Time Series of Workload on Railway Routes

Zdena Dobesova[1(✉)] and Michal Kucera[1,2]

[1] Department of Geoinformatics, Faculty of Science,
Palacky University, 17. listopadu 50, 779 00 Olomouc, Czech Republic
zdena.dobesova@upol.cz, misakucera@gmail.com
[2] Railway Infrastructure Administration,
Nerudova 773/1, 779 00 Olomouc, Czech Republic

Abstract. The article presents the processing of time series of the workload on railway routes in the Czech Republic. The data for railway stations and signal blocks on routes were processed. The aim is to describe some typical railway stations form the point of structure and workload changes. Both passenger and freight trains are recorded. The descriptive data contains the monthly aggregation of count and weight for passenger and freight trains. Monthly-length correction of data was processed before the evaluation of the time series. Examples of time series for selected stations show that passenger trains are mainly stationary time series otherwise the freight trains are non-stationary time series with a trend. Some stations have a sessional component of series in data about freight trains. In that case, it is possible to predict the time series from old previous data.

Keywords: Time series · Railway workload · Aggregation · Prediction

1 Introduction

The Railway Infrastructure Administration (SŽDC in Czech) is a state organisation that administrates the railway infrastructure in the Czech Republic. The administration collects the amount and weight of passing trains to monitor the workload of railway routes. The monitoring points are situated at mostly all railway stations and important points like signal blocks. The total number is nearly 3 350 monitoring points.

The Railway Infrastructure Administration needs this monitoring for evaluation of workload on rail routes, for economic assessment and technical maintenance. The data are collected primarily for the need of rental payments by individual carriers. Moreover, that data could be used as an important source for the calculation of the noise pollution of the population [1] and other tasks. The number and weights of the passing trains affect the operating load and technical condition of routes.

The presented investigation tries to describe the measured data from the point of time. The analysis of time series was the base idea. Firstly, selected stations were compared: the station located on the main railway route (transit corridor) and station on a regional route. Subsequently, some selected station were analysed to find if some seasonal effect exists during the year or if some unexpected deviations exist.

R. Silhavy (Ed.): CSOC 2019, AISC 985, pp. 370–380, 2019.
https://doi.org/10.1007/978-3-030-19810-7_37

2 Data and Methods

The Railway Infrastructure Administration provided the data in the followed structure for presented research. The attribute data contains this structure: time, the name of the station, the indicator of the type of train (passenger, freight, maintenance etc.), the number of trains, number of railway cars in train set, the average length of the train set, and the average number of axels. Only two descriptive attributes were taken for investigation – the number of trains and the weight of the train for selected measuring points. The monitoring data are accessible in monthly aggregation for each measuring point. The monitored time was three years: 2016, 2017 and 2018.

2.1 Calendar Effect and Monthly-Length Adjustments of Time Series

The data are affected by calendar influences. The are several calendar influences like different lengths of the month, different numbers of the weekend in a month, different working days in a month and movable feasts and state holidays in a year. This irregularity could have surprising consequences. It is necessary to clean the data before further processing. In the case of monthly aggregated data, the influence of different length of the month was discovered in the railroad workload. It is called monthly-length effect [2]. Especially, month February, as a shortest in length, is visible in the data (Fig. 1). There are declinations (minimum values) of the weight of train set in each February 2016, 2017 and 2018 in comparison with other months in the year.

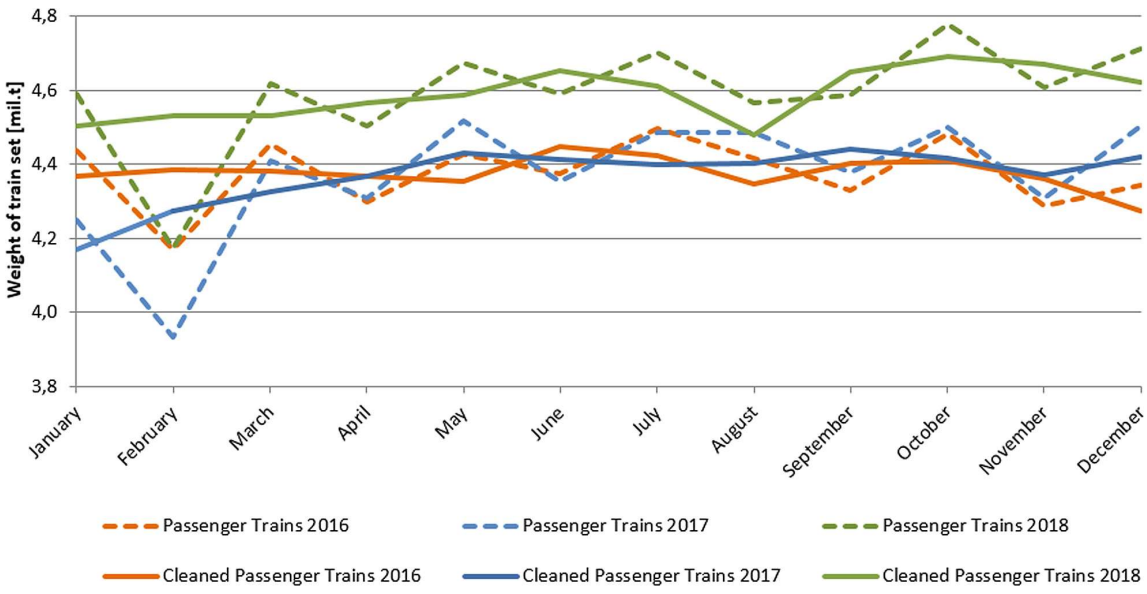

Fig. 1. Time series with origin data and cleaned data for Česká Třebová station

There are two possible solutions for recalculation and clearing data [3]. The first is recalculation to year with 360 days with constant length of month 30 days. The second

solution is recalculation for the punctual length of the year, and it is 365 or 366 days. The equation for recalculation values according to the real length of each month are [4]:

$$Ynew_t = \frac{Y_t}{N_t} * \overline{N}, \tag{1}$$

where $Ynew_t$ is a new value in a month t, Y_t is origin value in a month t, N_t is a count of days in a month t and N_t is the average length of the month in the year. In the year 2016 the average length of the month was 30.5 days (leap-year) and 30.41667 days in the year 2017 and 2018. All origin values in time series were recalculated according to Eq. (1). The monthly-length effect was removed from data. The differences of values are visible at Fig. 1 where both origin values (dot lines) and cleaned values (solid lines) are depicted. The Česká Třebová station was taken for a demonstration of time series in 2016, 2017 and 2018. The values are the total weight of the passenger train set in each month. The time series of cleaned values are more smoothed than origin values. Origin data contain more oscillation caused by the different length of the months.

2.2 Steps of Data Processing

After cleaning data, the data of various stations were displayed in graphs to investigate time series. The first step was a visual analysis. Visual analysis is the recommended initial step in analyzing the time series [5]. The first was comparing the number of passenger trains and freight trains. Subsequently, the weight of the train sets was compared. Some big differences exist. The count of passenger train set is many times higher than freight trains in most stations. However, some exception exists. We select some typical stations on the transit corridor, on regional routes and stations near border crossings.

The visual analysis discovered the variance of values in years and months. We try to find the trend and cycles in the selected time series. Because some railway stations are near the limit of capacity [6], there is not possible to expect a strong rise in the number of trains set. To decompose the time series and detect the trend we used the moving average with a 12-month window [3]. Microsoft Excel 2016 was used for data processing. Also, the graphs were prepared in this software.

In minor cases, the time series with seasonal cycles were discovered for freight trains. They are influences of industrial production in some localities. In the case of Rýmařov station, we try to predict the weight of the train set for the 2019 year. For prediction, the software WEKA v. 3.8 with the implementation of Holt-Winters method [7] was used. The selected examples of railway stations and the description of time series are shown in the next section.

3 Results

3.1 Comparison of Passenger and Freight Trains

The count and weight of train sets depend on the type and locality of the railway station. To imagine differences, the short overview is in Table 1. There are monthly average values of weight and count of train sets separately for passengers and freight

trains in 2016. The Česká Třebová is a station located on the transit corridor from Prague to Ostrava. The frequency and amount of traffic are very high. The count of passenger trains is nearly seven times higher than freight trains. Despite that, the average weight of passenger train sets is only two times higher. The next examples show that in all case the average count of passenger trains is higher than the count of freight trains. The average weight of passenger trains is also higher than the weight of the freight train set, but the difference is not so high than in the case of a count of trains. It is evident that freight transport is much more massive.

Table 1. Comparison of average count and weight of the passenger and freight trains in 2016.

Station	Type and location	Weight of passenger trains [t]	Weight of freight trains [t]	Count of passenger trains	Count of freight trains
Česká Třebová	Transit corridor	4 377 056	2 287 381	13 489	2 084
Břeclav	Border crossing	2 199 604	5 953 243	6 614	5 194
Cheb	Border crossing	836 381	577 939	3 892	1 155
Rýmařov	Regional route	12 487	11 724	518	34
Senice na Hané	Regional route	15 747	445	560	2

One exception was presented in Table 1, and it is the border crossing Břeclav. There is the weight of freight trains higher than the weight of passenger trains (two times higher). The reason is transport to neighbour countries Slovakia and Austria. Also, the count of freight trains in Břeclav is higher in comparison with Česká Třebová station. The second presented a border crossing is Cheb station in Table 1. The dominance of average weight and count of freight trains is the same as other domestic stations. It is evident that the transport of freight is high to neighbouring Germany.

Table 1 presents two stations on a regional route; there are Rýmařov and Senice na Hané. The numbers are lower in comparison to previous stations located on main routes. Both stations have a low amount of workload. In the case of terminal station Rýmařov, the average monthly weight of passenger trains and freight trains is nearly the same (around 10% difference). It is evident that there is relatively huge transportation of freight. It was an impulse to investigate the changes in transport in that station.

For all presented stations the graphs of time series were constructed. The example of Česká Třebová station is presented in Fig. 2. There are visible the same high difference between passenger trains and freight trains in 2016, 2017 and 2018 years.

The fluctuation of the weight and count is not high in the category of passenger trains. More fluctuations are in weight in the category of freight trains. The question was if the workload is stationary or non-stationary time series in Česká Třebová station. The time series is stationary if it does not have a trend or seasonal effects, in addition the average is the same in selected periods [4]. The stationarity was checked by

calculation the mean and variance for each year. Comparison of the means and variance showed that the times series are stationary in case of freight trains. Moreover, according to recommendation [8], the histograms were constructed for all six times series (weight and count for passenger and freight trains). The histograms have the bell curve-like shape of the Gaussian distribution for freight trains.

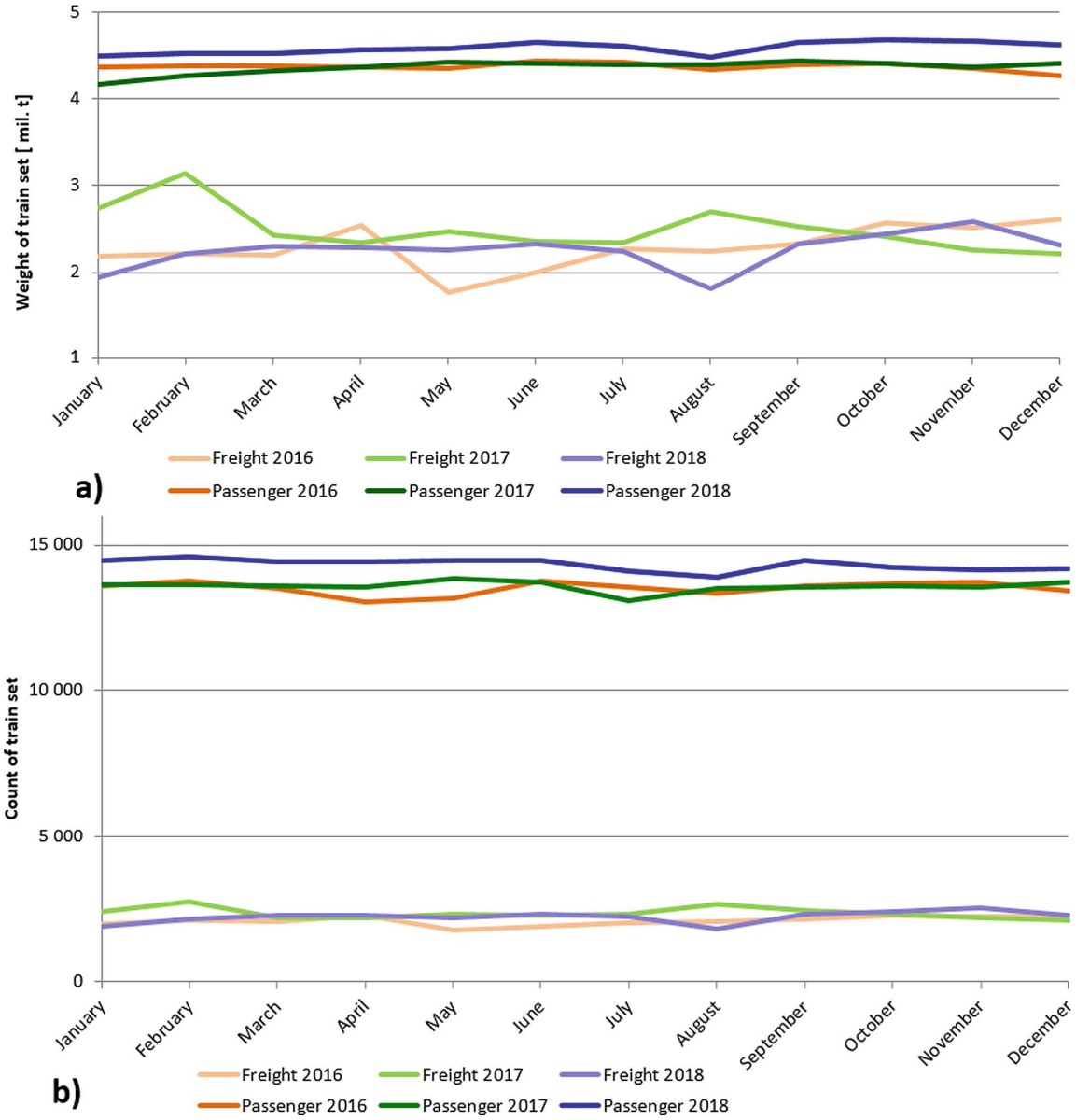

Fig. 2. Time series of weight (a) and count (b) of the trains in Česká Třebová station

The time series are stationary in case of the freight trains. It means that the workload in this station on the corridor has no trend and seasonal effect for a freight train. The time series of passenger train weight has a small increasing trend, especially influenced by data in 2018. So weight and count passenger trains are non-stationary time series in Česká Třebová station.

The Břeclav station is mentioned in Table 1 and the previous text. It is an example of the station where is an inverse portion of the passenger to freight trains from that majority of stations in the Czech Republic. Figure 3 shows the time series of train weight for Břeclav station. The weight of freight trains (light colours) is higher than the weight of passenger trains (dark colours). The oscillation of the freight train weight is higher than oscillation of passenger train weight (like Česká Třebová station). The passenger train weight is nearly stationary time series. The time series of freight trains is non-stationary time series. The decomposition of the time series for freight trains is presented in the next Sect. 3.2.

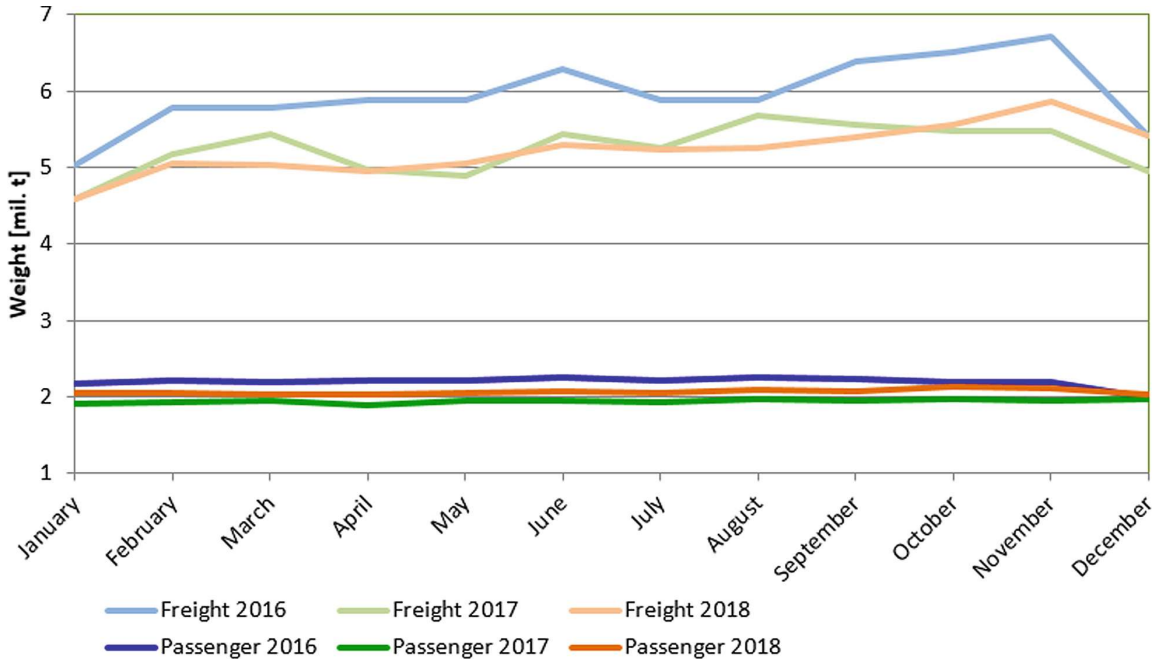

Fig. 3. Time series of the weight of train sets in Břeclav station in three years

3.2 Decomposition and Seasonal Part in Time Series

Time series data can exhibit a variety of patterns, and it is often helpful to split (decompose) a time series into several components, each representing an underlying pattern category. Three types of time series patterns exist - trend, seasonality and cycles [5, 7].

If an additive decomposition is assumed then it is described by Eq. (2):

$$y_t = S_t + T_t + R_t, \tag{2}$$

where y_t is the data, S_t is the seasonal component, T_t is the trend-cycle component, and R_t is the remainder component, all at period t.

Alternatively, a mõultiplicative decomposition is described by Eq. (3):

$$y_t = S_t \times T_t \times R_t \tag{3}$$

The additive decomposition is the most appropriate if the magnitude of the seasonal fluctuations, or the variation around the trend-cycle, does not vary with the level of the time series. When the variation in the seasonal pattern, or the variation around the trend-cycle, appears to be proportional to the level of the trend of time series, then a multiplicative decomposition is more appropriate.

After the visual analysing and the counting of variations, the additive model is more appropriate for time series of investigated workload on railway routes. Namely, the time series of weight are non-stationary on the contrary of stationary time series of train count (both passenger and freight train). The trend has various fluctuation in three-year time series. Only some time series have an evident seasonal part. The remainder component is mostly present in the time series. Two examples of interesting decompositions of time series are presented: Břeclav and Rýmařov stations.

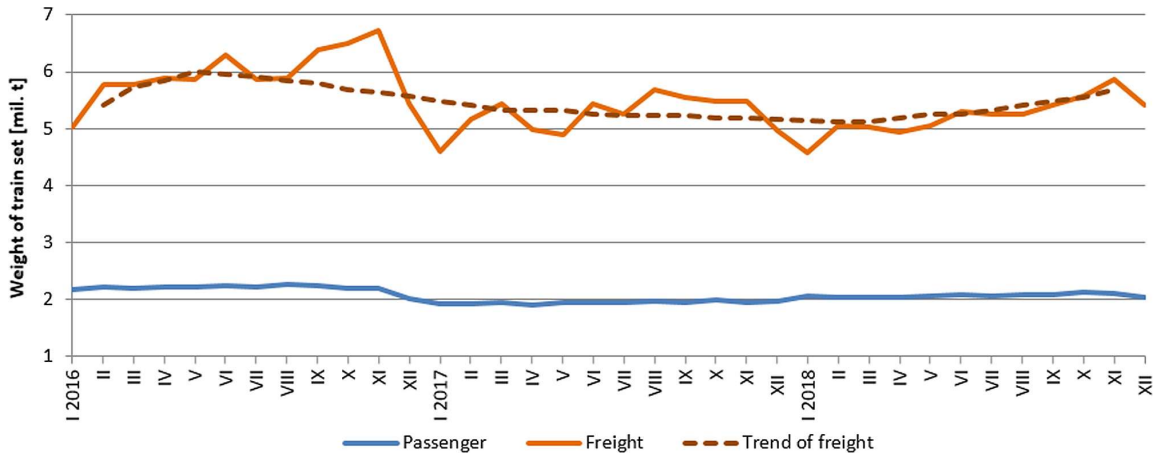

Fig. 4. The trend in time series of weight for freight trains in Břeclav station

To decompose the time series the classical method of moving averages was used. For smoothing data with expected seasonal component, the period for smoothing is recommended to the length equal the length of the season. In the case of monthly aggregated data, the length is 12 [3]. Figure 4 is a graph where the trend calculated by a central moving average by this equation:

$$T_t = \frac{y_{t-6} + 2 * \left(y_{t-5} + y_{t-4} + y_{t-3} + y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2} + y_{t+3} + y_{t+4} + y_{t+5}\right) + y_{t+6}}{24} \tag{4}$$

The equation was applied for time t_7 to t_{29} for Břeclav station. The start and end of time series are calculated by shorter moving averages (3, 5 and 7 element windows). The weight of freight train has a stable declining trend in 2017, and increasing trend in 2018 with remainder component R. There is no repetition of the seasonal part except a regular decrease in December and January for each year.

The interesting decomposition is for Rýmařov station. In Fig. 5 is a graph of weight both for the freight (light colours) and passenger trains (dark colours). The weight of passenger trains is nearly stationary series. Otherwise, freight trains have a strong

seasonal part. The lower amount of weight is in the winter. The trend increases in the spring with the highest value in summer (from May to September) followed by decreasing in the autumn. The series in 2016 and 2017 are more similar during the year than time series in 2018. The weight of freight trains in winter is also high in 2018; there is not so big difference in winter-spring time and summertime in 2018. All presented time series have a decrease in December that is probably caused by Christmas holidays and non-working days.

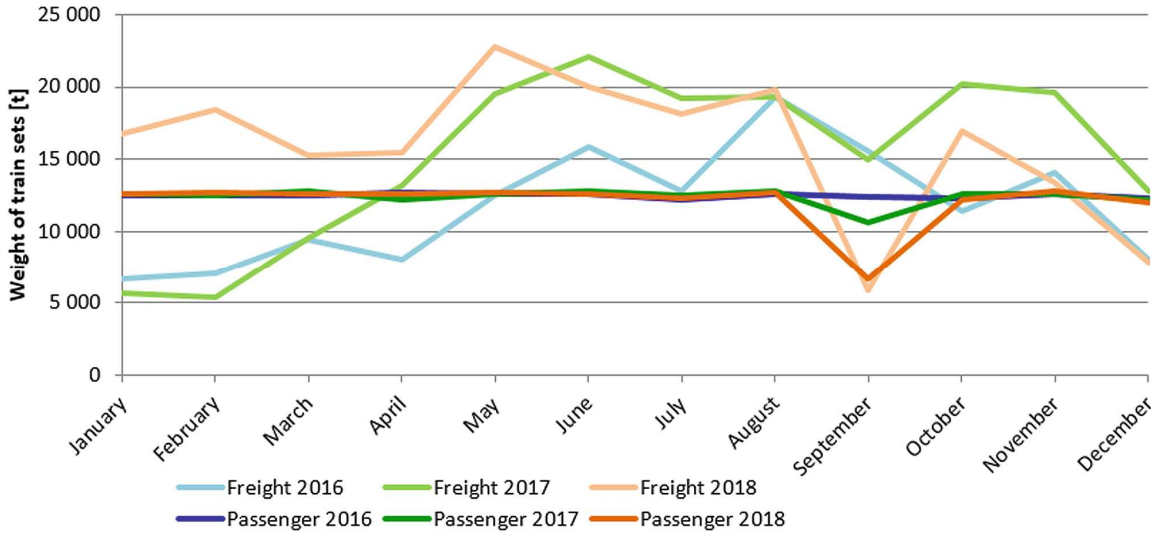

Fig. 5. Time series for Rýmařov station with a seasonal component

The interpretation of the strong seasonal part is followed. The railway station Rýmařov is a service station for the large wooded area. It is located in the Jeseníky Mountains in the Bruntál District, Moravian-Silesian Region. The surrounding area is typical for high activities in logging, cutting and preparing the timber. Rýmařov station is a terminal of the regional railway route. The timber industry is a reason for this seasonal component of a time series of weight. Because the count of freight trains has not this noticeable seasonal cycle, the length of train sets must be much longer in summer than in winter season. The weight of train sets in the year 2018 has noticeably hight values during all year. It is probably caused by high logging and cutting timber after bark beetle calamity and windfallen of trees.

Moreover, there is visible declination in September 2018. It could be caused by a temporary reconstruction of the route. It is supported by declination both freight and passenger trains. This declination was also found in the count of freight and passenger trains. The data about the count of trains verified the hypothesis about the reconstruction of the railway route.

3.3 Identification of Unusual Observations

Time series could also reveal some extraordinary situation like some reconstruction on the rail route. Figure 6 shows the time series for Senice na Hané station. It is a regional route with a small number of trains. Especially, freight trains have some zero values of a count in some months.

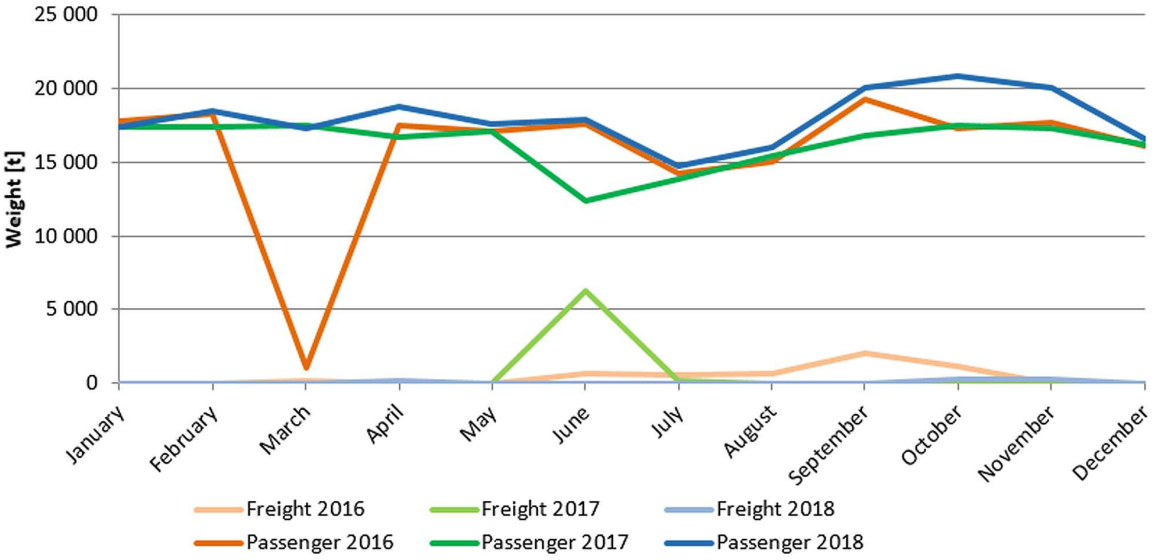

Fig. 6. Time series for Senice na Hané station with the identification of an outlier

The high decline of passenger trains is in spring of the 2016 year. In comparison with the situation in spring 2017, it is an unexpected situation. There are two possible reasons. The first reason could be wrong source data. The second could be a temporary reconstruction of the route. The second is valid for Senice na Hané station in March. Visual analysis of time series is a very illustrative way to reveal some unusual situation or outliers in data. The automatic identification of unusual observation could be a hint for some extraordinary situation. The exclusion of this data is necessary before finding a seasonal component. Otherwise, the results of decomposition could be misleading. Besides the visual analysis of time series, a boxplot and statistics are other options how to find any outliers or unusual observation in data.

3.4 Predictions of Train Weight

On the base of old data, it is possible to predict data for the future. The software WEKA v. 3.8. was used for prediction. WEKA is a collection of machine learning algorithms for data mining tasks that is freely accessible for use. The additional package *time series forecasting* by Mark Hall was installed into WEKA [9]. The package implements the Holt-Winters triple exponential smoothing method for prediction. The Holt-Winters method, implemented in that package, needs minimally three-year long time series to predict time series for 12 months in future.

The experimented time series for Rýmařov station had a strong seasonal part in spring and summer when the weight of freight train increased caused by the timber industry. The interpretation of data is depicted in the chapter about the decomposition of time series. The input parameters for Holt-Winter method were set like these - value smoothing factor 0.1, trend smoothing factor 0.2 and seasonal smoothing factor 0.1 for data from Rýmařov station. Figure 7 presents the fluctuation of time series from 2016 to 2018. Last part of the graph is a prediction for 12 months - the year 2019 (dot line). The prediction also predicts the seasonal increase in weight in summer time (Fig. 7). The prediction is wrongly influenced by decreasing in September 2018 due to the

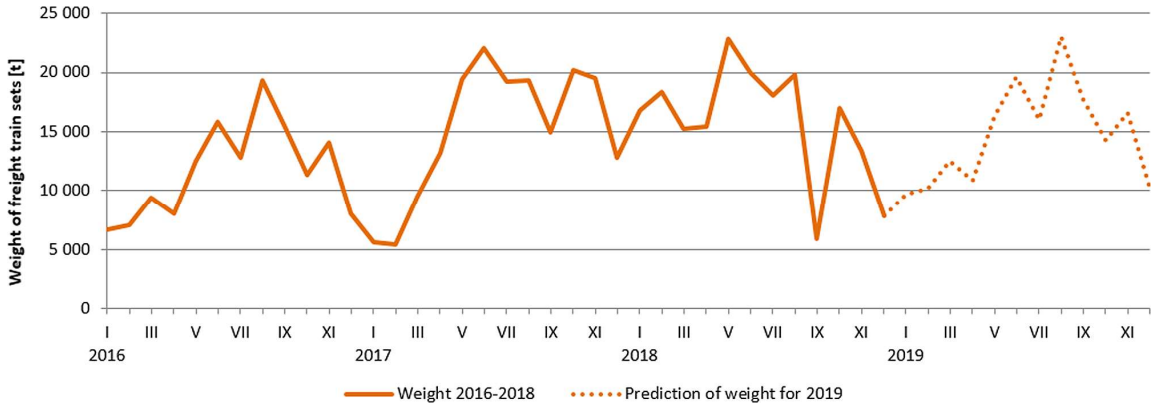

Fig. 7. Prediction of weight for freight trains at Rýmařov station for 2019

reconstruction of the route. Also, the data is influenced by the hight amount of cutting timber after calamity in 2018. The longer and older data in time series would be more precise in the prediction for the future.

4 Conclusion

The paper presents time series analyzes as the first analysis of the huge data set of the railway infrastructure measuring in the Czech Republic. Before evaluating the monthly time series, it is necessary to clear the data from the monthly-length effect as shown in the time series for Česká Třebová station. The presented analyses show that greater seasonal fluctuations during the year can be expected in freight transport than in passenger transport, which is given by valid timetable. Conversely, freight rail transport depends on customer orders for shipment. An analysis in a more detailed time series (for a day, a week, days of month) cannot be performed because the data is aggregated and provided only in monthly summaries.

It is already apparent from the examples that the stations differ in the number and weight of the trains during the year. Therefore, we plan to concentrate on analyzing the data at the border crossings to all neighbouring countries of the Czech Republic. The article presents the time series of Břeclav station near the border crossing, which is passed by freight trains to the neighbouring countries Slovakia and Austria.

It is planned to continue by further analysis of the data. One of the planned tasks is a classification of stations according to the annual load by number and weight both passenger and freight trains. Classification means to find all stations with similar workloads thus belong to one class (group) of similar stations. It is a task of clustering of time series [10].

It will also be interesting to compare the predictions made for 2019 with actual measured values in future. In order to evaluate and interpret variations and trends in time series, it will be necessary to continue consultation with rail experts to interpret data better. Data sometimes are influenced by temporary reconstruction of routes performed by a decrease of number and weight of train sets.

The presented time series analysis will also be used as examples of real-world data from the practice for the teaching at the study branch Geoinformatics at the Palacký University in Olomouc. The author of the article has a positive experience with the application of the knowledge that the teacher has acquired in solving practical problems. This practical knowledge could be added in the syllabus. This experience is mentioned in the article about the relational database design for the botanical garden plant database (BotanGIS project) [11]. Time series analysing is lectured in Data Mining for Geoinformatics. Students are familiar with the software WEKA for data mining and tasks. As a result, next year will be added some practical examples into the syllabus of subject Data Mining.

Another planned research is to compare rail workload with the state of surface and rail wear. SŽDC measures the technical condition (wear and tear) with a special moving measuring vehicle. These data are further used to evaluate rail damage and maintenance planning. The data obtained by technical vehicle could strongly relate with the weight of trains monitored in point stations. This type of research could be valuable for discovering the dependencies in data and prediction of technical maintenance.

Acknowledgement. This article has been created with the support of the student project IGA_PrF_2019_014 of the Palacky University Olomouc.

References

1. SŽDC: Annual Report 2017. SŽDC (2018)
2. Cleveland, W.S., Devlin, S.J.: Calendar effects in monthly time series: modeling and adjustment. J. Am. Stat. Assoc. **77**, 520–528 (1982)
3. Litschmannová, M.: Introduction to the time series analysis. VŠB-TU, Faculty of Electrical Engineering, Department of Applied Mathematics, Ostrava (2010)
4. Hančlová, J., Tvrdý, L.: Introduction to the time series analysis. VŠB-TU, Ostrava (2003)
5. Křivý, I.: Analysis of Time Series. University of Ostrava, Ostrava (2012)
6. Cestr, A.: Railway infrastructure for serviceability of Ustecky Region. In: Conference Železniční dopravní cesta. SŽDC, Ústí nad Labem (2018)
7. Hyndman, R.J.: Forecasting: Principles and Practice. Monash University, Australia (2018)
8. How to Check if Time Series Data is Stationary with Python. https://machinelearningmastery.com/time-series-data-stationary-python/. Accessed 15 Jan 2019
9. Hall, M.: Time Series Analysis and Forecasting with Weka. http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka. Accessed 10 Jan 2019
10. Samé, A., Chamroukhi, F., et al.: Model-based clustering and segmentation of time series with changes in regime. Adv. Data Anal. Classif. **5**, 301–321 (2011)
11. Dobesova, Z.: Teaching database systems using a practical example. Earth Sci. Inform. **9**, 215–224 (2016)