A method has been proposed to predict the expected departure time for a cargo dispatch at the marshaling yard in a railroad system without complying with a freight trains departure schedule. The impact of various factors on the time over which a wagon dispatch stays within a marshaling system has been studied using a correlation analysis. The macro parameters of a transportation process that affect most the time over which a wagon dispatch stays within a marshaling system have been determined. To improve the input data informativeness, it has been proposed to use a data partitioning method that makes it possible to properly take into consideration the impact of different factors on the duration of downtime of dispatches at a station. A method has been developed to forecast the expected cargo dispatch time at a marshaling yard, which is based on the random forest machine learning method; the prediction accuracy has been tested. A mathematical forecasting model is represented in the form of solving the problem of multiclassification employing the processing of data with a large number of attributes and classes. A classification method with a trainer has been used. The random forest optimization was performed by selecting hyperparameters for the mathematical prediction model based on a random search. The undertaken experimental study involved data on the operation of an out-of-class marshaling yard in the railroad network of Ukraine. The forecasting accuracy of classification for dispatching from the wagon flow "transit without processing" is 86 % of the correct answers; for dispatching from the wagon flow "transit with processing" is 54 %.

The approach applied to predict the expected time of a cargo dispatch makes it possible to considerably improve the accuracy of obtained forecasts taking into consideration the actual operational situation at a marshaling yard. That would provide for a reasonable approach to the development of an automated system to predict the duration of operations involving cargo dispatches in a railroad system

Keywords: railroad, marshaling yard, cargo dispatch, expected departure time, machine learning

UDC 656.222
DOI: 10.15587/1729-4061.2020.209912

PREDICTING THE ESTIMATED TIME OF CARGO DISPATCH FROM A MARSHALING YARD

A. Panchenko
Department Artificial Intelligence and Software
V. N. Karazin Kharkiv National University
Svobody sq., 4, Kharkiv, Ukraine, 61022
E-mail: artempanchenkotop@gmail.com
A. Prokhorchenko
Doctor of Technical Sciences, Associate Professor*
E-mail: andrii.prokhorchenko@gmail.com
S. Panchenko
Doctor of Technical Sciences, Professor
Department of Automation and
Computer Telecontrol of Train Traffic**
E-mail: info@kart.edu.ua
O. Dekarchuk
Postgraduate Student*
E-mail: uer@kart.edu.ua
D. Gurin
Postgraduate Student*
E-mail: dmitriy.gurin1990@gmail.com
I. Medvediev
PhD
Department of Logistics Management
and Traffic Safety in Transport
Volodymyr Dahl East Ukrainian National University
Tsentralnyi ave., 59-a, Severodonetsk, Ukraine, 93400
E-mail: medvedev.ep@gmail.com
*Department of Operational Work Management**
**Ukrainian State University of Railway Transport
Feierbakha sq., 7, Kharkiv, Ukraine, 61001

Received date 10.07.2020
Accepted date 03.08.2020
Published date 31.08.2020

1. Introduction

Given the conditions of automation and digitalization of all industries, increasing the level of predictability of the technological process of a rail system transportation is a necessary requirement to reduce the cost of supply chains [1]. The operating model of railroad systems without following the schedule of freight trains possesses a considerable degree of uncertainty in the cargo transportation process. In such systems, the possibility of predicting the duration of operations involving cargo dispatch along a transportation chain to the destination becomes more complicated. This leads to an increase in the operating costs of a railroad system compared to railroads that follow the system of train traffic on schedule. To improve the competitiveness of a railroad system within which the traffic of trains fails to follow a timetable, it is important to build the functional for a periodic notification about the state of train dispatch. One of the important notification functions is the calculation of the estimated time of arrival (ETA) of a cargo dispatch [2]. Within the framework of the ETA calculation, which is determined for a cargo dispatch at each railroad station

along a destination route, the key function implies setting the estimated time of departure (ETD) of a cargo dispatch from the marshaling yard. Marshaling yards are the key elements in the technological process of cargo transportation. In those railroad systems within which the traffic of freight trains fails to follow a timetable, the time spent by wagon dispatches at the marshaling yards reaches 40 % of the total duration of transportation. Thus, it is a relevant task to resolve the issue of constructing more accurate forecasting methods for the most complex and unpredictable stage in a transportation process – passing a cargo dispatch through a marshaling yard.

2. Literature review and problem statement

Papers [3–7] report the results of studying the operation of marshaling yards in railroad systems that follow the timetable of freight trains. Each study is aimed at improving the planning of the operation of marshaling yards based on information technology and data analysis methods. However, the authors consider the specificity of railroad systems in terms of station operations, as well as their purpose.

It is shown that in those railroad systems that follow the timetable of freight trains, in such countries as Germany (Mannheim marshaling yard), Sweden (marshaling yard Hallsberg), Belgium (marshaling yard Antwerpen-Nood), detailed planning of operations takes place in advance (not less than a day) and the need for decisions on changes in existing plans is negligible. In those railroad systems within which the traffic of freight trains fails to follow a timetable, in such countries as CPR, India, Belarus, Kazakhstan, planning is based only on the actual plan of the traffic of freight trains towards a marshaling yard. A similar situation occurs in Ukraine's railroad system – the marshaling yards Osnova, Darnitsa, Znamianka work without coordinating the schedule of freight trains. This leads to a constant change in plans and significantly complicates setting the forecast time of cargo dispatches from marshaling yards. Given the high degree of uncertainty in the duration of operations at marshaling yards, there are no, up to now, any effective methods to forecast the time of operations involving cargo dispatches for those railroads within which the traffic of freight trains fails to follow a timetable. There are examples of studies aimed at predicting the arrival time of trains to a station in order to select the track for a train [8]. There are separate works [9–11] that propose mathematical models for constructing the daily operational plans for marshaling yards, among the results of solving which is determining the indicator of idle cars at a station. The main disadvantage of such mathematical models is the importance of the pre-knowledge about the arrival timetables at a station, which is very difficult to establish due to the accidental arrival of trains. This results in a significant level of inaccuracies in the results obtained. Moreover, the cited studies do not solve the task of determining the time that cargo dispatches are predicted to spend at a marshaling yard.

Paper [12] addressed the tasks to predict the freight train arrival time for U.S. railroads. That research is aimed at improving the accuracy of ETA forecasting by using machine learning methods. A given paper resolves the issue of predicting ETD for marshaling yards. The US railroads use specialized IT platforms that enable the detailed planning of a marshaling yard operation, as well as determining the time that wagons are expected to spend at a station [13, 14]. For example, studies [15, 16] report the search for the operational plans of railroad marshaling yards in North America, applying which makes it possible to determine the time that wagon dispatches are expected to spend at a station. The main drawback of using this approach for ETD prediction is similar to that inherent in Ukraine – the high dependence of accuracy on the knowledge about the arrival time of a train flow to a station.

The ETD forecasting for terminals and ports has been widespread in the maritime shipping industry [17–19]. The cited papers address the search for more accurate methods of ETD prediction; their authors note the complexity of predicting the duration of operations in terminals. Studies [20, 21] emphasize the importance of forecasting functions to reduce the costs of port terminals. The ETA and ETD functions are employed by port information systems [22]. Work [23] proposed machine learning (ML) methods in order to predict the travel time of cargo dispatches in multimodal container transportation. Several mathematical models have been built to resolve the set task. The algorithms of extremely randomized trees (ExtraTrees), Adaptive Boosting (AdaBoost), as well as the support vector regression (SVR), were suggested. When checked against the actual data on transportation, the forecasting model based on SVR proved to be the best in terms of the accuracy and adequacy of results. The accuracy of the forecast with an average absolute error is 17 hours when the duration of transportation is up to 30 days.

Approaches that involve machine learning methods are used for a similar prediction task in the aviation industry [24, 25] and automobile transportation [26, 27]. The research results prove that ML methods work more accurately than classical forecasting methods.

The complexity of an ETD forecasting task for marshaling systems at railroad stations requires the search for new approaches to constructing forecasting methods. An analysis of earlier studies has revealed that machine learning methods are promising. These methods make it possible to implement the inductive method of training on big data. This allows the description of complex and non-linear connections between the input attributes of a transportation process and the forecasts of a cargo dispatch at a station. The complexity of mathematical model construction implying the detailed plan of operations at those marshaling yards at which the traffic of trains fails to follow a timetable leads to significant inaccuracies in the developed methods of ETD prediction. This allows us to argue that it is advisable to undertake research, involving machine learning methods, aimed at finding those hidden dependences that affect the time that a cargo dispatch is expected to spend at a marshaling yard at the macro-level of its functioning.

3. The aim and objectives of the study

The aim of this study is to improve the efficiency of cargo transportation by rail, based on forecasting the expected time of a cargo dispatch using machine learning methods.

To achieve the set aim, the following tasks have been solved:

– to analyze the technological features of cargo handling in marshaling systems and to define the macro parameters for a transportation process, which influence the time the cargo dispatches spend at a station;

– to build a method to forecast the expected time of a cargo dispatch from a marshaling yard taking into consideration the determination of hidden dependences in the macro characteristics of a transportation process in those railroad systems within which the traffic of freight trains fails to follow a timetable;

– to check the accuracy and adequacy of results provided by the developed method for forecasting the expected time of a cargo dispatch.

4. Technological features in the construction of a system to forecast the time that a cargo dispatch is expected to spend at a marshaling yard

To solve the set task, the current work has considered in detail the sequence of operations involving cargo dispatches in a marshaling system at the typical non-categorized marshaling yard on a railroad network within which the traffic of freight trains fails to follow a timetable. The chosen object of analysis and experimental study is the non-categorized two-system marshaling yard Osnova from the regional branch of the JSC "Ukrzaliznytsya" Southern Railroad. The present study is a continuation of the research reported in [28]. Based on earlier results, we paid attention to a railcar flow from the marshaling yard Kupyans′k-Sortuval′nyy via the Osnova station to the Lyubotyn station of the Kharkiv railroad hub in the direction of the station Poltava-Sortuval′na. This makes it possible to combine the current research with the pre-conducted experiments on the construction of a predictive mathematical model of a train's movement through a railroad section. A given flow of cars passes the Northern marshaling system of the Osnova station, which has a consistent arrangement of yards. It should be noted that at a technical station the task of forecasting is complicated by the passing of a cargo dispatch involving different stages of technological processing in accordance with the category of the train arrived. This distinguishes this study from the earlier one, in which, within the framework of predicting the ETA for a cargo dispatch at the stage of passing a railroad section the object of forecasting was the time of a freight train travel between stations. It is proposed to consider two types of car flows within the scope of the implementation of a prediction function – transit without processing and transit with processing. Cargo dispatches arriving at a marshaling yard in trains that are not to be disbanded at the marshaling hill are accepted in a special transit yard where the sequence of operations is based on the regulatory technological process. After that, a cargo dispatch departs from the station by the same train in which it has arrived at the station. In this case, the time the cargo spends at the station is much shorter. Otherwise, a cargo dispatch arrives in the trains that imply the delivery to a station of the transit railcar flow with processing. The time it spends there is prolonged due to the execution of successive stages of processing, which implies its disbanding at a marshaling hill. The technological routes of processing the different types of railcar flow at a marshaling yard are shown in Fig. 1.

To analyze the time that wagons are expected to spend at a marshaling yard, we used data from the automated system ASK VP UZ-E [29]. The data covered the period from July through September 2017 on the time points of the implementation of all operations involving wagons from their arrival to the station until the moment of departure. The formed dataset included 6026 cargo dispatches, which were pre-processed and grouped. Each dispatch corresponds to the information on the number of the train that delivers the cargo, the number of wagons in the train, the mass, time, and date of arrival and departure from the station, the number of wagons in the dispatch, their destination stations, etc. It should be noted that a cargo dispatch is understood to be one and more freight cars, which, based on one transportation document, are delivered to one station. Using the data acquired, we performed a statistical analysis of the time that wagon dispatches are expected to spend at a station. The density distribution diagram of the time that wagons are expected to spend at a station according to the type of a railcar flow is shown in Fig. 2.

The maximum time that a cargo dispatch spends at a station is 2,874 minutes, minimum – 30 minutes, the average time spent at a station is 741.8 minutes, the median is 610 minutes. The time that wagons from the railcar flow without processing spend at a station is much smaller in comparison with the railcar flow with processing (Fig. 2). The average time for the transit without processing is 128 minutes, for the transit with processing is 920 minutes. In the general distribution between the types of transit, the share of transit without processing is 62 % of the total quantity of dispatches in the marshaling system, which leads to greater uncertainty in the railcar idling time.

An analysis of the number of wagons in the cargo dispatch reveals that the type of railcar flow affects this indicator. A dispatch can include from one to 63 cars, which corresponds to the composition of a route train, the average number of cars in the dispatch is 6.99 wagons, the median is 2 cars. The distribution density diagram of the number of wagons in a cargo dispatch by the types of a railcar flow is shown in Fig. 3. The dispatches belonging to the railcar flow "transit without processing" have a significantly higher number of wagons in their composition; their distribution demonstrates a "long tail". The distribution of the number of wagons shows a much smaller tail, which corresponds to not more than 40 wagons in a dispatch, whereas the mathematical expectation is 4.2 wagons.

An important factor affecting the duration of downtime is the condition of wagons in a cargo dispatch. The distribution density of the time that a cargo dispatch spends at a station (Fig. 4) reveals that empty wagons spend a longer time at a station than those loaded. This fact can be taken into consideration in the construction of a mathematical forecasting model.

The complexity and speed of operations involving cargo dispatches in a marshaling system are affected by the degree to which railroad tracks in various yards are filled with wagons. When almost all tracks in a marshaling yard are busy, there is a delay in the operation the disbanding of a train, which includes a cargo dispatch. At the same time, the presence of a sufficient number of wagons on the tracks of a marshaling yard accelerates the process of accumulation of wagons for the possibility of forming a freight train and its dispatch. To account for this dependence, it is proposed to take into consideration the impact of the number of wagons – the working fleet of wagons in a sorting system at the moment of arrival of a cargo dispatch to the station.

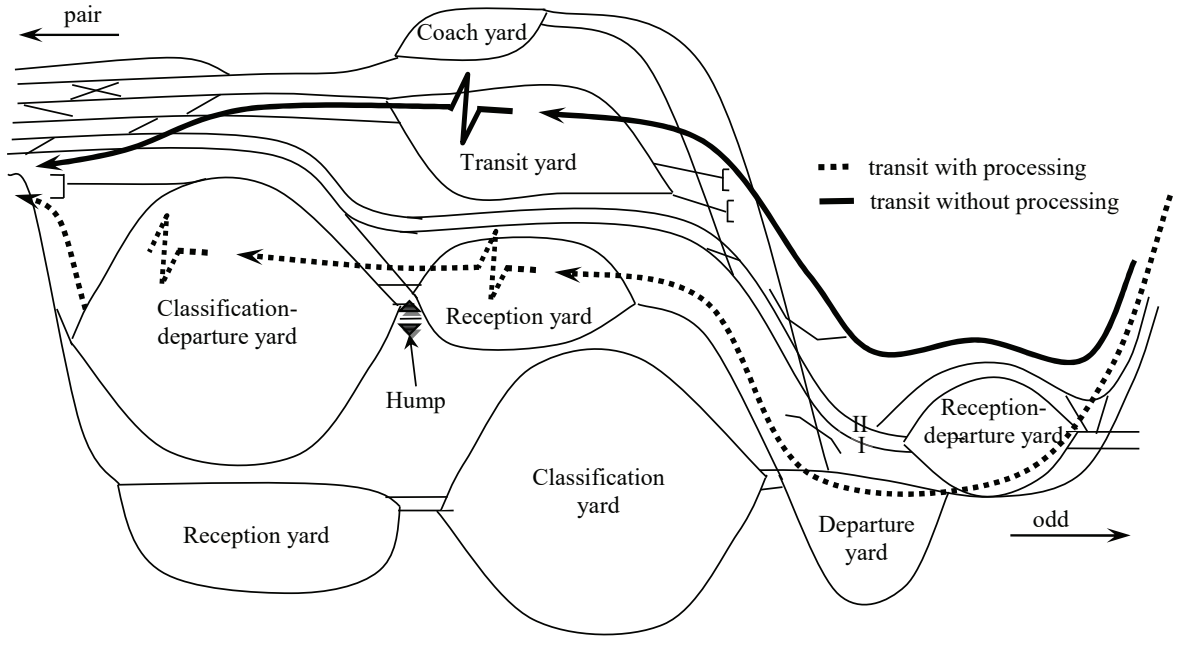

Fig. 1. Technological route for the routing of railcar flow involving the transit with processing and without processing in the direction of the station Poltava via the northern system of the Osnova station

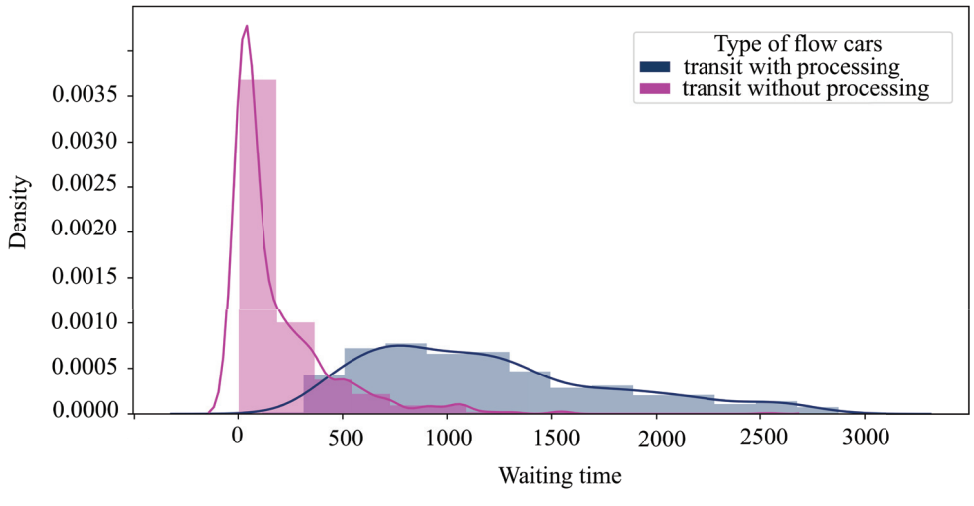

Fig. 2. The density distribution diagram of the time that wagons spend at a station according to the type of a railcar flow without processing and transit with processing, respectively

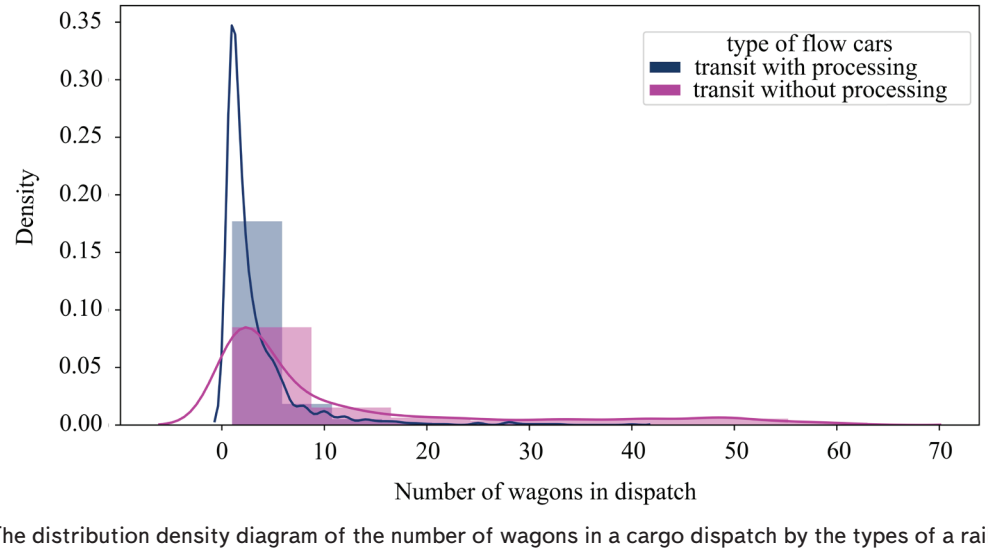

Fig. 3. The distribution density diagram of the number of wagons in a cargo dispatch by the types of a railcar flow

According to the technological process of transportation, different types of railcar flows follow a different plan of transportation that affects their downtime at the station. The previous analysis produced 76 destination stations for cargo dispatches (Fig. 5). To reduce their number and simplify the analysis, we examined the normative document based on which a train forming planning (TFP) is executed [30], developed and approved for each station in the network for a freight year. This document defines the direction and category of the train, which would include wagons from the railcar flow "transit with processing". According to the technology of operations, after passing a sorting hill (Fig. 1), wagons follow the tracks in a marshaling yard where the duration of downtime depends on the total number of wagons for a destination according to TFP. The more the railcar flow at the destination station, the faster the wagons accumulate to comply with the norm for mass and length, which makes it possible to form a freight train of the predefined category according to the TFP and dispatch it from the station. The visualization of the destination graph, according to the TFP for the Osnova station in the direction of a railcar flow with processing via the Lyubotyn station is shown in Fig. 5.

Fig. 6 show the diagram of dependence of destination stations on the number of dispatches. An analysis of Fig. 6, *a* reveals that destination station No. 40 has the largest number of dispatches while a significant quantity of stations has a small enough share in the total volumes of dispatch.

Based on comparing a transportation plan with actual data on the cargo dispatch destination stations, it has been proposed that all dispatches in the systems should be distributed according to five destinations (Fig. 6, *b*). This could simplify the analysis and systematize the impact exerted by a factor that accounts for the dispatch destination on the time that a cargo dispatch spends at a station.

Given the significant deviation in the time that cargo dispatches spend at a station on the average value, in this study we proposed splitting a series of continuous quantities into fifteen segments with the uniform density distribution. The dependence of the interval duration of downtime, corresponding to fifteen time segments, on the number of dispatches of various types of a railcar flow is shown in Fig. 7.

The application of a data partitioning method makes it possible to better take into consideration the impact of different factors on the duration of downtime of dispatch at a station. Based on a similar approach, we split the data on the number of wagons in a dispatch into five segments. To analyze the effects of the above-mentioned factors on the time that a cargo dispatch spends at a marshaling yard, a correlation analysis was conducted [31]. The Pearson correlation matrix is given in Table 1. Marked correlations are significant at $p<0.05$.

The analysis of connections' density can be characterized as moderate and weak, confirming the weakly-structured character of the problem being solved. It should be noted that obtaining other data, which could improve the quality of information about the time that cargo dispatches spend at a marshaling yard, is quite problematic. In accordance with this, it is important to conduct research and try to obtain the results of forecasting based on available data.

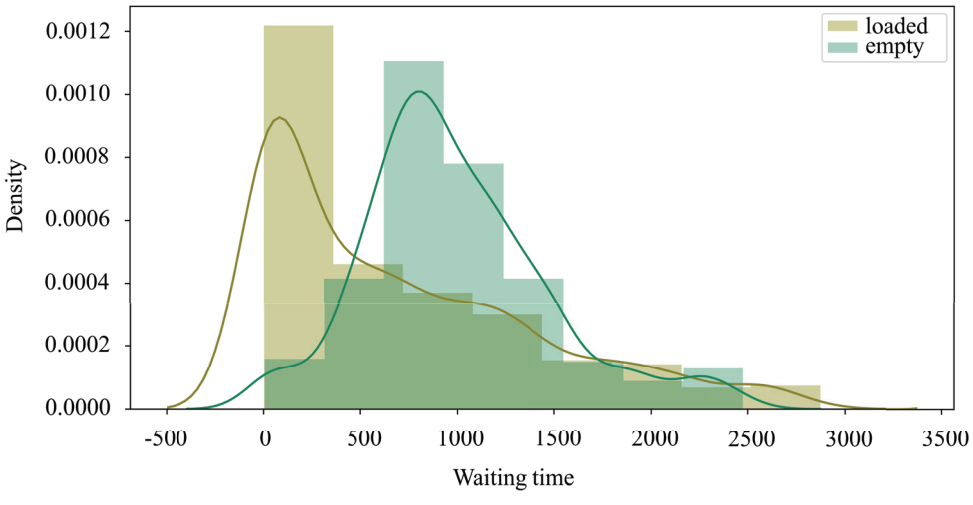

Fig. 4. The distribution density of the time that a cargo dispatch spends at a station for the loaded and empty states

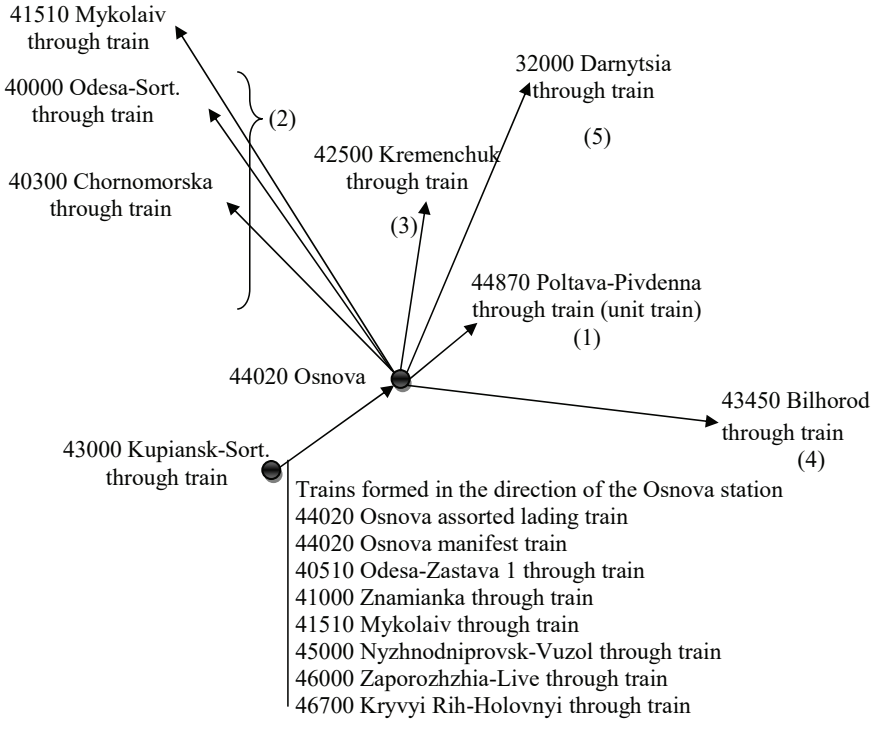

Fig. 5. The destination graph for dispatching a railcar flow in transit with processing from the Osnova station according to the train forming plan for the period of 2017–2018

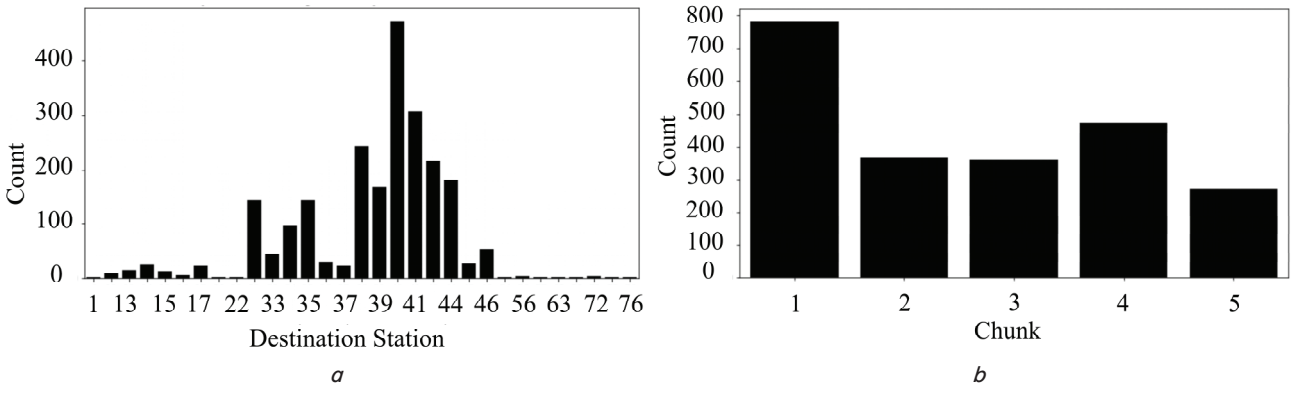

Fig. 6. Diagram of the number of dispatches for destinations from the Osnova station: *a* — to the destination stations of cargo dispatch; *b* — to the destination stations according to the train forming plan for the Osnova station

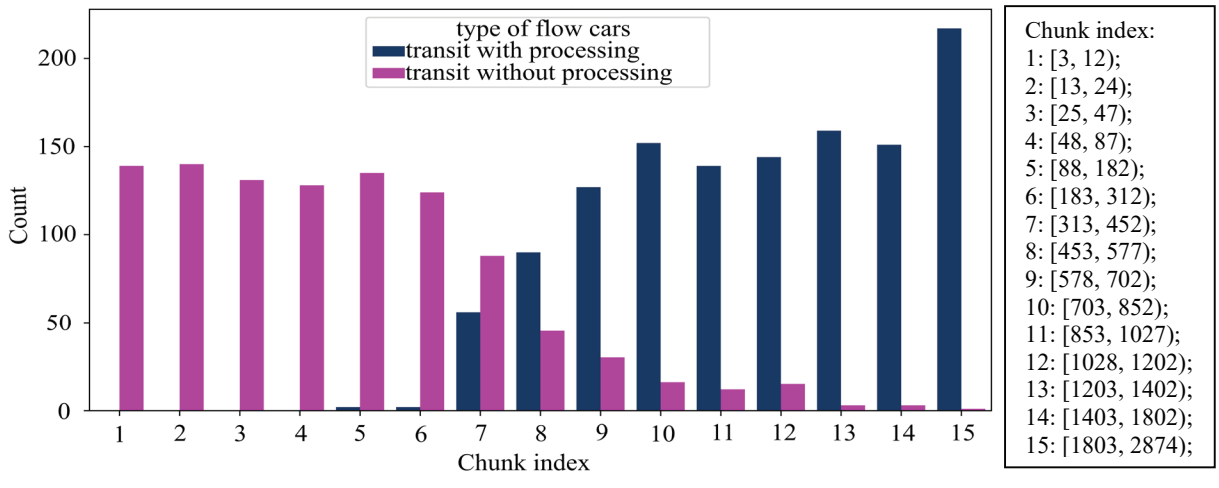

Fig. 7. Dependence of the interval values of a downtime duration that correspond to fifteen segments on the number of dispatches of various types of a railcar flow

Table 1

Pearson correlation matrix

Analyzed factor	The time that a cargo dispatch spends at a station	Number of wagons per a dispatch	Number of wagons at a station	Wagon status	Railcar flow type	Destination based on TFP
The time that a cargo dispatch spends at a station	1	–0.31816	–0.25245	–0.41958	–0.81176	–0.43157
Number of wagons per a dispatch	–0.31816	1	0.149888	0.283187	0.35957	0.94857
Number of wagons at a station	–0.25245	0.149888	1	0.118531	0.378865	0.159512
Wagon status (loaded/empty)	–0.41958	0.283187	0.118531	1	0.335086	0.259449
Railcar flow type	–0.81176	0.35957	0.378865	0.335086	1	0.375723
Destination based on TFP	–0.43157	0.94857	0.159512	0.259449	0.375723	1

5. Construction of a method for predicting the expected time of a cargo dispatch from the marshaling yard

Given that the purpose of building a mathematical model for predicting the ETD a wagon dispatch is the comprehensive implementation of ETA prediction function along the entire route, it is important to ensure the versatility of the approach. This could ensure its scalability for other technical stations within a railroad network. Earlier analysis proved the hierarchy of the influence exerted by various factors, which requires that their distribution should be based on the groups according to the predefined characteristics – the interval values of a downtime duration, corresponding to fifteen time segments. In such a statement, the prediction method requires the implementation of a classification problem implying the processing of data with a large number of attributes and classes.

The most appropriate approach that matches the set task is to use a Random Forest (RF) algorithm. A given algorithm refers to the methods of machine learning (ML) [32]. A prediction method based on the random forest algorithm has a series of disadvantages regarding the complexity of

solving problems with acceptable accuracy on the input data with much noise. This requires conducting an experimental study into the possibility of using a given method to the problem of such a class.

The predictive model building algorithm is implemented in the Python programming environment [33]. To build the predictive model, we chose the type of the random forest involving the classic splitting criterion for a classification problem – the gini coefficient. The random forest was optimized by the selection of hyperparameters for a mathematical predictive model based on a random search using the RandomizedSearchCV algorithm from a Scikit-Learn library [33]. Some results of the selection of random forest model parameters are given in Table 2. Based on the optimization results, we defined the parameters of the random forest – 46 iterations. The number of trees in the forest is 373 (n_estimators). The minimum number of splitting is 5 ('min _samples_ split'). The maximum number of splitting ('max_features') is defined by sqrt(n). The maximum tree depth ('max_depth') is 145. The limit on the number of objects in the leaves ('min_samples_leaf') is 1.

To test the accuracy and adequacy of the mathematical predictive model, we highlighted 33 % of the test lines for incoming data from the total examined sample size.

Table 2

Results of selecting the parameters for a mathematical predictive model using the RandomizedSearchCV algorithm

iteration	mean score on a test sample	mean score on a training sample	maximal tree depth	maximal splitting number	the limit for the number of objects in leaves	minimal splitting number	number of trees in a forest
135	0.612099644	0.990796473	23	log2	1	5	2,805
41	0.602016607	0.977443138	159	log2	1	7	3,326
67	0.585409253	0.957573607	118	log2	3	5	2,110
141	0.573546856	0.935632036	50	log2	1	10	894
79	0.559905101	0.892937461	132	log2	1	13	1,068
53	0.556939502	0.89263802	159	log2	1	13	3,152
88	0.534994069	0.875433117	172	log2	1	10	894
57	0.53024911	0.840441822	200	log2	5	13	2,805
116	0.529655991	0.883426927	159	log2	3	7	547
20	0.529062871	0.83362562	172	log2	1	18	894
115	0.517200474	0.799522536	200	log2	1	21	2,805
77	0.515421115	0.8386717	105	log2	3	10	1,415
98	0.514827995	0.84045236	23	log2	3	10	3,326
46	0.697827995	0.974493074	173	sqrt	1	5	373
145	0.513048636	0.798939921	77	log2	7	7	721
134	0.510676157	0.797437341	23	log2	7	7	2,978
131	0.509489917	0.939192368	200	auto	1	10	2,631
90	0.509489917	0.945120504	77	sqrt	3	7	1,589
142	0.507117438	0.954005847	105	sqrt	3	5	2,631
89	0.507117438	0.953713762	37	auto	3	5	2,631

The input parameters defining the dependence of time that a cargo dispatch spends at a station are the following: to what type of a railcar flow it refers – transit without processing and transit with processing; the number of wagons per a dispatch; the status of wagons – empty or loaded; a working yard when a cargo dispatch arrives at a station; its destination based on TFP. The outputs are the segments' numbers, corresponding to the accepted time intervals that a cargo dispatch spends at a station. The classifier converts the vector of attributes of the technological process of transportation at a marshaling yard into an integer that corresponds to the segment number, which correlates the time interval to the possible time that a cargo dispatch spends at a station until the moment of its departure.

6. Checking the accuracy and adequacy of results of predicting the expected departure time of a cargo dispatch from a marshaling yard

Based on the results of testing the constructed mathematical model, the average 70 % accuracy was obtained on the test sample. However, to assess the accuracy of prediction results, it is important to verify the accuracy and adequacy of the results obtained separately for each type of a railcar flow to which a cargo dispatch belongs.

For a dispatch from the railcar flow of transit without processing, we have obtained the accuracy of the mathematical model on a test sample [31] at the level of 0.863, or 86 % of correct answers. Given the imbalances of classes, we have determined the balanced accuracy in order to improve the quality of accuracy assessment, which is 0.797, or 80 % of the correct results. The results of predicting the classification of intervals of the time that a cargo dispatch from the transit flow without processing spends at a marshaling yard are shown in Fig. 8.

The estimation of the classification prediction accuracy on the test sample for a cargo dispatch from the transit flow with processing is 0.541, or 54 % of the correct answers. The balanced accuracy was 0.483, or 48 % of the correct results. The results of the forecast for a cargo dispatch from the railcar flow transit with processing are shown in Fig. 9.

The obtained accuracy of predicting the time that a cargo dispatch spends at a station is acceptable from a practical point of view. Given the acting regulations regarding the time that a cargo dispatch spends at a station based on the type of a railcar flow according to the technological process at the Osnova marshaling yard, the accuracy is high enough. The average downtime of a wagon from the transit without processing, according to the technological process, is 1.82 hours (109.2 minutes), and transit with processing – 20.5 hours (1,230 minutes). If we evaluate the forecast of the time that a cargo dispatch spends at a station according to the norms, the estimation of the accuracy of predicting the delay of a dispatch from the norm is 98 % of the correct results for the transit without processing, and 91 % for the transit with processing.

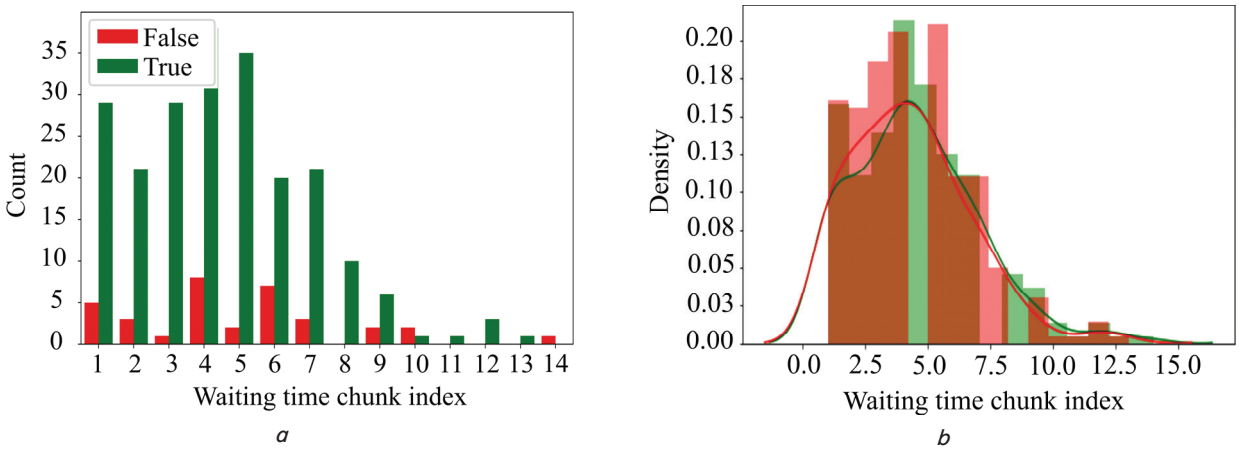

Fig. 8. The results of predicting the classification of intervals of the time that a cargo dispatch from the transit flow without processing spends at the Osnova marshaling yard: *a* — comparative chart of the frequency of true and false values while performing the forecast; *b* — the density of distribution of the true and false values while performing the forecast

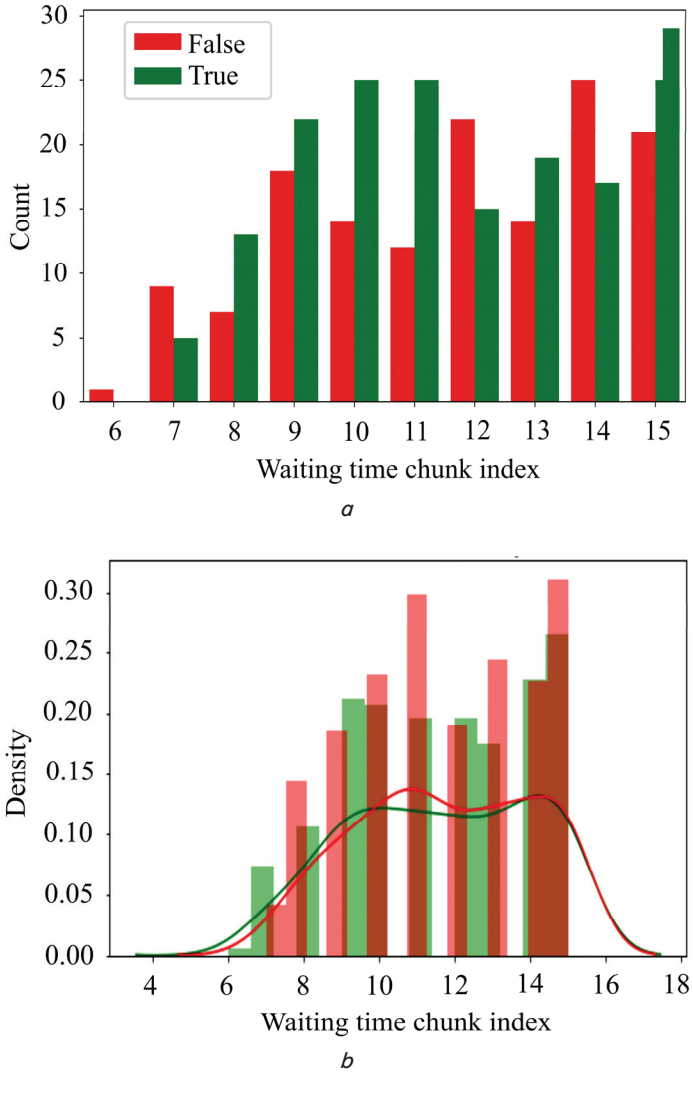

Fig. 9. The results of predicting the classification of intervals of the time that a cargo dispatch from the transit flow with processing spends at the Osnova marshaling yard: *a* — a comparative chart of the frequency of true and false values while performing the forecast; *b* — the density of distribution of the true and false values while performing the forecast

7. Discussion of results of studying the prediction of the expected time of a cargo dispatch at a marshaling yard

The obtained forecasting results of the expected cargo dispatch time at a marshaling yard indicate that the proposed method of prediction based on the random forest algorithm is more accurate compared to the classic approaches based on microparameters [9–11]. The results have been confirmed by our experiment using actual data on the operational work of a non-categorized marshaling yard in the railroad network of Ukraine. The accuracy of classification prediction in terms of the accuracy indicator for a cargo dispatch from the railcar flow "transit without processing" is 86 % of the correct answers; a cargo dispatch from the railcar flow "transit with processing" is 54 %. In a railroad system within which the traffic of freight trains fails to follow a timetable, it is difficult to estimate the time that a cargo dispatch spends at a marshaling yard for reasons of the significant uncertainty in a transportation process. It is possible to rely on the set norms for the downtime of cargo dispatches from different types of railcar flows at a station but such estimates could produce a significant error on the actual data about the transportation process, which might amount to several days.

The approach applied to predict the time that a cargo dispatch spends at a station makes it possible, under the limited data informativeness, to significantly improve the accuracy of the forecasts obtained. That has been implemented owing to the use, for solving the problem on forecasting, of the ensemble method of machine learning – a random forest. A given method can identify complex nonlinear interrelations among the macro parameters in the operational work of a marshaling yard.

The advantage of using the method that forecasts the expected time of a cargo dispatch is the possibility to use data from automated traffic dispatching systems, which can be acquired from any railroad system within which the traffic of trains fails to follow a timetable. The data on the parameters underlying the developed method have low detailing of the transportation process, which complicates the forecasting process. This renders difference to stating the problem in comparison with the accuracy

of data on the time that cargo dispatches spend at marshaling yards in railroad systems that follow the timetable of freight trains [3–7]. Consequently, the estimated accuracy of the predictive mathematical model is quite acceptable under practical conditions for the problems of such a class of uncertainty. No similar accuracy results, obtained for railroad systems within which the traffic of trains fails to follow a timetable, have been reported. The only method of checking the results of this study was the expert estimation by station personnel, based on the standards for downtime of cargo dispatches of different types of railcar flows at a station. The inaccuracy of such projections was calculated in days.

The limitations of the developed forecasting approach include the weakly structured input data, underlying decision-making, and a possibility to retrain the random forest using data from another marshaling yard in the network. To improve a given approach, additional research should be performed to find the generic algorithm parameters in order to obtain predictive results at all route stations with an acceptable accuracy level. The limitation inherent in a given approach is the difficulty in finding additional input parameters of a transportation process, which could improve the results of forecasting.

In further research, it is important to test the performance of the developed mathematical predictive model in combination with the pre-built model of forecasting the route of a cargo dispatch in trains along a railroad section [28]. In conjunction, this would make it possible to develop, for practical application, an automated prediction system for notifying about the expected arrival time of a cargo dispatch. The proposed method adequately reproduces the technology of transportation in a railroad system within which the traffic of freight trains fails to follow a timetable. In practical terms, the advantage of the developed forecasting method involving a classifier based on the random forest is its scalability and versatility. This allows its application in an integrated automated system for predicting the estimated time of arrival (ETA) for a cargo dispatch along the entire route of transportation.

8. Conclusions

1. We have investigated the technological features in the construction of a system for predicting the time it takes for a cargo dispatch to pass a marshaling yard. It has been proposed, to improve the accuracy of forecasting, to take into consideration which type of a railcar flow each dispatch belongs to – transit without processing and transit with processing. We have determined those macro parameters of a transportation process that characterize the processing of cargo dispatches in a sorting system. These parameters include the following: which type of a railcar flow a dispatch belongs to; the number of wagons per a dispatch; the status of wagons – empty or loaded; the working yard when a cargo dispatch arrives at the station; the TFP-based destination. Our correlation analysis has proven the moderate and weak connections in the influence of these attributes on the time that wagon dispatches spend in a sorting system. To improve the input data informativeness, it has been proposed to apply a data partitioning method, which makes it possible to better take into consideration the impact exerted by different factors on the duration of downtime of dispatches at a station.

2. We have constructed a method to forecast the time that a cargo dispatch spends at a marshaling yard, based on machine learning, specifically a random forest algorithm. The method of forecasting is proposed in the form of a solution to the multiclassification problem implying the processing of data with may attributes and classes. The input parameters of the model are: which type of a railcar flow a dispatch belongs to – transit without processing and transit with processing; the number of wagons per a dispatch; the status of wagons – empty or loaded; the number of wagons when a cargo dispatch arrives at a station; the TFP-based destination. The outputs are the segments' numbers, corresponding to the accepted time intervals when a cargo dispatch is at the station. A classification method with a trainer has been applied. We have chosen the parameters for a random forest algorithm and performed an experimental study involving data on the operation work of a non-categorized marshaling yard in the network of Ukraine, which proved the possibility of obtaining acceptable results of forecasting for the problems of such a class.

3. To confirm the accuracy and adequacy of the developed method to forecast the expected departure time for a cargo dispatch, we have tested the performance of the generated random forest on a test data sample. The evaluation of accuracy of the classification prediction on the test sample for a cargo dispatch from the transit flow without processing is 86 % of the correct answers; for a cargo dispatch from the transit flow with processing is 54 % of the correct answers. The prediction of the time that a cargo dispatch spends at a station has been estimated according to the standards of the technological process at the Osnova marshaling yard. The accuracy of forecasting the delay of dispatch from standards is 98 % of the correct results for transit without processing and 91 % for transit with processing. From a practical point of view, the obtained accuracy of forecasting is acceptable and quite high.

References

1. Sapronova, S., Tkachenko, V., Fomin, O., Hatchenko, V., Maliuk, S. (2017). Research on the safety factor against derailment of railway vehicless. Eastern-European Journal of Enterprise Technologies, 6 (7 (90)), 19–25. doi: https://doi.org/10.15587/1729-4061.2017.116194
2. Cameron, M., Brown, A. (1995). Intelligent transportation system Mayday becomes a reality. Proceedings of the IEEE 1995 National Aerospace and Electronics Conference. NAECON 1995. doi: https://doi.org/10.1109/naecon.1995.521962
3. Deliverable D2.2. Draft Recommendations for Improved Information and Communications for Real-Time Yard and Network Management. Available at: https://optiyard.eu/wp-content/uploads/2019/02/OptiYard_Deliverable2.2_Final.pdf
4. Khoshniyat, F. (2012). Simulation for Planning Strategies for Track Allocation at Marshalling Yards. Stockholm.
5. Alzén, C. (2015). Trafikeringsplan Hallsbergs rangerbangård. Handbok BRÖH 313.00700. Banverket.
6. Jacobsson, S., Arnäs, P. O., Stefansson, G. (2017). Access management in intermodal freight transportation: An explorative study of information attributes, actors, resources and activities. Research in Transportation Business & Management, 23, 106–124. doi: https://doi.org/10.1016/j.rtbm.2017.02.012

7. Simonović, M., Vitković, N., Miltenović, A., Ristić, D. (2017). Report No. 730836. Deliverable D6.1 Architectural design of the information system for supervision and management of marshalling yards. SMART Smart Automation of Rail Transport.
8. Bardas, O. O. (2016). Improving the intelligence technoligies of train traffic's management on sorting stations. Transport Systems and Transportation Technologies, 11, 9–15. doi: https://doi.org/10.15802/tstt2016/76818
9. Erofeev, A., Fedorov, E. (2016). Planning for the formation of trains in the system of intellectual management of transportation process. Transport Systems and Transportation Technologies, 12, 16–24. doi: https://doi.org/10.15802/tstt2016/85881
10. Kozachenko, D. M., Vernigora, R. V., Korobyova, R. G. (2008). The Software Package for Simulation of Railway Stations Based on Plan-Schedule. Zaliznychnyi transport Ukrainy, 4, 18–20.
11. Prokhorchenko, A. V., Prokhorov, V. M., Postolenko, A. Yu. (2011). Rozroblennia modeli formuvannia planu roboty sortuvalnoi stantsiyi na osnovi teoriyi rozkladu. Zbirnyk naukovykh prats Ukrainskoi derzhavnoi akademiyi zaliznychnoho transportu, 120, 38–43.
12. Barbour, W., Samal, C., Kuppa, S., Dubey, A., Work, D. B. (2018). On the Data-Driven Prediction of Arrival Times for Freight Trains on U.S. Railroads. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). doi: https://doi.org/10.1109/itsc.2018.8569406
13. RailConnect*. Rail Yard Management System. Available at: https://zerista.s3.amazonaws.com/item_files/7a8b/attachments/10454/original/rc_rail_yard_management_get.pdf
14. GE Transportation's Digital Solutions, Movement Planner Network Viewer. Available at: https://www.ge.com/digital/sites/default/files/download_assets/GE-Transportation-Movement-Planner-NV-20160824_0.pdf
15. Zhou, W., Yang, X., Qin, J., Deng, L. (2014). Optimizing the Long-Term Operating Plan of Railway Marshalling Station for Capacity Utilization Analysis. The Scientific World Journal, 2014, 1–13. doi: https://doi.org/10.1155/2014/251315
16. Lin, E., Cheng, C. (2009). YardSim: A rail yard simulation framework and its implementation in a major railroad in the U.S. Proceedings of the 2009 Winter Simulation Conference (WSC). doi: https://doi.org/10.1109/wsc.2009.5429654
17. Xiao, Z., Ponnambalam, L., Fu, X., Zhang, W. (2017). Maritime Traffic Probabilistic Forecasting Based on Vessels' Waterway Patterns and Motion Behaviors. IEEE Transactions on Intelligent Transportation Systems, 18 (11), 3122–3134. doi: https://doi.org/10.1109/tits.2017.2681810
18. Lechtenberg, S., Braga, D., Hellingrath, B. (2019). Automatic Identification System (AIS) data based Ship-Supply Forecasting. Digital Transformation in Maritime and City Logistics: Smart Solutions for Logistics. Proceedings of the Hamburg International Conference of Logistics. doi: https://doi.org/10.15480/882.2487
19. Pani, C., Vanelslander, T., Fancello, G., Cannas, M. (2015). Prediction of late/early arrivals in container terminals – A qualitative approach. European Journal of Transport and Infrastructure Research, 15(4), 536–550. doi: http://doi.org/10.18757/ejtir.2015.15.4.3096
20. Fancello, G., Pani, C., Pisano, M., Serra, P., Zuddas, P., Fadda, P. (2011). Prediction of arrival times and human resources allocation for container terminal. Maritime Economics & Logistics, 13 (2), 142–173. doi: https://doi.org/10.1057/mel.2011.3
21. Salleh, N. H. M., Riahi, R., Yang, Z., Wang, J. (2017). Predicting a Containership's Arrival Punctuality in Liner Operations by Using a Fuzzy Rule-Based Bayesian Network (FRBBN).The Asian Journal of Shipping and Logistics, 33 (2), 95–104. doi: https://doi.org/10.1016/j.ajsl.2017.06.007
22. Parolas, I., Tavasszy, L., Kourounioti, I., van Duin, R. (2017). Predicition of Vessels' estimated time of arrival (ETA) using machine learning: A port of Rotterdam case study. Proceedings of the 96th Annual Meeting of the Transportation Research. Washington, 8–12.
23. Servos, N., Liu, X., Teucke, M., Freitag, M. (2019). Travel Time Prediction in a Multimodal Freight Transport Relation Using Machine Learning Algorithms. Logistics, 4 (1), 1. doi: https://doi.org/10.3390/logistics4010001
24. Wang, Z., Liang, M., Delahaye, D. (2018). Automated Data-Driven Prediction on Aircraft Estimated Time of Arrival. Eighth SESAR Innovation Days.
25. Shang, Y., Dunson, D., Song, J.-S. (2017). Exploiting Big Data in Logistics Risk Assessment via Bayesian Nonparametrics. Operations Research, 65 (6), 1574–1588. doi: https://doi.org/10.1287/opre.2017.1612
26. Altinkaya, M., Zontul, M. (2013). Urban bus arrival time prediction: A review of computational models. International Journal of Recent Technology and Engineering (IJRTE), 2 (4).
27. Sun, X., Zhang, H., Tian, F., Yang, L. (2018). The Use of a Machine Learning Method to Predict the Real-Time Link Travel Time of Open-Pit Trucks. Mathematical Problems in Engineering, 2018, 1–14. doi: https://doi.org/10.1155/2018/4368045
28. Prokhorchenko, A., Panchenko, A., Parkhomenko, L., Nesterenko, G., Muzykin, M., Prokhorchenko, G., Kolisnyk, A. (2019). Forecasting the estimated time of arrival for a cargo dispatch delivered by a freight train along a railway section. Eastern-European Journal of Enterprise Technologies, 3 (3 (99)), 30–38. doi: https://doi.org/10.15587/1729-4061.2019.170174
29. Naumenko, P. P., Minenko, V. D., Zemlyanov, V. B. (2007). Single automated control system of "Ukrzaliznytsia" freight transportation as the basis for the integration of automated control systems of freight railway transport of Ukraine. Visnyk Dnipropetrovskoho natsionalnoho universytetu zaliznychnoho transportu imeni akademika V. Lazariana, 17, 35–40.
30. But'ko, T., Prokhorchenko, A. (2013). Investigation into Train Flow System on Ukraine's Railways with Methods of Complex Network Analysis. American Journal of Industrial Engineering, 1 (3), 41–45.
31. Aris, S. (1999). Probability Theory and Statistical Inference: Econometric Modeling with Observational Data. Cambridge University Press. doi: https://doi.org/10.1017/cbo9780511754081
32. Breiman, L. (2001). Random Forests. Machine Learning, 45 (1), 5–32. doi: https://doi.org/10.1023/a:1010933404324
33. Richert, W., Coelho, L. P. (2013). Building Machine Learning Systems with Python. Birmingham: Packt Publishing, 290.