Tamara SAVCHUK[1], Sergiy PETRISHYN[1], Laura SUGUROVA[2], Andrzej SMOLARZ[3]

Vinnitsa National Technical University (1), Kazakh National Technical University (2), Lublin University of Technology (3)

Identification of technogenic emergency situations in railway transport using cluster analysis

Abstract. *The antropogenic load on natural environment is continuously growing. One of important issues is the influence of transport of goods, especially when the cargo is hazardous. Railroad transport in Ukraine shares about 60% of all means of transport so identification of potential emergency situations in this area is an important issue. The article discusses some methodology approaches and describes cluster analysis as a tool for Identification of technogenic emergency situations in railway transport.*

Streszczenie. *Antropogeniczne obciążenie środowiska naturalnego stale rośnie. Jedną z istotnych kwestii jest wpływ transportu towarów, zwłaszcza gdy ładunek jest niebezpieczny. Transport kolejowy na Ukrainie ma około 60% udziału wszystkich środków transportu, więc identyfikacja potencjalnych sytuacji awaryjnych w tej dziedzinie jest ważnym zagadnieniem. W artykule przedstawiono wybrane podejścia metodyczne i opisano analizę skupień jako narzędzie do identyfikacji technogennych sytuacji awaryjnych w transporcie kolejowym.* ***Identyfikacja technogennych sytuacji awaryjnych w transporcie kolejowym***

Keywords: cluster analysis, k-means algorithm, railway transport.
Słowa kluczowe: analiza skupień, algorytm k-średnich, transport kolejowy.

doi:10.12915/pe.2014.11.46

Introduction

Growing anthropogenic loading on the environment causes a dramatic worsening of the ecological situation, bringing humanity to a critical point in its relationship with nature and raises a question about the possibility of survival [1].

World's environment if falling into the state of crisis due to the insufficiently deliberate strategy of the state concerning environmental safety. Consequently, every year more and more material and financial resources are directed at the elimination of natural and anthropogenic disasters. One of the troubling issues arise when transporting dangerous / hazardous cargo by rail [2].

Railroad transport is now an important part of the economy in Ukraine, its share is 60% of transportation (including transportation of hazardous and dangerous goods), carried out in the country by all means of transport [3].

The increase in rail freight traffic, including hazardous and dangerous cargo, analyzes the problems of information analysis about emergencies that may occur during the transportation of this cargo.

Since the data on the emergencies in railroad transport is often unclear and unstructured, it is appropriate to use Data Mining technologies for their analysis with the peculiarities of the studied subject, and as a result, the synthesis of new information technology analysis of an emergency on the railroad that takes into account both its characteristics and peculiarities of the stage of development [4].

Thus, the relevance of information technology analysis of emergencies in the railway transport, based on the analysis of the situation that has developed as a result of such situations, using modern intelligent technologies necessitates automation of operational management of the emergency arising from the dangerous cargo trafficking in order to reduce the time for their analysis and increase the objectivity and effectiveness of management decisions.

Current level of means and methods for identification of emergency in railways

Rapid development of industry, the development and expansion of cities and the development of railway transport, increased the possibility of emergencies. Therefore, it is necessary to implement technology that intelligently analyzes emergencies to increase their rate of elimination, minimize their consequences in a short period of time. This task requires rapid processing of large amounts of information. Among the newest technologies that can quickly and efficiently process large volumes of information are: KDD, OLAP, Data Mining [4-5].

Depending on the nature of the problem that is being solved (problem description and prediction problems) [6], all data analysis algorithms are divided into supervised learning (learning with teacher) and unsupervised learning (learning without a teacher). In the first case the problem is solved by several steps: data models under analysis is built using specific algorithm. The synthesized model learns until it begins to work correctly. Unsupervised learning is used when there is no prior knowledge of the analyzed data [6].

The main tasks of the Data Mining are: classification, regression, association rules search and clustering (Table 1) [2, 5]. Lets consider them on the example of the analysis of emergency situations on the railroad.

The task of classification is reduced to the determination of the class of an emergency on the railway based on its characteristics. In this problem, a set of such situations' classes, which can be classified as an object of study, is known in advance [5]. It should also be noted that there is a number of disadvantages when it comes to using classification [1]:

1) the number of the training samples must be large enough;
2) the study sample should include emergency situations on the railway, representing all the problematic classes in the analysis of such situations;
3) there must be a sufficient set of emergencies in the study sample for each class, which is difficult to obtain in the analysis of NA emergencies on railways;
4) the problem of overfitting, the essence of which is that the classification function is highly adaptable to the data, and if among them there are errors or abnormal value, the function interprets them as part of the internal data structure that is inappropriate for analyzing of emergency situations;
5) the problem of underfitting, which occurs when checking classifier revealed a large number of errors that are unacceptable to the subject area, which is analyzed.

The problem of regression as well as the problem of classification allows to determine the value of a parameters for the known characteristics of an emergency on the railway. Unlike the classification, the value of this parameter is a set of real numbers, but not the set of classes, which is irrelevant to the analysis of emergency in railroad transport [3].

Table 1. Characteristics of Data Mining methods, which may be used in the analysis of emergency in rail transport

Method name	The essence of the method	Advantages of the method	Disadvantages of the method
Classification	The class definition of an emergency on the railway from its prominent characteristics.	Easy to use, the presence of a large number of effective approaches to solving this problem.	Power training sample must be large enough, the learning sample should include emergencies, representing all classes, for each class must be sufficiently powerful set of emergency situations in the study sample, the problem of overfitting; issue underfitting.
Regression	Allows the known characteristics of an emergency on the railway, to determine the value of a parameter.	Easy to use, the presence of a large number of approaches to solving this problem.	Failure to identify a disaster on the railway.
Finding association rules	Determining dependencies frequent among emergency.	The possibility of finding certain patterns between emergencies; easy perceptions rules; simple interpretation of programming languages.	Rules are not always useful, as there are three types of association rules: useful, trivial, unintelligible.
Clustering	Finding clusters in independent set of data analyzed emergencies on the railway.	Iterative search optimal results, the possibility of using the methods of formation of clusters, selecting informative features and proximity measures, building multidimensional classification of observations based on the selected set of indicators and identify hidden links between emergency on the railroad.	Defining the input number of iterations in finding solutions.

The advantages of finding association rules is that they allow to find certain patterns specific to emergencies that are relevant for this subject area. Also the complexity of human perception of these rules and their interpretation by programming languages is irrelevant in the analysis of such situations [7].

The objective of finding the associative rules is the determination of the dependences which often repeat among the emergency situations. The found dependences are presented as the rules and may be used both, for better understanding of the nature of the data under analysis, as well as for the prediction of certain events.

The advantages of finding associative rules is that they allow to find certain patterns specific to emergencies that are relevant to this subject area. And the complexity of human perception and interpretation of the rules of programming languages is irrelevant in the analysis of such situations [7].

The disadvantages of finding associative rules are that the rules found in the analysis of emergency are not always useful. There are three types of association rules: useful, trivial, obscure, which is unacceptable for the analysis of emergency on the railroad.

The task of clustering is to find clusters in the set of independent data on the emergency of the railroad transport, which are analyzed. It helps to understand the data. In addition, the grouping of similar data can reduce their number to simplify the analysis further.

The advantages of clustering is an iterative search for the optimal results, which increases the probability of finding such a solution, the possibility of using the methods of formation of clusters and select informative features and measures of proximity between two objects, and object cluster, two clusters that are relevant in the analysis of emergency situations on railways, building multidimensional classification of observations based on the selected set of indicators and identifying the internal connections between emergencies under analysis [6].

The complexity of clustering is regulated by the definition of the input number of iterations in finding solutions which is essential to the subject area under analysis, since this is the way to determine the accuracy of predicted results of the identification algorithm emergency, subject analysis, and the state of its development [8, 9].

Thus, the result of the analysis allow to make a conclusion as for the expediency of using the Data Mining for the evaluation of the technogenic emergency situation, which arise during the transportation of hazardous materials by rail.

Problem statement of emergency identification in railways using cluster analysis

In formalized form, the problem of emergency identification is reduced to the determination of its relation to the current state (causes, conditions and factors) to one of the formal conditions set D_i from the set D. It analyzes the factors of the situation X_i, which are widely used in practice, and reflect the current state of the j-th status i-th emergency S_{ij} [3].

Let put Y - matrix in which each column $\{y_{i1},...,y_{ij},...y_{im}\}$ describes certain emergency, ie y_{ij} is certain characteristics of a particular emergency situation (1).

$$Y = \{Y_1, Y_2, ..., Y_n\} = \begin{Bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{Bmatrix}, \quad (1)$$

where Y_i - specific emergency situation on the railway transportation;

y_{ij} - the value of a particular j parameter of i-th emergency situation;

m - the number of parameters of emergencies that are stored in the database.

The above allows to formulate the problem of identifying emergency situation and its development as follows: a set of emergency situations on the railways $Y = \{Y_i\}\ (i = \overline{1,n})$, presented in a matrix Y. Each of these situations has m characteristics. We split the set S into k ($k \leq n$) clusters so that a particular emergency situation Y_i belongs to one and only one cluster, with the following conditions [10]:

- emergencies that belong to the same cluster should be as similar as possible;

- emergency belonging to different clusters should be treated as dissimilar as possible.

Analysis of the emergency situations on the railroad transport as the task of clustering Data Mining

Prior to clustering of emergencies on the railway transport it is necessary to evaluate its various methods and decide which of them to prefer that the result was reliable. Choosing between the hierarchical and non-hierarchical methods, it is necessary to consider the following features [11-14]:

- Non-hierarchical methods show high resistance to noise and discharges, improper metric selection, insignificant variables in the set which is involved in clustering, which is relevant for the analysis of emergency on the railway transport for the purpose of identification, because the data will be entered as on line, which increases the probability of errors in data under processing. Also, the advantage of such methods is that they help process huge database, which is also necessary for processing information about emergencies, as in larger sample might get more reliable solution. However, the main disadvantage of non-hierarchical methods is that the input of the algorithm needs to have predefined number of clusters or the number of iterations, although the analysis to identify possible emergencies accurately determine the number of clusters, because only one current emergency situation in rail transport will be analyzed, and all other members sample will be taken from the data warehouse and the number of clusters to be determined.
- If there is no assumption about the number of clusters involved in emergencies on the railway or the number of iterations in the cluster analysis, use hierarchical algorithms. However, if handled by a powerful database, a possible way to solve a set is a series of experiments with different number of clusters, such as start partitioning the data set of the two groups, and gradually increasing their number of experiments to compare the results with each other. Due to this change the results of cluster analysis is achieved quite a Flexibility clustering. Hierarchical methods, as opposed to non-hierarchical, involving the rejection of determining the number of clusters and build a complete tree of nested clusters. The complexity of hierarchical clustering methods: limitation of the data set as for processing large amounts of data to store and handle many times the similarity matrix, which is unacceptable for a given subject area, the choice of proximity measure, which is a tricky task, with high probability of errors in this stage; inflexibility obtained classifications. The advantage of this method compared to non-hierarchical methods - their visibility and to get a detailed view of the data structure, which is also important for the analysis to identify the emergency of the railways, as well, you can track all processes combination / separation of clusters and therefore perhaps the wrong track split / connections / in clusters, which is useful for developers. But this feature is not principal for users, as only the result is important.

The results of the analysis of clustering methods was found that acceptable method of cluster analysis for the solution of this problem is the non-hierarchical method is characterized by:

- using the minimum possible number of characteristics and parameters of emergency in rail transport in the performance of their clustering and thus obtain the highest possible quality analysis;
- to provide the highest quality analysis using a small number of already clusterized emergencies on the railway, all data is contained in a data warehouse, but in turn the opportunity to work with a powerful database;
- simplicity;
- clarity;
- high performance.

Therefore, we analyzed the clustering methods and found that the appropriate method of cluster analysis for the solution of this problem is the non-hierarchical method.

The distance and the degree of intimacy as basic characteristics predictive analysis emergencies in rail transport

When cluster analysis of technological emergencies problem determining the distance between individual situations such as emergencies values of the railways have different units and different weights, and for reliable clustering to accurately determine the similarity of each pair of such situations. The main difficulties encountered in this [4]:

- ambiguity in choice of normalization;
- ambiguity in determining the distance between objects.

If there are certain similarities between true emergencies in rail transport will be conducted significant splitting their set into clusters, and thus made accurate identification of each specific emergency taken on the railway.

In general uniformity of two-and i-th j-th emergencies in rail transport is determined by objective rules of calculating the quantity ψ_{ij} characterizing or distance $a(Y_i, Y_j)$ between objects Y_i and Y_j a researched set of emergency in rail transport (2)

$$Y = \{Y_i\} \ (i = \overline{1,n}) \tag{2}$$

or the degree of closeness $\omega(Y_i, Y_j)$ between the same situations.

If set function, approximated this metric $a(Y_i, Y_j)$ emergencies are homogeneous, ie those that belong to the same cluster. But this should be compared $a(Y_i, Y_j)$ with a certain threshold, as may be the case that the cluster belongs to only one emergency situation on the railway.

Definitions approach is useful in determining the degree of intimacy $\omega(Y_i, Y_j)$ in the formation of homogeneous taxons emergencies on the railway. This must be met such requirements [4]:

- the requirement of symmetry ($\omega(Y_i, Y_j) = \omega(Y_j, Y_i)$);
- the requirement of maximum similarity emergencies themselves with a ($\omega(Y_i, Y_i) = \max(\omega(Y_i, Y_j))$);
- requiring consistency between the distance between emergency on the railway and measure the proximity between them (if $a(Y_1, Y_2) \geq a(Y_2, Y_3)$ that $\omega(Y_1, Y_2) \leq \omega(Y_2, Y_3)$) [15].

Characterization of cluster analysis emergencies in rail transport

Distance between emergency Y_i and Y_j used in the cluster analysis to identify such situations is called non-negative real function $a(Y_i, Y_j)$ which has the following properties [4]:

1) $a(Y_i, Y_j) \geq 0$ for all Y_i and Y_j from the set $Y = \{Y_i\} \ (i = \overline{1,n})$;
2) $a(Y_i, Y_j) = 0$ if $Y_i = Y_j$;

3) $a(Y_i,Y_j)=a(Y_j,Y_i)$;

4) $a(Y_i,Y_j)\le a(Y_i,Y_k)+a(Y_k,Y_j)$, where, Y_i, Y_j and Y_k - any three emergencies on the railway from the set $Y=\{Y_i\}\,(i=\overline{1,n})$.

Distances between emergency and railways provide for their representation in the form of points *m*-dimensional space.

For the reliable cluster analysis emergencies on the railway for the purpose of identification is necessary to consider certain characteristics of input data (the possibility of emission weighting component input vector data, etc.), It needs to use its various characteristics.

Euclidean distance as one of the commonly used metrics in cluster analysis, corresponds to the intuitive notion of proximity and can be defined by the expression (3) [3]:

$$a_E(Y_i,Y_j)=\sqrt{(y_{i1}-y_{j1})^2+(y_{i2}-y_{j2})^2+...+(y_{im}-y_{jm})^2}, \quad (3)$$

where $a_E(Y_i,Y_j)$ - the Euclidean distance between two emergencies on the railway Y_i and Y_j;

y_{i1}, y_{i2}, …, y_{im} - vector of values of characteristics that describes th emergency situation on the railway;

y_{j1}, y_{j2}, …, y_{jm} - vector of values of characteristics that describes th emergency situation on the railway.

Dan metrics are useful in the following cases:

- parameter values y_{i1}, y_{i2}, …, y_{im} homogeneous physical meaning, and found that all of them are equally important in terms of solving the problem of classifying emergencies on the railway to a cluster;
- space characteristics coincides with the geometric space of reality and the notion of proximity emergency coincides with the notion of geometric proximity in this space.

To distinguish distant objects using squared Euclidean distance (4) [4]:

$$a_E(Y_i,Y_j)^2=(y_{i1}-y_{j1})^2+(y_{i2}-y_{j2})^2+...+(y_{im}-y_{jm})^2. \quad (4)$$

If necessary, consideration of "importance" λ_l of each *i*-th characteristics of emergency in rail transport (for example, the temperature in the tank for transportation of flammable substances and humidity) to be proportional to weight ratio in terms of assigning a particular emergency situation on the railway to a particular cluster advisable to use a weighted Euclidean distance (5) [4]:

$$\left[a_{зE}(Y_i,Y_j)\right]^2=\lambda_1\cdot(y_{i1}-y_{j1})^2+\lambda_2\cdot(y_{i2}-y_{j2})^2+...+\lambda_m\cdot(y_{im}-y_{jm})^2, \quad (5)$$

where $a_{зE}(Y_i,Y_j)$ - the weighted Euclidean distance between emergency Y_i and Y_j;

λ_1, λ_2, …, λ_m $(0\le\lambda_l\le 1(l=\overline{1,m}))$ - vector of values of weighting coefficients that meet specifications y_1, y_2, …, y_m emergencies in rail transport;

y_{i1}, y_{i2}, …, y_{im} - vector of values of characteristics that describes *i*-th emergency situation on the railway;

y_{j1}, y_{j2}, …, y_{jm} - vector of values of characteristics that describes *j*-th emergency situation on the railway.

To determine the vector of values of weights λ_1, λ_2, …, λ_m using training samples of emergencies on the railway or the experience of experts. Attempts to determine the weights λ_1, λ_2, …, λ_m only on the information contained in the input data does not give the desired result and may increase the accuracy of the result.

Distance from Hemet module is the difference coordinates. In most cases, this measure of closeness leads to the same results as the Euclidean distance, but its impact large emission decreases as they rise to the square, which may also be useful in emergencies cluster analysis to identify them.

General view of the formula for the Hamming distance has the form (6) [15]:

$$a_{\mathrm{H}}(Y_i,Y_j)=|y_{i1}-y_{j1}|+|y_{i2}-y_{j2}|+...+|y_{im}-y_{jm}|, \quad (6)$$

where $a_H(Y_i,Y_j)$ - according to Hamming distance between two emergencies on the railway Y_i and Y_j;

y_{i1}, y_{i2}, …, y_{im} - vector of values of characteristics that describes *i*-th emergency situation on the railway;

y_{j1}, y_{j2}, …, y_{jm} - vector of values of characteristics that describes *j*-th emergency situation on the railway.

Peak distance assumes independence between random variables, suggesting the orthogonal distance in space, but in practical applications, these variables are not independent.

Formula peak distance is of the form (7) [8]:

$$a_{\mathrm{L}}(Y_i,Y_j)=\frac{1}{m}\cdot\left(\frac{|y_{i1}-y_{j1}|}{y_{i1}+y_{j1}}+\frac{|y_{i2}-y_{j2}|}{y_{i2}+y_{j2}}+...+\frac{|y_{im}-y_{jm}|}{y_{im}+y_{jm}}\right) \quad (7)$$

where $a_L(Y_i,Y_j)$ - queen of the distance between two emergencies on the railway Y_i and Y_j;

y_{i1}, y_{i2}, …, y_{im} - vector of values of characteristics that describes *i*-th emergency situation on the railway;

y_{j1}, y_{j2}, …, y_{jm} - vector of values of characteristics that describes *j*-th emergency situation on the railway.

So, from analyzed metrics appropriate for cluster analysis emergencies on the railway for the purpose of identification is the weighted Euclidean distance because it takes into account the "importance" of each characteristic of such a situation that increases the reliability of the results of cluster analysis.

Normalization parameters in cluster analysis emergencies in rail transport

In cluster analysis to identify emergencies on the railway used parameters with different units of measurement, and this means that you need to bring them to a standardized form (normalized), especially when using such an extent close as Euclidean distance. This decision should be made taking into account the characteristics of the problem being solved.

Rationing is the introduction of a new conventional units, allowing a formal comparison of emergencies on the

railway. The main methods of valuation parameters (8 - 12) [5, 16]:

(8) $$q = \frac{(y - \bar{y})}{\sigma},$$

where q - normalized value y;
y - the setting for emergencies on the railway;
$\bar{y}$ - mean value of the parameter y;
σ - average deviation.

(9) $$q = \frac{y}{\bar{y}},$$

(10) $$q = \frac{y}{y'},$$

where y' - standard (reference) parameter y emergencies on the railway.

(11) $$q = \frac{y}{y_{\max}},$$

where $y_{\max}$ - the largest value of parameter y emergencies on the railway.

(12) $$q = \frac{(y - \bar{y})}{(y_{\max} - y_{\min})},$$

where $y_{\min}$ - the smallest value of the parameter y.

Development of the modified algorithm k-means clustering emergencies in railways

Basic concepts in the framework of this algorithm [6, 7]:

- training sample $M = \{m_j\}_{j=1}^{d}$, where d - power training sample;
- metric distance;
- the set of centers of clusters $C = \{c^{(i)}\}_{i=1}^{c}$, where

$$c^{(i)} = \frac{\sum_{j=1}^{d} u_{ij} \cdot m_j}{\sum_{j=1}^{d} u_{ij}},$$

- matrix decomposition $U = \{u_{ij}\}$, where

$$u_{ij}^{(l)} = \begin{cases} 1 \; at \; d(m_j, c_i^{(l)}) = \min\limits_{1 \le k \le c} d(m_j, c_k^{(l)}), \\ 0 \; in \; other \; cases. \end{cases}.$$

- the objective function (13). In the standard method does not take into account the particular subject area. In the analysis to identify the emergency of the railways belonging to one of nine classes [2] - it is appropriate to use the standard method. But to use the method for the analysis of multiple emergencies on the railway belonging to different classes - you need to enter in the objective function coefficient η will significantly distinguish different classes of emergencies. The coefficient depends on the class of emergency and must be determined by an expert. For these classes it has a smaller difference (eg, explosives and compressed, liquefied and dissolved gases under pressure) than for fundamentally different (eg, flammable solids, hypergolic substances and substances that emit flammable gases in contact with water and radioactive substances).

(13) $$J(M, U, C) = \sum_{i=1}^{c} \sum_{j=1}^{d} \eta u_{ij} d(m_j, c^{(l)}),$$

- A set of constraints (14):

(14) $$u_{ij} \in \{0,1\}; \; \sum_{i=1}^{c} u_{ij} = 1; \; 0 < \sum_{i=1}^{d} u_{ij} < d,$$

which specifies that each vector data can belong to only one cluster and owned by others. Each cluster contains at least one point, but less than the total number of points.

Scheme modified k-means algorithm for analyzing emergency on the railway to their identification is shown in Fig. 1

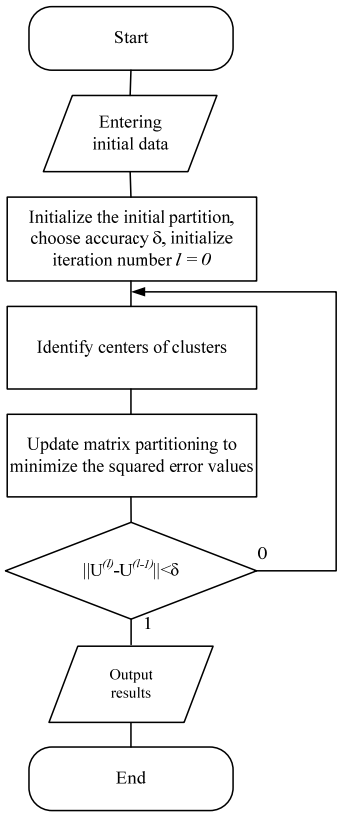

Fig. 1. Diagram of the modified k-means algorithm for analyzing the emergency situations on the railway transport with the aim of their identification

Formalizing the process of evaluating the results of the identification of emergency situations on railway transport

Evaluation of clustering results emergencies in rail transport can be achieved through the use of criterion variables.

There is a powerful set of criteria that can be used to analyze the emergency of the railways, among them was selected proportion of the total variation, point-biserial correlation method and the generalized variance in the classes that will be sufficient to assess the quality partitioning results [5].

Let the set of emergency on the railway is divided into k clusters G_i $(i = \overline{1,k})$.

To determine the value of T portion of the total variation between clusters emergency) to enter the following three

characteristics of the degree of dispersion of Emergency matrix *Y*, where the stored data about emergencies presented as points in a multidimensional space [3, 8, 9]:

- Total scattering *S* shall be defined (15)

$$S = \sum_{i=1}^{n} a^2(Y_i, \overline{Y}), \tag{15}$$

where Y_i - vector data about *i*-th emergency;

$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ - a common center of gravity;

n - number of emergencies that are analyzed;

$a^2(Y_i, \overline{Y})$ - square of the distance between the *i*-th emergencies and common center of gravity.

- Variation in linkage clustering *B*(16)

$$B = \sum_{z=1}^{k} n_z a^2(\overline{Y_z}, \overline{Y}), \tag{16}$$

where $\overline{Y_z} = \frac{1}{n_z}\sum_{Y_i \in G_z} Y_i$ - the center of gravity z-th cluster emergencies on railways;

n_z - the number of emergencies in rail transport in the cluster G_z;

$a^2(\overline{Y_z}, \overline{Y})$ - square of the distance between the centroid of the *z*-th cluster and the overall center of gravity.

- Variations within clusters emergencies in rail transport (17)

$$W = \sum_{z=1}^{k} W_z, \tag{17}$$

where $W_z = \sum_{Y_i \in G_z} a^2(Y_i, \overline{Y})$

Since the cluster analysis emergencies using Euclidean distance, it really is equality (18)

$$S = W + B. \tag{18}$$

The value share of the total variation T emergencies in rail transport is calculated by the formula (19) [4, 5]

$$T = 1 - \frac{W}{S}. \tag{19}$$

The value of *T* ranges from 0 to 1 $(0 \le T \le 1)$, if its value is close to 0 - this indicates lower quality breakdown emergencies in rail transport on clasters, and if the value is close to 1 - on the contrary.

Point-biserial correlation coefficient between R_b emergencies analyzed is defined as follows. Each pair of emergencies on the railway Y_i and Y_j is associated with two values - the distance between them and index equivalence δ_{ij} (20) [4, 5]

$$\delta_{ij} = \begin{cases} 1, \text{if } \mathrm{Y_i} \textit{ and } \mathrm{Y_j} \textit{ belong to the same cluster}; \\ 0, \textit{otherwise}. \end{cases} \tag{20}$$

R_b is calculated as the correlation coefficient between a_{ij} and binary and value δ_{ij} for all pairs of emergencies, which are analyzed, which gives (21) [8, 9]

$$R_b = \frac{(\overline{a_b} - \overline{a_w})\sqrt{\frac{f_w f_b}{n_a^2}}}{s_a}, \tag{21}$$

where $\overline{a_b}$ - the average distance between emergencies from different clusters;

$\overline{a_w}$ - the average distance between the emergency of a cluster;

f_w - the number of distances between emergencies, trapped in the singular cluster;

f_b - the number of distances between emergencies from different clusters;

n_a - total number of distances;

s_a - standard deviation of the distances.

Generalized variance in grades emergencies in rail transport *H* is one of the characteristics of the degree of dispersion of emergencies on the railway, which belong to the same class across its center. The size is calculated by the formula (22) [8, 9].

$$H = \det(\sum_{l=1}^{k} n_l W_l), \tag{22}$$

where $\det(\sum_{l=1}^{k} n_l W_l)$ - the determinant of the matrix, and the elements $w_{qm}(l)$ of the sample covariance matrix W_l calculated by the formula (23)

$$w_{qp}(l) = \frac{1}{n_l}\sum_{Y_i \in G_l}(y_i^{(q)} - \overline{y}^{(q)}(l))(y_i^{(p)} - \overline{y}^{(p)}(l)), \quad q, p = 1,2,...,m, \tag{23}$$

where $y_i^{(p)}$ - *p*-and characterization of an emergency on the railway Y_i;

$\overline{y}^{(p)}(l)$ - the average *p*-th component, calculated in emergencies of *l*-th class.

Relative quality division multiple emergencies at taxon is calculated by the formula (24)

$$K = \frac{\overline{T'} + \overline{R_b'} + \overline{H'}}{3}, \tag{24}$$

where T', R_b' and H' - the relative values of T, R_b and H.

Table 2. Evaluation of the results of cluster analysis of the emergency situations on the in railroad transport

Remedy	Clustering algorithm	Number of clusters			Share of total variation, T			Precision-biserial correlation coefficient, Rb			Generalized variance in classroom, H			Relative quality score partition, %
Number of experiment		1	2	3	1	2	3	1	2	3	1	2	3	
DEDUCTOR software tool for analyzing emergency	k-means	3	3	4	0,8005	0,8018	0,8399	0,586	0,516	0,469	0,721	0,744	0,782	75,4
PAES	Modified k-means	3	3	3	0,8199	0,8233	0,8475	0,563	0,512	0,482	0,672	0,689	0,715	78,31

Thus, the process of evaluating clustering results emergencies in rail transport will include the following steps (Figure 2):

1) simulation of cluster analysis by using the software DEDUCTOR and using PAES;
2) calculate the proportion of the total variation *T* emergencies between clusters;
3) calculate the point-biserial correlation coefficient between R_b emergencies analyzed;
4) calculating generalized variance in grades emergencies in rail transport *H*;
5) calculating the relative quality indicator set partitioning emergencies on clusters.

Evaluation of the results of the identification of emergency situations on railway

If you use the software to analyze DEDUCTOR emergencies and PAES, three experiments were conducted:

- at first - was elected five important parameters and characteristics of emergency, which fully reflects the state of such situations;
- the second - but these five were added one more - or less meaningful;
- the third - but significant added another 5 not important characteristics and parameters for identification of emergency.

Proportion of the total variation *T* (Table 2) determined from the results of modeling of cluster analysis emergencies showed that the quality of the tile, which is made PAES is higher because $T_{\text{PAES}} > T_{Deductor}$ (Figure 3).

Precision-biserial correlation coefficient R_b (Table 2) determined from the results of modeling of cluster analysis emergencies showed that the quality of the splitting up, which made by the software tool for analyzing DEDUCTOR for the analysis of the emergency situations is higher since $R_{bDeductor} > R_{b\text{PAES}}$ but the difference between the obtained values are low and under certain conditions we can assume that, $R_{bDeductor} \geq R_{bPAES}$ or $R_{bDeductor} = R_{b\text{PAES}}$, therefore, we can conclude that the quality of the partition in both cases (by this criterion) is the same (Figure 3).

Generalized variance in the classes *H* (Table 2) determined from the results of modeling of cluster analysis emergencies showed that the quality of the tile, which made software tool for analyzing DEDUCTOR emergency is lower, since $H_{Deductor} > H_{\text{PAES}}$ that indicates a deviation from the center of clusters emergencies when splitting DEDUCTOR software tool for analyzing emergency (Fig. 3).

Relative Quality division multiple emergencies at clasters (Table 2) determined from the results of modeling of cluster analysis emergencies showed that the quality of the tile, which made software tool for analyzing DEDUCTOR emergency is lower by 2.91%, indicating a lower quality division multiple emergencies at clusters DEDUCTOR software tool for analyzing emergency (Fig. 3).

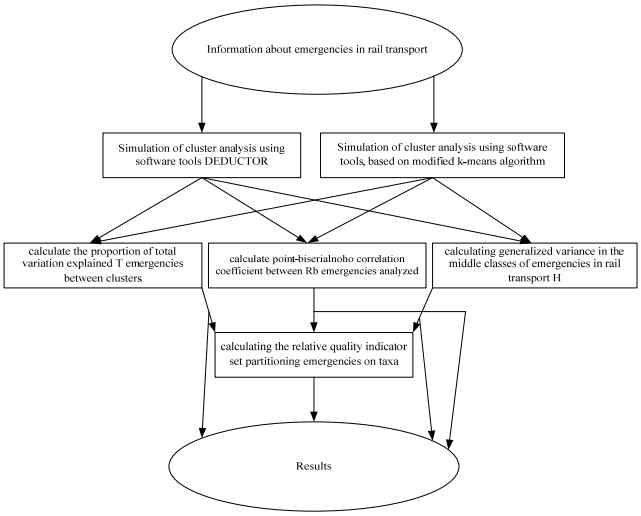

Fig. 2. The stages of the evaluation process of clustering results of the emergency situations on the in railroad transport

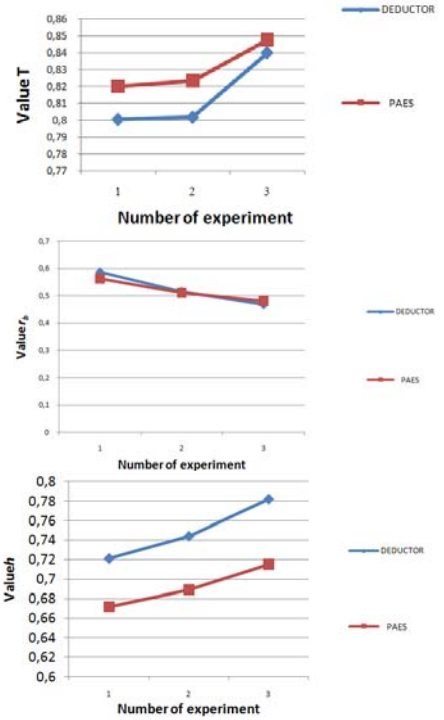

Fig. 3. Comparison of simulation of cluster analysis emergencies on the railroad transport

Conclusion

The conducted research showed that the improvement of the classical k-means algorithm by introducing the new

parameters to the target function caused better quality of railway transport emergency clusterization by 2.9%, and thus, a modified algorithm is useful for analysis of emergency on the railroad transport.

REFERENCES

[1] Youkhimchukl S.V., Kazman M.D., Models for the automation in making recommendations to the commander of fire extinguishers on the railroad transport: monograph, *UNIVERSUM*, Vinnytsia, 2008

[2] Savchuk T.O., Petrishyn S.I., Determination of the Euclidean distance between the emergency situations on the railroad transport during claster analisis, *Naukovi Pratsi Vinnytskogo Nationalnogo Technichnogo Universitety*, 3 (2010), http://www.nbuv.gov.ua/e-journals/vntu/2010_3/2010

[3] Savchuk T.O., Petrishyn S.I., Comparative analysis of using clastering methods for the identification of the emergency situations on the railroad transport, *Naukovi Pratsi Donetskogo Nationalnogo Technichnogo Universytety,* 11(2010), 135-140

[4] Savchuk T.O., Petrishyn S.I., Distance and the degree of proximity as the basic characteristics of intellectual analysis of the emergency situations on the railroad transport, *Conference proceedings, «INTERNET-EDUCATION-SCIENCE-2010», 7-th international conference IOH-2010*, Vinnytsia, 2010, 258-261

[5] Savchuk T.O., Petrishyn S.I., Normalization of the parameters' values during the cluster analysis of the emergency situations on the railroad transport, *Conference proceedings, International conference «Information computer technologies, simulation, control»*

[6] Barsegian A.A., Kuprianov M.S., Stepanenko V.V., Holod I.I., Methods and models for data analysis: OLAP and Data Mining, *BHV-Peterburg*, 2004

[7] Savchuk T.O., Petrishyn S.I., Peculiarities in selecting cluster parameters during the analysis of emergency situations on the railroad transport, *Measuring and calculating equipment in technological processes*, 2 (2010), 144-149

[8] Aivazian S.A., Buhstaber V.M., Eniukov I.S., Applied statistics: Classification and decrease in dimensionality, *Finances and statistics*, Moskva, 1989

[9] Mandel I.D., Cluster analysis, *Finances and statistics*, Moskov, 1988

[10] Savchuk T.O., Petrishyn S.I., Comparative analysis of using clustering methods for the identification of the emergency situations on the railroad transport, *Conference proceedings, System analysis and information technologies SAIT2010*, 2010, 485

[11] Methods for structure analysis. Access mode: www.sati.archaeology.nsc.rustatmethods_info.php

[12] Petrishyn S.I., Cluster analysis of emergency situations on the railroad transport using distances and degrees of of proximity between such situations, *Proceedings of the XL scientific and technical conference*, 2011

[13] Duran B., Odel P., Cluster analysis, *Statistika,* Moskov, 1977

[14] Holand S.M., Cluster Analysis, *Department of Geology, University of Georgia*, GA 30602-2501, 2006

[15] Nevin L., Zhang Hierarchical Latent Class Models for Cluster Analysis, *Journal of Machine Learning Research*, 5 (2004), 697-723

[16] Kormen T., Leiserson Ch., Rivest R., Stein K., Algorithms: building and analysis, 2-d edition, Printing house «Williams», 2009

***Authors**: Ph. D. Tamara Savchuk, Vinnitsa National Technical University, E-mail: savchtam@rambler.ru; M. Sc. Sergiy Petrishyn, Vinnitsa National Technical University, E-mail: petrishyn@gmail.com; M. Sc. Laura Sugurova, Kazakh National Technical University, E-mail: sla-taraz@mail.ru; Ph.D. Andrzej Smolarz, Lublin University of technology, E-mail: a.smolarz@pollub.pl*