помилки в одній ВТ із певного класу ВТ конкретним користувачем</td></tr>
<tr><td>Маркування в дереві перегляду розрізів ВТ, які пройшли контролі (з помилками чи без них — відповідно)</td><td>Визначення функції агрегування при групуванні по полю запита («Група», «Найбільше», «Найменше», «Середнє», «Кількість»))</td><td rowspan="4">Адаптивний доступ до нерегламентних запитів користувачів групи</td><td rowspan="2">Впорядкування запитів за автоматично сформованими рангами</td></tr>
<tr><td rowspan="2">Формування протоколу результатів контролю ВТ</td><td rowspan="2">Запис в протокол контролю повідомлення про виявлену помилку та розташування в розрізі ВТ помилкового значення</td><td>Редагування списку умов, що накладаються на поле запиту (з визначенням атрибутів «Поле», «Умова», «Значення», «Зв'язок» та дужок)</td></tr>
<tr><td>Зберігання чи вилучення запиту</td><td>Впорядкування запитів за рангами, наданими користувачами групи</td></tr>
<tr><td>Запис в протокол контролю загальної інформації про виявлені помилки чи їх відсутність</td><td rowspan="2">Виконання нерегламентного запиту та перегляд результату</td><td>Запуск нерегламент-ного запиту на виконання</td><td rowspan="2">Впорядкування запитів за особисто визначеними рангами</td></tr>
<tr><td>Перегляд результату виконання запиту</td></tr>
</table>

Висновки. Системи обробки демографічної інформації, насамперед, – переписних даних, є великими масштабними інформа-ційними системами, в яких отримання остаточних результатів перепису може тривати роками. Ефективна реалізація локалізації та пошуку помилок в первинних і зведених даних, тобто найбільш трудомістких операцій під час обробки даних перепису, може значно покращити кінцеві результати та прискорити терміни їх отримання. Наприклад, описані в статті відповідні інформаційні технології дозволили ліквідувати більш ніж трьохмісячне відставання, що виникло із-за організаційних причин під час обробки даних Всеукраїнського перепису населення 2001 р.

Запропонований в статті формалізований опис інформаційних технологій локалізації та пошуку помилок дозволив з єдиних позицій підійти до їх реалізації як при обробці первинних, так і зведених демографічних даних. Відповідні інформаційні технології наразі застосовуються в Державній службі статистики України для обробки даних пробного перепису населення України 2010 р. про 98,5 тисяч респондентів Дергачівського району Харківської області.

ЛІТЕРАТУРА

1. Информационные технологии: приоритетные направления развития : монография / [Л. Н. Абуталипова, А. Г. Гусейнов, А. С. Дулесов и др.] ; под общ. ред. О. Р. Чертова. – Новосибирск : СИБПРИНТ, 2010. – Кн. 4. – 194 с.
2. Чертов О. Р. Учетные и аналитические информационные системы в области статистики населения / О. Р. Чертов // Наукові праці. – Миколаїв : Вид-во ЧДУ ім. Петра Могили, 2010. – Т. 134. Вип. 121. Комп'ютерні технології. – С. 225—229.
3. Автоматизовані системи. Терміни та визначення : ДСТУ 2226-93. – К. : Держстандарт України, 1994. – 94 с.
4. Про Національну програму інформатизації [Текст]: закон України від 4 лютого 1998 року № 74/98-ВР // Відомості Верховної Ради України. – 1998. – № 27. – С. 181.
5. Информатика : Учебник / [Макарова Н. В., Матвеев Л. А., Бройдо В. Л. и др.] ; под ред. Н. В. Макаровой. – [3-е изд., перераб.]. – М. : Финансы и статистика, 2009. – 768 с.
6. Павлов А. А. Информационные технологии и алгоритмизация в управлении / А. А. Павлов, С. Ф. Теленик. – К. : Техніка, 2002. – 344 с.
7. Мескон М. Х. Основы менеджмента : пер. с англ. / М. Х. Мескон, М. Альберт, Ф. Хедоури. – [2-е изд.]. – М. : Дело, 2001. – 800 с.
8. Кабушкин Н. Л. Основы менеджмента : учеб. пособие / Н. Л. Кабушкин. – [11-е изд., испр.]. – М. : Новое знание, 2009. – 336 с.
9. Годун В. М. Інформаційні системи і технології в статистиці : навч. посібник / В. М. Годун, Н. С. Орленко, М. А. Сендзюк ; за ред. В. Ф. Ситника. – К. : КНЕУ, 2003. – 267 с.
10. Чертов О. Р. Адаптивная поддержка сотрудничества при поиске информации / О. Р. Чертов, Д. В. Райчук // Штучний інтелект. – 2009. – № 3. – С. 97—104.

© Чертов О. Р., 2012

Дата надходження статті до редколегії 03.05.2012 р.

ЧЕРТОВ О. Р. – к.т.н., доцент, доцент кафедри прикладної математики Національного Технічного Університету України «Київський політехнічний інститут».