Russian Freight Flights Time Prediction

Ilya Makarov
School of Data Analysis and Artificial Intelligence
National Research University Higher School of Economics
125319, 3 Kochnovskiy Proezd, Moscow, Russia
iamakarov@hse.ru

***Abstract*—We present a model for freight train time prediction based on station network analysis and specific feature engineering. We discuss the first pipeline to improve the freight flight duration prediction in Russia. While every freight company use only reference book made by RZD (Russian Railways) based on railroad distances with accuracy measured in days, we argue that one could predict the flight duration with error less than twenty hours while decreasing error to twelve hours for certain type of freight trains.**

Keywords: Rail freight operations; train time prediction; operation management; machine learning.

I. Introduction

Railway transportation is one of the most common types of cargo transportation in the world [1]. For Russia, rail transport is provide over 80% of freight turnover according to Russian Control Engineering reports [2]. With the help of railways, the most important resources are transported: oil and oil products, coal, iron ore, construction materials, chemical and mineral fertilizers [3].

In view of the fact that most of the freight is carried out by several trains without a clearly defined stop schedule, forecasting the length of a loaded freight train flight is a hard problem. In addition, its duration may increase due to unforeseen breakdowns of the railway carriage, train or railway tracks thus reducing possibility to accurately plan train schedule for the carrier company and the consignee company. As a result, incorrect planning increases the transport companies' costs and leads to non-optimal train control.

Analysis and prediction of the railway flight duration is a well-studied problem. One of the first observations connecting railway networks with social network analysis was mentioned in [4] fact that the distribution of delays in passenger railroads is close to the power law, which is a usual case for real-world social and interaction networks. In simple worlds, such distribution means that there are many trains with small delays, however, there are certain number of long delays.

One of the directions for railway flight research was made using the Time Events graph based on the Petri network [5]–[7]. The railway flight was presented as a set of processes and events that are markers of the beginning or end of the process. As a process, the movement of the train was accepted, the train was waiting, and as an event the authors consider the arrival of the train and the departure of the train. Dependencies between processes and events were described by the graph for each section of the railway. Then, this graph was filled with historical data, and the delay propagation model was applied. As a continuation of this study, a system combining prediction of flight duration and schedule changes was suggested in [8], while simultaneously incorporating prediction component and evaluating the impact of these predictions on decision making made by using predicted durations and traffic load changes.

In the studies presented above, the problem of predicting the duration of passenger flights with a certain timetable is solved, which is quite different from the problem posed in this paper, since flights without a clear timetable are highly variable [9].

One of the approaches for scheduling cargo flights is using game theory approaches for bidding price of certain solutions [10]. The problem of cargo transportation planning was also studied using the construction of a network in which cargo flows are distributed [11], [12]. However, these models do not allow to predict the duration of the flight.

The problem of decision making for dependent train arrivals at intermodal freight transfer terminals was considered in [13]. The analysis of transport terminals efficiency using Timed Petri networks was suggested in [14].

Probabilistic approach for stability of average train time arrival was suggested in [15]. On the other hand, there are many negative factors affecting the possibility of accurate prediction in our task [16]. We also take into account interconnections between traffic routing problems for Internet subnetworks based on Recurrent neural networks and Time series analysis for the signal transmission time prediction [17].

However, it is difficult to use these methods to solve the problem posed in this paper, since there is much less data in the field of rail transportation, and it is impossible to compile a representative sample of time series for any two stations due to lack of trustful data and digitalized expertise at Russian railway stations widely spread through the largest country in the world.
For further study we address a reader to a survey on methods for train flight duration prediction in transportation decision and control [18].

II. Preliminaries and Problem Statement

In this paper we study the problem of freight traffic duration estimation, which even for two different freight trains with the same route may vary depending on transportation and several unobservable factors.

Any transportation is described by invoice, which in case of Russian Railways contains the departure station, destination

The work was supported by the Russian Science Foundation under grant 17-11-01294 and performed at National Research University Higher School of Economics, Russia.

978-1-7281-0108-8/19/$31.00 ©2019 IEEE

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on January 20,2022 at 10:50:06 UTC from IEEE Xplore. Restrictions apply.

station, the name and weight of the cargo, the type of train (RPS), the date of departure and arrival of the train. In addition, in the process of transportation, certain dislocation events are formed characterized by the date, station of dislocation and the name of the operation, which was carried out with the train at this station (proceeded without stopping, loading, unloading, casting).

In this paper, we overview a regression problem for prediction of the freight train duration based on stations network analysis, historical data of freight flights and factor analysis of invoice data.

III. Proposed Model Description

The factors of invoices listed earlier were transformed into semantic signs, which we describe below.

A. Railway Station Network

Since railway transportation is largely determined by the railway tracks laid, a large role is assigned to the construction of the graph of railway stations, in which railway stations are considered nodes, and corresponding direct path between two stations will be the edge of the graph. The loop connecting the node to itself (it does not matter whether it is considered directed or undirected) represents the fact that train passes through the station as intermediate station (meaning it is not start or target station). Corresponding weights may be assigned from dislocation events as the median (or mean) duration of traffic on the edge for exact day date.

Now, we formulate the problem of predicting future duration of transportation along the edge as regression of edge weights corresponding to the exact date t based on previous weights for m previous dates $(t-1), \ldots, (t-m)$. For such a task one may use such methods as linear regression methods, kNN, Random Forest or even Time Series analysis if there are enough temporal data.

To determine which edges the train will travel from the departure station to the destination station, one have to calculate the shortest path based on computed and predicted weights for dates when the algorithm estimate the training to process along the corresponding network edge. The only two differences from Dijkstra's shortest path algorithm are that we adapt edge weights based on the time needed for train to reach the current node in the graph using predictive model based on historical data for edge latency, and that we also enforce taking loops weights into account, except for start and target nodes of the route.

The estimated duration of the flight will be determined as the sum of the predicted duration values on the edges corresponding to the shifting arrival time to the intermediate stations. With the help of the graph, one can calculate the duration of the flights between all the stations that are in the graph.

B. Invoice Factor Analysis

However, not only network properties, but also many factors that describe transportation invoice should be taken into account. Below, we describe list of parameters that are usually recorded for Russian railway transportation: Weight (w) of cargo in the train (tons, numerical), Length (d) the shortest path between the departure station and the destination station (km., numerical), Duration of flight (dur, numerical), Type of train (RPS, nominal), and binary indicators of whether the cargo is bulk (sand), liquid, freezing, and whether the route is fixed.

It is easy to think that one would need to remove bias from original duration data and predict not the duration time itself but the difference between the calculated and the observed duration values based on invoice parameters described above without taking into account network structure. One may use linear regression or gradient boosting models as the simplest baselines for this type of trains' invoice data.

C. Baseline Model

In order to have some basic model we could think of very basic idea how to predict the train duration time based on the average daily run. The normalizing values of the average daily run for each type of trains may be defined as a fraction of the shortest distance between the departure station and the destination station, and the average daily run (km.), which in practice is almost twice times greater for "route" flight than "non-route" flight. There is also a method for calculating the duration between the departure station and the destination station based on historical data using time series, but unfortunately, it may be not possible to use it due to lack of data because of low level of digitizing transport services in Russia.

IV. Required Dataset

In order to implement the suggested above models, one needs to construct a directed graph for around four hundreds stations of the West Siberian Railway. The example of core graph is shown at Figure 1.

For prediction of the duration model from the values for the previous days, dislocation data should be collected in the form of available invoice records data: Departure station, Destination station, Flight duration in hrs., Wagon identifier, Date and time of dispatch, Date and time of arrival, Distance in km., Type of the train (RPS), and binary identifiers, such as Is the train loaded? Is the cargo bulk? Is the cargo liquid? Is there route dispatch?

For each numerical data the model need to incorporate date/time attributes as differences in duration and category factors corresponding to exact data and season time affecting overall traffic load.

Each category factor is then binarized into a set of correlated binary factors. The overall model can use either binarized data together with numerical ones, or use interpretable methods such as Decision Trees or Random Forest on binary data, after which apply regression models for predicting exact duration time based only on numerical variables similar to [19].

V. Model Parameters

A. Prediction model for the attributes of edges and vertices of a graph

The data for certain period should be used for training and testing. Predictions of the train flight duration on the edge and

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on January 20,2022 at 10:50:06 UTC from IEEE Xplore. Restrictions apply.

Figure 1. Network representing Russian Railways

the idle time at the station for certain day are to be computed from previous several values of the attributes of the edges and vertices. The linear regression model for attribute prediction may be trained using standard parameters. The kNN method should be used by verifying what is the optimal $k = 1, 2, 3, \ldots$ while preserving low prediction computation time. The random forest regression is to be trained using the MAE loss for understanding the quality metrics measured in hours.

The results of these models should be compared with a simple prediction based on the mean value and the median value for the previous periods. In this study, metrics preserving measurement units were selected based on absolute error $MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid$ or the median absolute error $MedAE = Med(\mid y_1 - \hat{y}_1 \mid, \ldots, \mid y_n - \hat{y}_n \mid)$.

After constructing the attribute prediction model, the model should be evaluated on test dislocation data for following train data date period. Using the predicted values, the calculated durations for invoice data for next period are computed by summing values along the edges and vertices along the shortest path from the station of departure to the destination station. The data on the invoices and the estimated duration predicted for them should be divided into training and testing samples. Using the training sample, regression and gradient boost models may be trained.

The results of testing the attribute prediction model based on testbed artificial data are presented in Figure 2 (in hours). You can see that the linear regression model and the neighbor-based model show better results than the simpler models based on the calculation of the mean/median values and the random forest regression model.

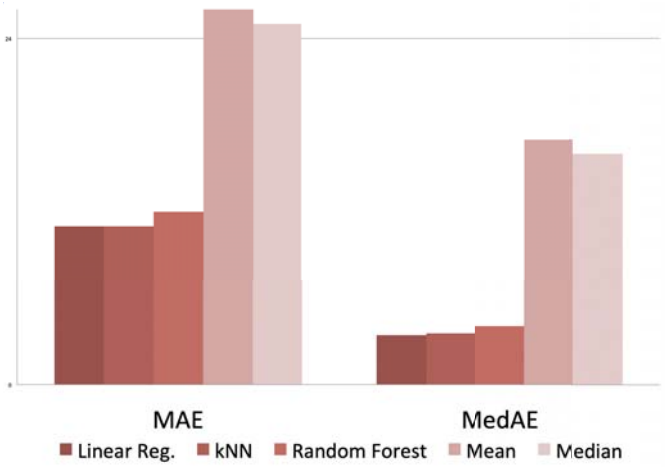

Figure 2. Results of Freight Flight Duration along the Route between Adjacent Railway Stations. It can be seen that machine learning regression models outperformed baselines, such as computer mean and median statistics

B. Network and Factor Models Evaluation

Using the predicted values for edge weights, the lengths of all existing routes may be calculated by finding corresponding

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on January 20,2022 at 10:50:06 UTC from IEEE Xplore. Restrictions apply.

shortest paths. The calculated and actual durations differ significantly for large values of the actual duration, while the smaller duration values are predicted quite well.

It can be noted that, when "routed", the actual value of the duration is often less than the calculated value, because of lower variance of average daily run. This may be also due to the fact that routing does not provide for stopping at some stations from those that were taken into account in the model. Also, it was observed that when applying interviewing persons at railway station that the wagons carrying bulk cargo often reach the destination station earlier than was calculated using the network based models.

Using correlation matrix we deduced that the difference between real and calculated duration was significantly influenced by variables such as the distance between the departure station and the destination station and the type of train. However, the models were tested on all the variables given.

The results of testing the factor model are presented in Figure 3 (in hours, upper line represent RZD report minimum error in 1 day). One can see that the best results are shown by the model based on gradient boosting. In addition, the basic model and model based on network analysis show similar results. This may be due to the fact that the basic model takes into account the routing type, and the model of network analysis is based just on railways network edge and nodes statistics. The coefficients of the trained models showed quite known fact that the variables describing the distance between the departure station and the destination station and the type of train are the mostly correlated with the predicted duration.

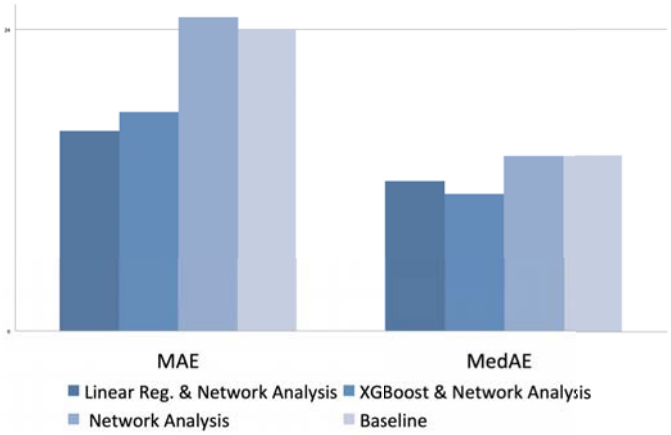

Figure 3. Results of Freight Flight Duration over all the network routes. We showed that network based model is not enough to outperform baseline based on average daily run (due to noisy data and inaccurate edge weights prediction in Figure 2 models results), but adding invoice data helped to achieve sufficient quality of predictions with absolute error less than 20 hours

VI. Conclusion

In the current study the model for forecasting the duration of a railway freight flight was described. The proposed model is based on the construction of stations network and analysis of its attributes: calculation of the duration of transportation on each of the railways between two adjacent stations. We also claim that it is important to consider transportation factors placed in Invoices for freight flights.

To check the quality of models, the corresponding railways graph should be built and data on freight traffic should be collected based on this graph. We computed several experiments using testbed and estimating possible statistics for these parameters, so real-world experiments should be made upon availability of datasets for real duration time for freight flights.

We formulate baseline model based on the normalization coefficients for average daily run and compare its performance to network based and combined network and factor analysis models. It was found that the model based on network analysis, not taking into account the factors describing the transportation, lacks the quality compared baseline due to not taking into account "routing" factor and train type, as well as several others important factors correlated with duration prediction quality. We argue that the proposed factor model together with network descriptors outperform the results of the basic model due to the use of the invoice parameters and should be further improved upon on access to temporal data for improving edge weights predictions.

Finally, it is important to mention that to our knowledge, it is first attempt to improve the freight flight duration prediction in Russia. Nowadays, every freight company use only reference book made by RZD (Russian Railways) with values being only function of railroad distances and not taking into account temporal and station statistics, with accuracy measured in days, while we make it less than one day and can improve to twelve hours for certain type of trains.

Additional Information for review process

Declarations of interest: none. The article was prepared as a proof of concept for the suggested topic while discussing the diploma work of N. Kostyakova. No results are obtained or are a part of RZD or other railways company proprietary ownership.

References

[1] O.-P. Hilmola, "European railway freight transportation and adaptation to demand decline: Efficiency and partial productivity analysis from period of 1980-2003," International Journal of Productivity and Performance Management, vol. 56, no. 3, 2007, pp. 205–225.

[2] A. Romanenko, "Overview of the russian transport sector in 2016," KPMG report, 2017, pp. 1–28 (in Russian). [Online]. Available: https://assets.kpmg.com/content/dam/kpmg/ru/pdf/2017/04/ru-ru-transport-survey.pdf

[3] J. Saranen, B. Szekely, O.-P. Hilmola, and T. Toikka, "Transportation strategy in international supply chains–the case of russia," International Journal of Shipping and Transport Logistics, vol. 2, no. 2, 2010, pp. 168–186.

[4] K. Briggs and C. Beck, "Modelling train delays with q-exponential functions," Physica A: Statistical Mechanics and its Applications, vol. 378, no. 2, 2007, pp. 498–504.

[5] R. M. Goverde, Punctuality of railway operations and timetable stability analysis. Trail, 2005.

[6] I. A. Hansen, R. M. Goverde, and D. J. van der Meer, "Online train delay recognition and running time prediction," in Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on. IEEE, 2010, pp. 1783–1788.

[7] W. Daamen, R. M. Goverde, and I. A. Hansen, "Non-discriminatory automatic registration of knock-on train delays," Networks and Spatial Economics, vol. 9, no. 1, 2009, pp. 47–61.

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on January 20,2022 at 10:50:06 UTC from IEEE Xplore. Restrictions apply.

[8] F. Corman, A. DAriano, A. D. Marra, D. Pacciarelli, and M. Samà, "Integrating train scheduling and delay management in real-time railway traffic control," Transportation Research Part E: Logistics and Transportation Review, vol. 105, 2017, pp. 213–239.

[9] H. Saeedi, B. Behdani, B. Wiegmans, and R. Zuidwijk, "Performance measurement in freight transport systems," SSRN, 2018.

[10] E. R. Kraft, "Scheduling railway freight delivery appointments using a bid price approach," Transportation Research Part A: Policy and Practice, vol. 36, no. 2, 2002, pp. 145–165.

[11] T. Crainic, J.-A. Ferland, and J.-M. Rousseau, "A tactical planning model for rail freight transportation," Transportation science, vol. 18, no. 2, 1984, pp. 165–184.

[12] T. G. Crainic, "Service network design in freight transportation," European Journal of Operational Research, vol. 122, no. 2, 2000, pp. 272–288.

[13] Y. Sun and P. Schonfeld, "Holding decisions for correlated vehicle arrivals at intermodal freight transfer terminals," Transportation Research Part B: Methodological, vol. 90, 2016, pp. 218–240.

[14] M. Dotoli, N. Epicoco, M. Falagario, and G. Cavone, "A timed petri nets model for performance evaluation of intermodal freight transport terminals," IEEE Transactions on Automation Science and Engineering, vol. 13, no. 2, 2016, pp. 842–857.

[15] M.-R. Namazi-Rad, M. Dunbar, H. Ghaderi, and P. Mokhtarian, "Constrained optimization of average arrival time via a probabilistic approach to transport reliability," PloS one, vol. 10, no. 5, 2015, p. e0126137.

[16] E. Demir, Y. Huang, S. Scholts, and T. Van Woensel, "A selected review on the negative externalities of the freight transportation: Modeling and pricing," Transportation research part E: Logistics and transportation review, vol. 77, 2015, pp. 95–114.

[17] S. Belhaj and M. Tagina, "Modeling and prediction of the internet end-to-end delay using recurrent neural networks," Journal of Networks, vol. 4, no. 6, 2009, pp. 528–535.

[18] M. Jacyna, P. Gołebiowski, and M. Krześniak, "Some aspects of heuristic algorithms and their application in decision support tools for freight railway traffic organization," Zeszyty Naukowe. Transport/Politechnika Śląska, 2017.

[19] Q. Zhao, Y. Shi, and L. Hong, "Gb-cent: Gradient boosted categorical embedding and numerical trees," in Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2017, pp. 1311–1319.

Authorized licensed use limited to: University of Electronic Science and Tech of China. Downloaded on January 20,2022 at 10:50:06 UTC from IEEE Xplore. Restrictions apply.