https://doi.org/10.33271/nvngu/2023-3/137

Yu. I. Klius[*1],
orcid.org/0000-0002-1841-2578,
M. A. Prodanchuk[2],
orcid.org/0000-0003-3504-4583,
O. O. Gololobova[3],
orcid.org/0000-0003-1857-8196,
L. V. Halan[4],
orcid.org/0000-0002-4118-9255,
M. O. Varhatiuk[5],
orcid.org/0000-0003-2357-1597

1 – Volodymyr Dahl East Ukrainian National University, Kyiv, Ukraine
2 – National Scientific Centre "Institute of Agrarian Economics", Kyiv, Ukraine
3 – Ukrainian State University of Science and Technology, Dnipro, Ukraine
4 – State University of Intelligent Technologies and Telecommunications, Odesa, Ukraine
5 – Odesa State Agrarian University, Odesa, Ukraine
[*] Corresponding author e-mail: klius@snu.edu.ua

EVALUATION AND CONTROL OF DATA RELEVANCE IN INFORMATION SYSTEMS OF THE TRANSPORT INDUSTRY MANAGEMENT

Purpose. Analysis of data coming to the information systems (IS) of the transport area (TA) and development of proposals for verification of their relevance to ensure reliability and efficiency of management. Development of the evaluation and control method of data relevance in information systems of the TA.

Methodology. Special and general methods of scientific knowledge are used: critical analysis to establish scenarios of the impact of data relevance on the level of effective interaction of the elements of the system and its management; content analysis for stratification of requirements for data from IS and management decisions at the stage of analysis; mathematical formalization for the development of the data evaluation method and control of their relevance, which is proposed as the mathematical base of the algorithm; method of sequential selection for automatic detection of functional dependence, which describes data with a smaller value of standard error.

Findings. Proposals to ensure data relevance have been developed. Data and management decision stratification has been proposed which will allow avoiding common errors due to non-normalization of data presentation. The scenarios of the impact of data relevance on the level of effective interaction of the elements of the transport system and its management have been established. For permanent data control, an algorithm for detecting deviations in consecutive data sets, automatic analysis of specified deviations, and detection of univariate and multivariate nonlinear patterns that cause deviations has been proposed. In case of inconsistencies of data with the identified non-linear patterns, the information is checked by the method of evolutionary programming. The effectiveness of the proposed methods is demonstrated using the example of analysis of the cargo turnover of seaports whose trends are characterized by significant changes over time and across ports. These trends may even have opposite characteristics.

Originality. The method of data evaluation and control of their relevance, which is based on the methods of evolutionary ("genetic") programming, is developed.

Practical value. Proposals for evaluation and control of data relevance in information systems of the TA will increase the efficiency of management activities.

Keywords: *transport area, management, information systems, data relevance control*

Introduction. The reliability and relevance of information systems (IS) data in the transport industry today not only determine the efficiency of IS work, but also, to a large extent, affect the efficiency of the industry as a whole, since transport is a structure with a high level of technological connectivity. The importance of ensuring the reliability and relevance of IS data increases significantly in the conditions of rapid growth in the amount of data used by transport IS, and the growth of digitization rates in the transport sector and other areas of the economy.

Since both the transport industry and each mode of transport separately represent complex systems, it is natural to use systems theory approaches for their analysis. From this point of view, errors due to the irrelevance of data in one of the segments of these systems have a significant impact on the functioning of the transport system as a whole due to the emergency nature of the specified structures.

The weight of this influence is significantly increasing today, because in the period of acute crisis and military operations, which generate a significant level of uncertainty in the functioning of transport systems, each subsequent reason for a failure in their work may serve as a trigger for bringing these systems out of a shaky state of temporary equilibrium. The reason for the irrelevance of the data received by the IS may be due to: an error ("human factor"), technical or technological reasons, a purposeful malicious action, a computer virus, etc.

Nowadays the situation in Ukraine is significantly complicated by the fact that the transport sector is served by information systems of different age and performance, which form a complex eclectic conglomerate, the functioning of which increases the level of unstructuredness of the general array of data. And this is a factor that itself can cause system errors.

Literature review. Many scientific works are devoted to the study on the reliability and relevance of data as a factor in ensuring the effective operation of information systems and the structures they serve. The volume of scientific literature on this issue is so large that separate scientific reviews are dedicated to its research, in particular, Iriarte, et al. [1] and Morcov, et al. [2]. Unfortunately, most of the authors of these reviews relied on expert and statistical methods of information relevance, a significant level of effectiveness of which is achieved only when processing statistical data and is unsatisfactory, for example, when processing other types of data.

In contrast to the mentioned works, Javed, et al. [3] studied Cooperative Intelligent Transport Systems (C-ITS), where different types of data are used. For data analysis and control, Javed, et al. [3] proposed simulation models, which is insufficiently reliable. Ellingsen, et al. [4] rely on the method of analogies when analyzing data, which significantly increases the level of subjectivity of the analysis. Moros-Daza, et al. [5] used a comparison of samples from the Sicex database, which also does not properly increase the relevance of the data, since the differences between the data of the samples may be caused by parameters of influence on one of the samples that have not been taken into account. In their work, Díaz-Gutiérrez and

Núñez-Rivas [6] used standard methods of BIS – operational data analysis, which may also not take into account impact parameters. Di Vaio, et al. [7] and Brunila, et al. [8] indicate the importance of normalizing the provision of data from port operations in inter-organizational relations, because this can lead to disruptions in work, which is taken into account in the presented study. Zerbino, et al. [9] propose a "mirror method" which ensures that a number of consecutive data are returned to the point of failure when an error is detected and all subsequent transactions are repeated for correction. Unfortunately, this method works only after an error is detected. Seo, et al. [10], to control and assess the reliability of data, offers a data mining software package – PolyAnalyst, which, unfortunately, requires significant computing resources because of extremely large volumes of data. Yueshuai, et al. [11] describes the application of synthesis of sets of alternative tours with diffusing probability for query-time sampling. But this algorithm is primarily focused on data protection rather than data control. Davidich, et al. [12] analyzes the modeling of traffic and passenger flows in urban transport systems for an intelligent decision support system. The relevance analysis here relies on the comparison of data flows, which is not always possible for other transport systems. Greis, et al. [13] analyzes the Digital Twin Framework for integrated transportation and logistics processes with the support of machine learning, which is also a specific method. Kim, et al. [14] studied the spatio-temporal density method. But this method is adapted for use on a large Automatic Identification System (AIS). Zhang, et al. [15] use a set of Data mining algorithms. Montoya-Torres, et al. [16] use Big Data Analytics methods for Intelligent Transport Systems (ITS). The same approach is implemented by Dong, et al. [17], which indicates its relevance. Unfortunately, Big Data Analytics methods require significant amounts of calculations. Bazaluk, et al. [18] proposed an original entropy method for optimizing multimodal transportation. Unfortunately, the use of this method to assess the relevance of data is not given enough attention in the article by Bazaluk, et al. [18]. The method of taking into account the simultaneous use of deterministic, stochastic and fuzzy values, proposed by Bazaluk, et al. [18], is used in the presented study.

Unsolved aspects of the problem. The given review of literary sources shows that in order to ensure the reliability and efficiency of the management of the transport industry, there is a need to develop a method for evaluating and controlling the relevance of data in information systems, especially in conditions of an increased level of threats.

The purpose of the article. Conducting an analysis of data entering the information systems of the transport industry and developing proposals for checking their relevance to ensure the reliability and efficiency of management.

Developing a method of evaluating and controlling the relevance of data in information systems of the transport industry.

Methods. By using the method of critical analysis, it was established that the introduction of IS into the management process of the transport industry determines new, stricter conditions for the presentation of operational and reporting information that comes to the elements of the system that belong to the number of management objects from the elements of the system – management subjects. The unreliability of the data on the basis of which the information system offers managers certain decision options is one of the main reasons that can lead to erroneous management decisions. The data according to which the decision is made are, first of all, the reporting information provided by the management of the controlled elements of the system to the control elements.

Since transport structures represent systems with a much higher level of connectivity than the structures of other industries, and the transport industry ensures the reliable and uninterrupted operation of other sectors of the economy, the requirements for the reliability and relevance of data entering the IS of the transport industry must be much stricter.

Using the method of content analysis, the requirements for the data entering the IS are stratified, and management decisions are stratified at the analysis stage. It is established that the reported information must be: timely, minimized (that is, not redundant), sufficient, have an appropriate level of data aggregation, relevant, reliable, comply with the stipulated data formats.

The groups of variables that are required for the implementation of the decision-making support process in the transport industry are: parameters that influence the choice of a management decision option and the rate of their change; decision-making constraints.

Restrictions on making each of the decisions can be permanent or temporary in nature; may be unique or typical; may be formulated by the analyst before the stage of setting the problem of choosing a solution option, or can be implemented automatically – using the appropriate algorithms and knowledge base.

Using the method of critical analysis, it was established that the level of relevance of the specified current information at significant time intervals affects the level of effective interaction of system elements and its management in two scenarios: the "domino effect" – as a result of one wrong decision or systematic deviation of a managed element (or subsystem) from the optimal state (or the path to it) as a result of the sequential action of a number of incorrect decisions made on the basis of irrelevant information. By the "domino effect" we mean a sequence of states of the system or its element, independent of the will of the manager or executive, which leads to significant losses or even a disaster.

Therefore, the current reporting data provided by the controlled elements of the system must pass the stage of automatic analysis and verification at the entrance to the control element. For this, a procedure for operational analytical data processing is proposed whose tasks are not identical to the classical approaches of OLAP data analysis [16, 17].

First of all, data control must be implemented, namely:

- integrity check and detection of missing values;
- finding relationships between data;
- checking the continuity of the specified relationships in reports for previous time periods;
- data analysis of previous and current reports with the finding of deviations in the series of consecutive data, greater than the permissible standard error and which are not identical to the level of deviations in reports for previous periods of time.

The next step is the automatic analysis of the indicated deviations and the finding of univariate and multivariate nonlinear patterns that cause the deviations.

For this purpose, it is proposed to use an algorithm that allows one:

- to determine in the series of consecutive data and their groups that can be described by the automatically selected functional dependence;
- to evaluate the compliance of the specified dependence by the size of the standard error;
- to identify the parameters that are most significant for the detected dependence and find those values in the series of consecutive data that have deviations greater than the specified interval values.

Using the method of mathematical formalization, a method of data evaluation and control of their relevance has been developed, which is proposed as the mathematical base of the mentioned algorithm.

This method is as follows. There is a set of m dependent variables that describe all aspects of the transport process and n independent variables, including those that, according to expert assessments, could previously be considered insignificant. Dependent variables are further considered as functions. Then the entire transport process can be considered as a set of functional dependencies that form the response area W in the $(n + m)$ – dimensional space.

The method of sequential selection automatically reveals a functional dependence that describes the data with a smaller value of the standard error and with a smaller number of points with a deviation greater than that of the parameter of the general population.

To find the specified functional regularities, which do not have to be specified by the user, a data mining software package adapted for use in the transport management system, which is based on the methods of evolutionary ("genetic") programming, can be used.

An important requirement for using this data mining software package is that it must simultaneously process different types of data: real, stochastic, fuzzy parameters, linguistic variables, etc. The implementation of this requirement provides an opportunity to form universal regularities based on data types. For this, the technique proposed by Bazaluk, et al. [18] is used.

Using evolutionary programming, the following approach is proposed.

From the set m, a group of functionals, the so-called "terms", is selected. Let us consider this process using the example of two terms ($t_{(1)}$, $t_{(2)}$)

$$t_{(1)} = f(x_i, a); \quad t_{(2)} = f(x_j, a), \quad x_i, x_j \in n, m,$$

where x_i, x_j are variables; a is the variable parameter of the evolutionary algorithm; i, j are indices; $i = 1, 2, 3, \ldots, n$; $j = 1, 2, 3, \ldots, m$.

That is, an allele can be represented by the values of constants, independent and, in general, dependent variables. Next, genetic methods are used: substitution, finding out the level of adaptability and selection. Substitution is implemented from sets of n, m, and constants and parameters expertly added to each of the input terms.

Next, the degree of similarity of these terms is determined. For this, the Levenshtein distance is used. The Levenstein distance is used because it is suitable for analyzing different numbers of unifiers (as it is not possible to specify a priori how many variables the function will depend on). If u is a string of length $1 - i$, and v is a string of length $1 - j$ then, as it is known, the Levenstein distance d_{lev} can be calculated recursively

$$d_{lev}(u\{1 \ldots i\}, v\{1 \ldots j\}) \rightarrow \min\{d(u\{1 \ldots i\}, v\{1 \ldots j-1\} +$$

$$+1) d(u\{1 \ldots i-1\}, v\{1 \ldots j\} + 1) d(\{u\{1 \ldots i-1\}, v\{1 \ldots j-1\})\} + \delta,$$

provided $d_{lev} \in (0 \ldots \vartheta)$,
where $d = \{0 \quad \text{if } u(i) = v(j) 1 \quad \text{if } u(i) \neq v(j)\}$.

The value ϑ is set by an expert.

If the distance $d \gg d_{lev}$, the relevance of the data is questionable. Data determined as erroneous are not taken into account during further processing in information systems for support and management decision-making.

Results. Reporting information in Ukraine is usually presented in a relational form, but for the future development and universalization of IS, it is proposed to foresee the possibility of using the form of a multidimensional cube (the so-called "hypercube"). The formation of a hypercube will provide the necessary level of heterogeneous data stratification and data clustering. This will eliminate data presentation errors and will further facilitate the performance of analysis tasks and finding implicit regularities, in particular in the behavior of elements of the transport management system.

Management decisions at the analysis stage should also be stratified, which will facilitate the determination of requirements for the normalization of data presentation, simplify the choice of decision options, reduce the volume of calculations and, in general, and reduce the time of IS work.

Stratification of management decisions should be carried out according to the type of decision, namely:

- for structured variants of events determined by deterministic problem parameters;
- for structured variants of events in the presence of an insignificant level of uncertainty;
- for poorly structured variants of events with a significant level of uncertainty;
- for unstructured variants of events with unclear problem parameters;
- for crisis situations, which are characterized by unclear parameters of the task, a significant level of uncertainty of consequences, and time limitations for decision-making.

Since there can be many elements in the transport management system, it is natural to move from a database distributed by element or subsystem, for example, by transport hubs, sea or river ports, by regions, etc., to a hierarchical corporate data warehouse and knowledge base of the type of transport or industry as a whole.

The creation of a corporate data repository and knowledge base should be accompanied by the allocation of access to system administrators from individual elements or subsystems only to certain areas of this repository and, if necessary, clarification by the security administrator of the specified access rights.

This is facilitated by the introduction of a project approach, the effectiveness of which has been proven by practice. With the implementation of the specified approach, the access of managers of different levels to certain data in information bases is automatically substantiated. This, in turn, requires a flexible response of system administrators to the provision of such access, or an appropriate response of management to the formation of a group of managers to implement the project and, if necessary, issuing an order to extend access rights to individual project executors.

IS forms solutions for each project. The choice of a decision option before implementation can be carried out in an expert manner (according to the decision of the management), according to the appropriate level of economic rationality of the interaction of the elements of the transport management system; according to the level of effectiveness, under the condition of minimization of consequences, etc. But a set of alternatives based on a set of parameters often makes it difficult to choose a solution. Therefore, it is recommended to introduce a block of analysis of metric binary relations for decision options and their pairwise comparison. In this case, the principle of invariance of selection with the change in the measurement scale should be applied.

Since the transport management system is multi-element with a possible cross-influence of elements and a significant number of parameters that can affect the level of economic rationality of the results of management actions, this may in some cases cause the need to replace the "tree of events" method with the "forest of decisions" method. When using the "forest of decisions" method, a group of "trees of events" should be formed for variants of the impact parameters. Decision options are submitted for consideration by the management according to the decreasing probability of the values of the specified parameters.

The found patterns of data series are stored in the information base and, if necessary, are sent to the analyst to confirm their reliability. If it is impossible to substantiate the cause of the named deviations, the identified series of data must be sent to the executor for additional verification, followed by an explanation of the deviation or its correction. The stored patterns of the data series form a knowledge base, with the help of which forecasts are formed and, in the event of failure to provide operative and reporting information within the specified time due to objective reasons, a predictive version of this report can be automatically generated. The forecast version of the report, in particular, is needed for the formation of management decision options.

In addition, it should be noted that the storage of primary, unprocessed – the so-called "atomic" data has a number of disadvantages. In particular, the storage of "atomic" data requires a significant increase in costs, for example, the need to correct them due to the constant detection of technical, human and other errors and causes an increase in the delay time when accessing the specified data. Therefore, it is worth introducing a subsystem of data processing, verification and normalization.

It was mentioned above that the transport industry is a complex system. As a result, there may be groups of factors whose impact on achieving the tasks of rational transport management is insufficiently studied. The proposed algorithm allows determining the presence of previously unknown influencing factors with a high level of relevance.

An unpredictable consequence of the interaction of the entire set of factors due to the appearance of new, previously undefined, factors can be a weakening of the effectiveness or, even, the reverse effect of the management action or, on the contrary, its strengthening in an unexpected form and size. This can lead to significant destabilization of the control element, their group, or even the entire system.

This makes it impossible to use methods where the set of influencing factors is determined a priori. An example of such methods is Moros-Daza, et al. [5], Díaz-Gutiérrez and Núñez-Rivas [6], factor analysis, which is a traditional tool for determining the significance of the influence of one or another factor.

However, since the amount of data that forms the response area W in the $(n + m)$ – dimensional space can be large, the permanent application of the specified algorithm leads to a significant increase in the level of use of IS resources, which, in turn, slows down or even makes work of the system impossible. Therefore, an alternative option for detecting previously unidentified factors in the formation of standard functional dependencies of influence factors is proposed. The specified method is rather qualitative in nature and can be used for preliminary data analysis for the economical use of computing resources.

This option involves comparing the scalable objective function of the system and the objective function for the controlled element or subsystem under analysis. On a larger scale, for example, at the level of analysis of transport data of the entire country, factors inherent to each individual type of transport or its substructure become insignificant. But when comparing homogeneous functions on the same scale, it is possible to immediately detect the difference in trends at separate time intervals.

This is proven by the study on sea transport parameters presented in Figs. 1 and 2, which is given as an example of the implementation of the proposed approach. The study on sea transportation parameters, whose results are presented in Fig. 1, was carried out using statistical data of the State Statistics Service [19], and the study on the parameters of sea transportation, whose results are presented in Fig. 2, using statistical data of the Administration of Sea Ports of Ukraine [20].

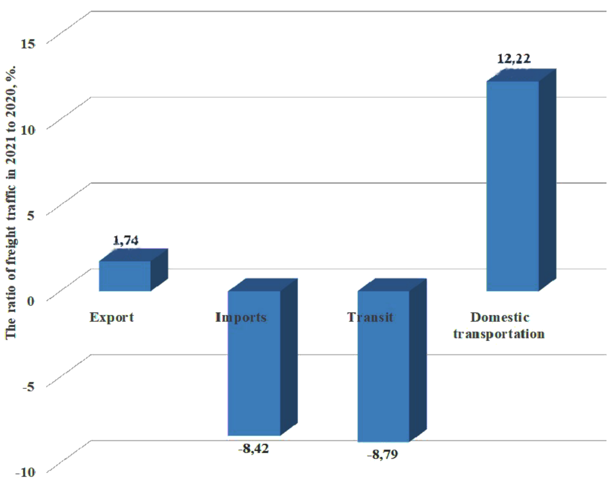

Fig. 1. The number of changes in the cargo turnover of sea transport from 2021 compared to 2020 in accordance with the types of transportation, %

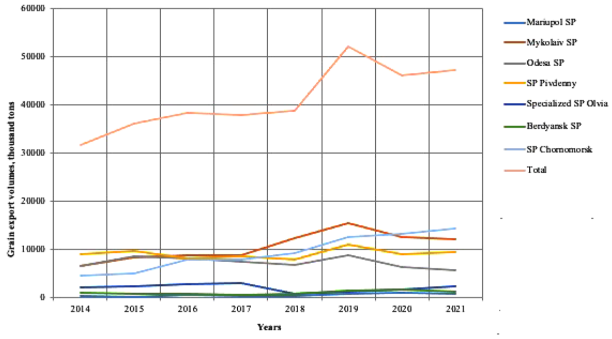

Fig. 2. Volumes of transshipment of grain cargoes through Ukrainian seaports for the period 2014–2021, million tons

The dynamics of freight turnover by types of transportation of maritime transport enterprises (Fig. 1) in 2021 compared to 2020 indicates significant fluctuations of this indicator even in a relatively short period.

The above said, accordingly, causes the need to analyze the data for relevance. The preliminary analysis shows that the values of cargo turnover for all ports and all types of cargo (Fig. 1) hide the influence of individual parameters. Let us detail the task, highlighting, for example, the dynamics of grain cargoes by maritime transport enterprises of Ukraine.

The analysis of data series of grain cargo turnover (Fig. 2) indicates signs of significant differences in data series for individual ports, even changes in grain cargo volume trends for them at significant time intervals.

At the same time, with significant changes in the studied parameters, there is a coincidence of breaks in the functional dependencies of these parameters for different seaports in different regions of the country (Fig. 2).

In particular, the analysis shows the similarity of the analytical dependencies of export dynamics for the country's sea transport and the sea ports (SP) of Mykolaiv, Pivdennyi, Odesa, Chornomorsk. A characteristic feature, in particular, is the coincidence of the peak times in 2019.

The indicated peaks of the analytical dependencies of the dynamics are caused by the high harvest of that year, which, accordingly, led to a jump in export deliveries. For Berdiansk seaport and Mykolaiv seaport, this coincidence is not so obvious due to the difference in scale – export deliveries through these ports were smaller than through the leading ports.

At the same time, a characteristic feature of these graphs is not only the coincidence of extreme values of export deliveries, but also a comparison of existing trends of changes in the studied parameters with periods of less significant changes in the specified parameters.

According to the mathematized approach, the concept of "rend" will be interpreted as the slope ratio of the tangent to the middle of the interval segment of the graph of the functional dependence of the studied parameter in time. The points limiting the specified segment are the values of the parameter at the beginning and at the end of the specified time period. As it is known, the slope ratio of the tangent is equivalent to the first derivative of the function of the studied parameter in time $\frac{\partial\delta}{\partial\tau}$ (where τ is time).

The analysis of information entropy, which according to research by Bazaluk, et al. [18] correlates with the traditional objective functions of sea freight transportation – cost, time and reliability of cargo transportation, is more sensitive, but requires significant amounts of computing resources and, accordingly, computing time. Therefore, the study on the change of the first derivative of information entropy over time can be a more sensitive indicator of data relevance.

The expediency of using information entropy as an indicator of data relevance is enhanced by the possibility to analyze in this case not only data represented by real numbers, but also data represented as stochastic or even uncertain parameters [18].

The transition to the use of the first derivatives of functions allows using these derivatives as a convenient indicator of the relevance of the studied parameters during automated data analysis for relevance. The discrepancy will be indicated by a change in the sign of the first derivative function, which will indicate the absence of a smooth nature of the functional dependence.

For example, in the period of 2018–2019, despite the coincidence of the signs of the trends, in contrast to the following periods, in particular the data for the period of 2020–2021 given in the table, when the trends had different signs, according to the value of the first derivatives of the volume functions export of grain cargoes for the Mykolaiv seaport, Odesa seaport, "Pivdennyi" seaport, "Olvia" specialized seaport and, in general, for all seaports of Ukraine, the trends (first derivatives) differed significantly from each other, as it can be seen from the following tables.

The first derivatives of the functions of grain cargo exports for the periods of 2018–2019 and 2020–2021 are calculated according to the statistical data of the Administration of Sea Ports of Ukraine [20]. The calculated data are shown in the Table below.

The values of the first derivatives (trends) of grain export volumes for the period of 2018–2019 for the indicated ports were, respectively, 1.271996; 1.301837; 1.366982; 1.381446 and 1.33684. That is, for the "Pivdennyi" Seaport and the "Olvia" Specialized Seaport, the tilting angles (defined by the arctangents, the values of which are given in the Table) of the trend were bigger than the tilting angle of the trend for all seaports of Ukraine, and for Mykolaiv and Odesa seaports they were smaller. This indicates the different magnitude of the influence of factors on cargo turnover for each of the ports, in particular, the technological level of their equipment and the competitive position in the rating of Ukrainian seaports for grain cargoes in the specified time period.

In particular, the analysis shows the similarity of the analytical dependencies of export dynamics for the country's sea transport and the sea ports (SP) of Mykolaiv, Pivdennyi, Odesa, Chornomorsk. A characteristic feature, in particular, is the coincidence of the peak times in 2019.

The indicated peaks of the analytical dependencies of the dynamics are caused by the high harvest of that year, which, accordingly, led to a jump in export deliveries. For Berdiansk seaport and Mykolaiv seaport, this coincidence is not so obvious due to the difference in scale – export shipments through these ports were smaller than through the leading ports.

A more detailed analysis indicated that the difference in the signs of the trends in grain traffic for export in the period after 2019 for SP Chornomorsk and other ports (Fig. 2) is not a mistake, because in August 2018, the Administration of Sea Ports of Ukraine in cooperation with the China company Harbor Engineering Company Ltd. started dredging sea approach channels to ports and their water areas. This provided opportunities for the approach of large-tonnage vessels and significantly increased the throughput capacity of Ukraine's seaports. At the same time, an integrated approach to the analysis of changes in grain transportation through Ukrainian seaports, which, for this type of cargo, must take into account not only the world market demand for grain, but also the sudden changes in the volume of its production in Ukraine, which are associated with the level of yield for the studied time period.

Insignificant changes in the studied parameters, which can be seen, in particular, from the analysis of grain cargo exports for the period of 2020–2021, shown in the table, indicate that even the change in the signs of the first derivatives (trends) for the Mykolaiv and Odesa seaports indicate only a certain redistribution of the volume of traffic of homogeneous cargo between ports in less productive years and, in this case, is not a confirmation of the irrelevance of the data.

Among other Ukrainian ports, the largest changes in the volume of grain cargo transportation through the specialized sea port of Olvia in 2020–2021 (+37.1 %) confirm the stability of the trend identified for the period of 2018–2019 (+41.7 %) for this port.

The indicated results were further confirmed using the proposed method of genetic programming, and the following approach was proposed. Levenstein's distance, regardless of the fact of dredging on the approaches and in the port of Chornomorsk, indicated a significant discrepancy with its reference value. After taking into account the increase in the interval of the seaport capacity factor, the difference between d and d_lev was only 1.7 %, which is acceptable. The step-by-step implementation of the qualitative and quantitative approaches described above significantly reduced the amount of calculations and their time.

Conclusions. According to the analysis of the data received in the information systems of the transport industry, the data requirements are stratified. A procedure for operational analytical processing of data, which implements the developed algorithm, is proposed.

Proposals to avoid errors due to the human factor, technical or technological reasons, malicious action, computer virus, are developed. Attention is paid to errors due to non-normality of data presentation, which is especially important for heterogeneous data. Solutions to facilitate data clustering are proposed. This will not only eliminate data presentation errors, but will also facilitate the performance of analysis tasks and finding implicit regularities, in particular in the behavior of elements of the transport management system.

The need for stratification of management decisions at the analysis stage is indicated. This will facilitate the determination of requirements for the normalization of data presentation, simplify the choice of solution options, reduce the volume of calculations and, in general, reduce the time of operation of the information system and the amount of computing resources involved. It is proposed to stratify management deci-

Table

Characteristics of grain cargo export volume trends for the periods of 2018–2019 and 2020–2021

No.	The name of the port	The value of the first derivative of grain exports for the period 2018–2019	Arc tangent of the 2018–2019 trend, radians	Change in grain exports volume in 2018–2019, %	The value of the first derivative of grain exports for the period of 2020–2021	Arc tangent of the 2020–2021 trend, radians	Change in grain exports volume in 2020–2021, %
1	Mykolaiv Sea Port	1.2719	0.9045	+24.8	–0.2679	–0.2821	–2.9
2	Odesa Sea Port	1.3018	0.9157	+29	–1.3123	–0.9195	–10.8
3	The Sea Port "Pivdennyi"	1.3669	0.9392	+39.4	+0,7909	0.6692	+6.7
4	Specialized Sea Port of Olvia	1.3814	0.9442	+41.7	+1.3554	0.9352	+37.1
5	On all seaports of Ukraine	1.3368	0.9285	+34.4	+0.2592	0.2567	+2.5

sions according to the type of decision, taking into account the level of structuring of event options and the level of determinism and uncertainty of their parameters. With regard to military actions, the types of crisis situations, which are characterized by unclear parameters of the task, a significant level of uncertainty of consequences, and a time limit for decision-making, are separately highlighted.

Since there can be many elements in the transport management system, it is natural to move from a database distributed by element or subsystem, for example, by transport nodes, sea or river ports, by regions, etc., to a hierarchical corporate data warehouse and knowledge base of the type of transport or industry as a whole. It is proposed to implement IS by using a corporate data warehouse and knowledge base and introducing a project approach, the effectiveness of which has been proven by practice.

A method for data evaluation and control of their relevance which uses methods of evolutionary ("genetic") programming, has been developed. An important requirement for using this data mining software package is that it must simultaneously process different types of data: real, stochastic, fuzzy parameters, linguistic variables, etc. The implementation of this requirement provides an opportunity to form universal regularities based on data types.

An algorithm for detecting deviations in consecutive data sets and automatic analysis of the specified deviations and finding univariate and multifactorial nonlinear patterns that cause deviations is proposed for the possibility of carrying out constant analysis and reducing the time spent and the amount of IS resources.

If the two described methods do not reveal factors that lead to deviation, the data with a high level of probability are determined as erroneous. Data determined as erroneous are not taken into account during further processing in information systems for support and management decision-making. This increases the level of reliability and efficiency of the specified solutions.

The effectiveness of the proposed methods is demonstrated using the example of the analysis of cargo turnover of sea ports, which, as it was shown, is characterized by significant changes not only over time, but also for individual ports, whose trends may even have the opposite character.

Further research should be aimed at approbation of the developed methods for the information system of coordinating the interaction of different types of transport under significant traffic disruptions.

References.

1. Iriarte, C., & Bayona, S. (2020). IT projects success factors: A literature review. *International Journal of Information Systems and Project Management*, *8*(2), 49-78. https://doi.org/10.12821/ijispm080203.

2. Morcov, S., Pintelon, L., & Kusters, R. (2020). Definitions, characteristics and measures of IT project complexity – a systematic literature review. *International Journal of Information Systems and Project Management*, *8*(2), 5-21. https://doi.org/10.12821/ijispm080201.

3. Javed, M.A., Zeadally, S., & Hamida, E.B. (2018). Data Analytics for Cooperative Intelligent Transport Systems. *Vehicular Communications, 15*, 63-72. https://doi.org/10.1016/j.vehcom.2018.10.004.

4. Ellingsen, O., & Aasland, K.E. (2019). Digitalizing the maritime industry: A case study of technology acquisition and enabling advanced manufacturing technology. *Journal of Engineering and Technology Management*, *54*, 12-27. https://doi.org/10.1016/j.jengtecman.2019.06.001.

5. Moros-Daza, A., Solano, N., Amaya, R., & Paternina, C. (2018). A multivariate analysis for the creation of Port Community System Approaches. *Transportation Research Procedia, 30*, 127-136. https://doi.org/10.1016/j.trpro.2018.09.015.

6. Díaz-Gutiérrez, D., & Núñez-Rivas, L.R. (2022). Digitalizing Maritime Containers Shipping Companies: Impacts on Their Processes. *Applied Sciences*, *12*, 2532. https://doi.org/10.3390/app12052532.

7. Di Vaio, A., & Varriale, L. (2020). Digitalization in the sea-land supply chain: experiences from Italy in rethinking the port operations within interorganizational relationships. *Production Planning & Control, 31*, 220-232. https://doi.org/10.1080/09537287.2019.1631464.

8. Brunila, O.P., Kunnaala-Hyrkki, V., & Inkinen, T. (2021). Hindrances in port digitalization? Identifying problems in adoption and implementation. *European Transport Research Review*, *13*, 62. https://doi.org/10.1186/s12544-021-00523-0.

9. Zerbino, P., Aloini, D., Dulmin, R., & Mininno, V. (2019). Towards Analytics-Enabled Efficiency Improvements in Maritime Transportation: A Case Study in a Mediterranean Port. *Sustainability*, *11*, 4473. https://doi.org/10.3390/su11164473.

10. Seo, J., Lee, B.K., & Jeon, Y. (2023). Digitalization strategies and evaluation of maritime container supply chains. *Business Process Management Journal*, *29*(1), 1-21. https://doi.org/10.1108/BPMJ-05-2022-0241.

11. Yueshuai, B., & Chow, J.Y.J. (2019). Optimal privacy control for transport network data sharing. *Transportation Research Part C Emerging Technologies*, *113*, 370-387. https://doi.org/10.1016/j.trc.2019.07.010.

12. Davidich, N., Galkin, A., Davidich, Y., Schlosser, T., Capayova, S., Nowakowska-Grunt, J., Kush, Y., & Thompson, R. (2022). Intelligent Decision Support System for Modeling Transport and Passenger Flows in Human-Centric Urban Transport Systems. *Energies*, *15*, 2495. https://doi.org/10.3390/en15072495.

13. Greis, N.P., Nogueira, M.L., & Rohde, W. (2021). Digital Twin Framework for Machine Learning-Enabled Integrated Production and Logistics Processes, (pp. 218-227). In Dolgui, A., Bernard, A., Lemoine, D., von Cieminski, G., Romero, D. (Eds). Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems. APMS 2021. *IFIP Advances in Information and Communication Technology, 630*. Springer, Cham. https://doi.org/10.1007/978-3-030-85874-2_23.

14. Kim, Y.-J., Lee, J.-S., Pititto, A., Falco, L., Lee, M.-S., Yoon, K.-K., & Cho, I.-S. (2022). Maritime Traffic Evaluation Using Spatial-Temporal Density Analysis Based on Big AIS Data. *Applied Sciences*, *12*, 11246. https://doi.org/10.3390/app122111246.

15. Zhang, D., Zhang, Y., & Zhang, C. (2021). Data mining approach for automatic ship-route design for coastal seas using AIS trajectory clustering analysis. *Ocean Engineering*, *236*, 109535. https://doi.org/10.1016/j.oceaneng.2021.109535.

16. Montoya-Torres, J.R., Moreno, S., Guerrero, W.J., & Mejía,G. (2021). Big Data Analytics and Intelligent Transportation Systems. *IFAC-PapersOnLine, 54*(2), 216-220. https://doi.org/10.1016/j.ifacol.2021.06.025.

17. Dong, Y., Lingxiao, W., Shuaian, W., Haiying, J., & Li, K.X. (2019). How big data enriches maritime research – a critical review of Automatic Identification System (AIS) data applications. *Transport Reviews, 39*(6), 755-773. https://doi.org/10.1080/01441647.2019.1649315.

18. Bazaluk, O., Kotenko, S., & Nitsenko, V. (2021). Entropy as an Objective Function of Optimization Multimodal Transportations. *Entropy, 23*(8), 946. https://doi.org/10.3390/e23080946.

19. State Statistics Service of Ukraine (2023). *Freight transportation by water transport and type of cargo*. Retrieved from https://www.ukrstat.gov.ua.

20. Ukrainian Sea Ports Administration (2023). *Volumes of cargo processing by stevedoring companies in seaports*. Retrieved from https://www.uspa.gov.ua/.

Оцінювання й контроль релевантності даних в інформаційних системах управління транспортною галуззю

Ю. І. Клюс[*1], *М. А. Проданчук*[2], *О. О. Гололобова*[3], *Л. В. Галан*[4], *М. О. Варгатюк*[5]

1 – Східноукраїнський національний університет імені Володимира Даля, м. Київ, Україна

2 – Національний науковий центр «Інститут аграрної економіки», м. Київ, Україна

3 – Український державний університет науки і технологій, м. Дніпро, Україна

4 – Держаний університет інтелектуальних технологій і зв'язку, м. Одеса, Україна

5 – Одеський державний аграрний університет, м. Одеса, Україна

[*] Автор-кореспондент e-mail: klius@snu.edu.ua

Мета. Проведення аналізу даних, що надходять до інформаційних систем (ІС) транспортної галузі (ТГ) і розроблення пропозицій перевірки їх релевантності для забезпечення надійності та ефективності управління. Розроблення методу оцінювання й контролю релевантності даних в інформаційних системах транспортної галузі.

Методика. Використані спеціальні та загальні методи наукового пізнання: критичного аналізу для встановлення сценаріїв впливу релевантності даних на рівень ефективної взаємодії елементів системи та управління нею; контент-аналізу для стратифікації вимог до даних, що надходять до ІС, та управлінських рішень на етапі аналізу; математичної формалізації для розроблення методу оцінювання даних і контролю їх релевантності, що запропонований як математичне ядро згаданого алгоритму; метод послідовного підбору для автоматичного виявлення функціональної залежності, що описує дані з меншим значенням стандартної помилки.

Результати. Розроблені пропозиції із забезпечення релевантності даних. Запропонована стратифікація даних та управлінських рішень, що дозволить уникнути поширених помилок через ненормованість представлення даних. Встановлені сценарії впливу релевантності даних на рівень ефективної взаємодії елементів транспортної системи та управління нею. Для перманентного контролю даних запропоновано алгоритм виявлення відхилень у послідовних наборах даних, автоматичного аналізу вказаних відхилень і знаходження одно- та багатофакторних нелінійних закономірностей, що викликають відхилення. За невідповідності даних виявленим нелінійним закономірностям інформація перевіряється методом еволюційного програмування. Ефективність запропонованих методів продемонстрована на прикладі аналізу вантажообігу морських портів, тренди якого характерні значними змінами за часом і за портами. Ці тренди можуть мати навіть протилежний характер.

Наукова новизна. Розроблено метод оцінювання даних і контролю їх релевантності, що спирається на методи еволюційного («генетичного») програмування.

Практична значимість. Пропозиції щодо оцінювання й контролю релевантності даних в інформаційних системах транспортної галузі збільшать ефективність управлінської діяльності.

Ключові слова: *транспортна галузь, управління, інформаційні системи, контроль релевантності даних*

The manuscript was submitted 05.12.22.