20th International Scientific Conference AMSE
Applications of Mathematics and Statistics in Economics 2017
Szklarska Poręba, Poland 30 August 2017 – 3 September 2017

MODELING THE NUMBER OF ROAD ACCIDENTS

PROCHÁZKA JIŘÍ, FLIMMEL SAMUEL, ČAMAJ MATĚJ, BAŠTA MILAN
University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability, Nám. W. Churchilla 4, Prague, Czech Republic
email: xproj16@vse.cz, samuel.flimmel@vse.cz, matej.camaj@vse.cz, milan.basta@vse.cz

Abstract

Modelling the number of daily road accidents can be beneficial not only for insurance companies but also for other institutions such as the national road administration, national insurers' bureau etc. Accurate predictions of the number of road accidents could be beneficial in terms of efficient liquidation planning, improving the reserving processes, streamlining the capital allocation and road maintaining. Consequently, it is relevant to build a viable model for predicting the number of daily road accidents. One of the most important parts of the model is the model of daily seasonality. Since this seasonality exhibits a long seasonal period, approaches based on basis expansion can be used for its modelling. We also investigate the multiple seasonality pattern and specific time events which could potentially affect the number of accidents. Furthermore, the impact of different external variables, such as the average daily temperature, rainfall and other factors influencing human driving skills, will also be investigated.

Key words: *road accidents, long seasonality modeling, multiple seasonality*

JEL Codes: *C53, G22*

DOI: 10.15611/amse.2017.20.29

1. Introduction

The main aim of this paper is to model the number of daily road accidents caused by uninsured drivers, capture the relationship between the external variables and the number of road accidents and create a prediction model that could be beneficial for lots of national institutions such as the national road administration, national insurers' bureau etc. Data used for the analysis are provided by the Czech insurers' bureau. Raw data can be found at http://actedu.vse.cz/wp-content/uploads/2016/03/data_nfvp.csv. Specifically, the daily number of road accidents caused by uninsured drivers in the period 2007 – 2011 is analyzed. To ensure consistency of the data, leap days have been removed from the data set. Data were also investigated in Procházka (2017).

In the following text a time series $\{X_t\}$ of length N, and with seasonality of length L will be considered. Further, an additive decomposition will be assumed, $\{X_t\}$ being decomposed as follows

$$X_t = S_t + B_t + E_t \qquad t = 1,\ldots,N, \tag{1}$$

where $\{S_t : t = 1,\ldots,N\}$ represents a deterministic seasonal component, $\{B_t : t = 1,\ldots,N\}$ represents other deterministic components such as deterministic trend, external variable etc. and $\{E_t : t = 1,\ldots,N\}$ is a stationary ARMA(p,q) process, where p is the order of the AR part and q is the order of the MA part of the process. Specifically E_t is given as

$$E_t = \phi_1 E_{t-1} + \ldots + \phi_p E_{t-p} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t \quad t = 1,\ldots,N, \tag{2}$$

where $\phi_i : i = 1, \dots p$ and $\theta_j : j = 1, \dots q$ are parameters that must be estimated and $\{e_t : \quad t = 1, \dots, N\}$ is a Gaussian white noise (see Hyndman and Athanasopoulos, 2014). Because of the character of the time series a distribution different from the normal distribution should be preferably used as the distribution of the stochastic part of E_t. However for simplicity we will assume normal distribution as an approximation. The seasonal component, S_t, represents seasonal fluctuations caused by varying traffic exposure. B_t represents long term trend and other external variables such as the average rainfall, unexpected extreme weather conditions such as frost etc. The last component, E_t, represents the residual randomness of the time series. Because of the character of the data additive decomposition is assumed in Equation (1). The reason for choosing an additive model is also visible in Figure 1.

2. Seasonal part of the model

To capture the seasonal part of the model we used the multiple seasonality model. To be more specific, we considered seasonality due to changing seasons of the year, i.e. seasonality with the length equal to 365 ($L = 365$), and also seasonality due to different days of the week, i.e. seasonality with the length the equal to 7 ($L = 7$). To represent the seasonal part with $L = 365$ we use Fourier representations (e.g Ramsay and Silverman, 2002; Gould et al., 2008; Procházka et al., 2016), whereas the seasonal part with $L = 7$ will be represented using dummy (indicator) variables. As a result, the seasonal component $\{S_t : t = 1, \dots, N\}$ will be decomposed into two parts

$$S_t = S_{1,t} + S_{2,t} \qquad t = 1, \dots, N, \tag{3}$$

where $\{S_{1,t} : t = 1, \dots, N\}$ represents seasonality with the length equal to 365 and $\{S_{2,t} : t = 1, \dots, N\}$ represents seasonality with the length equal to 7. $S_{1,t}$ can be written as

$$S_{1,t} = \sum_{k=1}^{K} \alpha_k \sin\left(2\pi \frac{k}{365} t\right) + \sum_{k=1}^{K} \beta_k \cos\left(2\pi \frac{k}{365} t\right), \qquad t = 1, \dots, N, \tag{4}$$

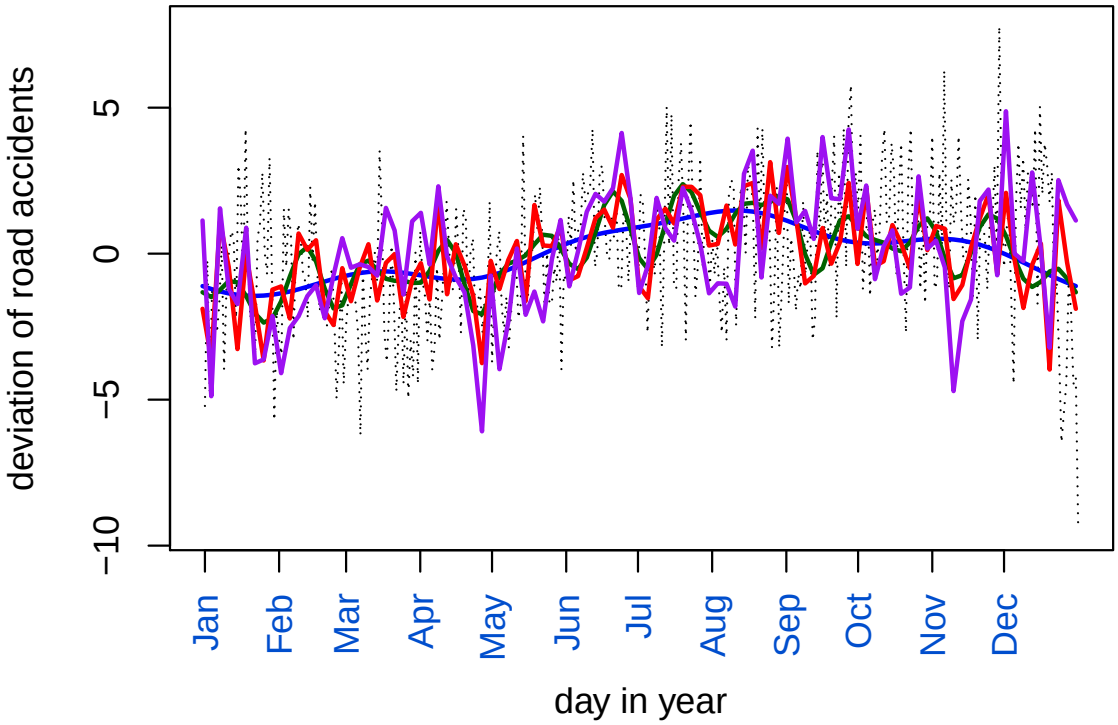
where α_k and β_k for $k = 1 \dots, K$ are parameters that need to be estimated. K can potentially take any value from 1 to 182 (which is effectively equivalent to the situation where 364 dummy variables would be used to represent $S_{1,t}$). In the final model we will not use $K = 182$; instead we will use $K << 182$, in order to represent the seasonal pattern using a smaller number of parameters. $S_{2,t}$ can be written as

$$S_{2,t} = \sum_{m=2}^{7} \psi_m D_{m,t}, \tag{5}$$

where ψ_m for $m = 2, \dots, 7$ are parameters that needs to be estimated and $D_{m,t}$ are dummy (indicator) variables which are equal to one if t corresponds to the m-th day of the week, and equal to zero otherwise.

As we have already mentioned, to get a good final model for $S_{1,t}$, it is crucial to choose the right value of K . A low value of K may not lead to a good approximation of the seasonal cycle; on the other hand, a large value of K may cause the estimated model to be highly variable. In Figure 1 you can see the representation of $S_{1,t}$ using different numbers of basis functions.

Figure 1: The average daily deviation (meaning the difference of daily observations from average of time series) of the number of road accidents is shown as a dotted line. The blue line is an estimate of $S_{1,t}$ (for a period of length 365 from January through December) for $K = 5$. The green line is an estimate of $S_{1,t}$ for $K = 20$, the red line an estimate for $K = 50$ and the purple line for $K = 100$.



Source: Own elaboration.

As you can discern from Figure 1, a large value of K (and thus a large number of basis functions) leads to a more precise in sample fit, but at the same time implies a large number of parameters and hence the estimated model is highly variable. Consequently, we must be very cautious while selecting K. There are several possibilities which can be helpful in this aspect. In our case, we will choose K according to the information criterion, namely the Bayesian Information Criterion (also called Schwarz criterion), abbreviated as BIC. BIC is defined as

$$\mathrm{BIC} = \log(N)k - 2 * \hat{l}, \tag{6}$$

where N is the number of observations, k the number of parameters and $\hat{l}$ is the natural logarithm of the maximized likelihood function. We prefer BIC to AIC (Akaike Information Criterion), because BIC penalizes the number of parameters more strongly in our situation.

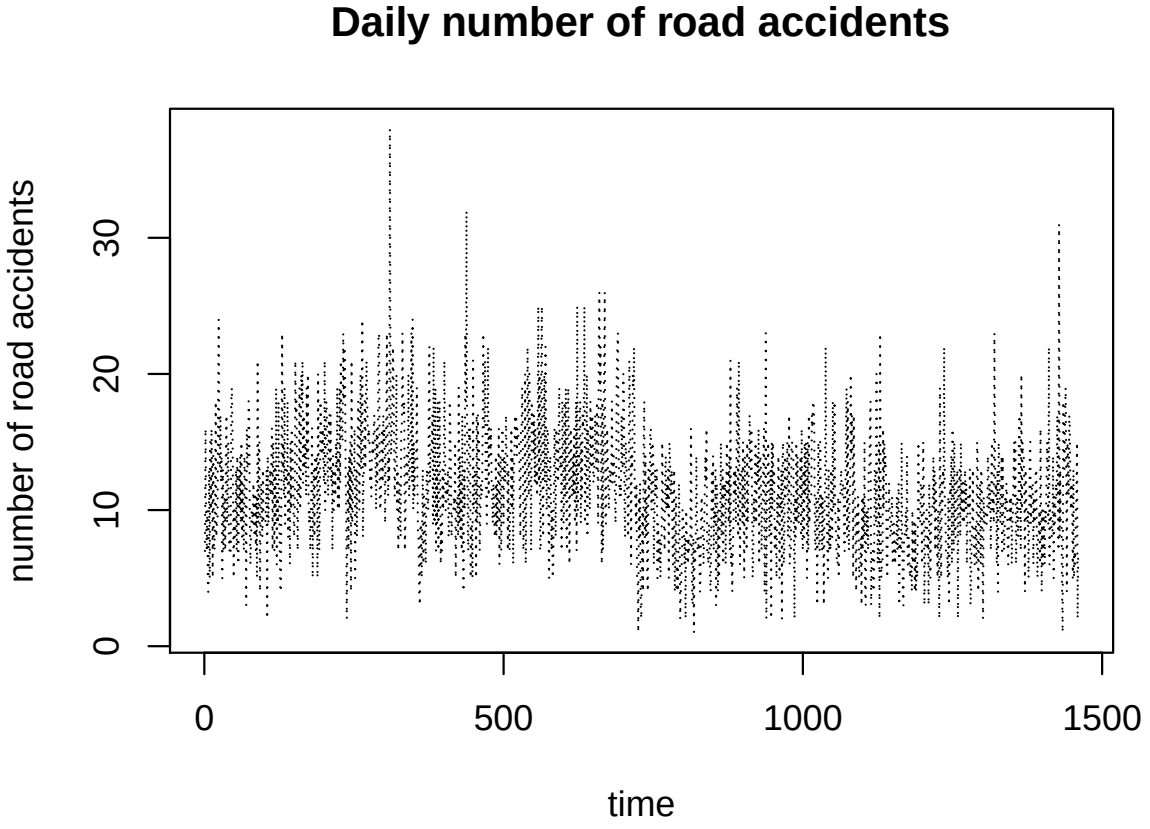
3. Other deterministic parts of model

The deterministic component, $\{B_t : t = 1,\dots,N\}$ contains a deterministic trend and climate variables such as the average rainfall and frost. We will also include the effect of events which could be unexpected for drivers such as the first frost or frost in non-winter months. $\{B_t\}$ can be written as

$$B_t = T_t + O_t, \tag{7}$$

where $\{T_t : t = 1,\dots,N\}$ is the trend part and $\{O_t : t = 1,\dots,N\}$ is the part associated with the other regressors (climate variables and unexpected events).

Figure 2: Dotted line represents number of daily road accidents.



Source: Own elaboration.

In Figure 2 there is a visible moderate parabolic trend that holds approximately for the first 810 observations. Afterwards, there is a visible break in the trend after which the trend is more or less constant. We will assume the following model for the trend part

$$T_t = \beta_0 + (\beta_1 + \beta_2 t + \beta_3 t^2) * \delta_t \qquad t = 1,\dots,N, \tag{8}$$

where β_0, β_1, β_2 and β_3 are parameters that need to be estimated and δ_t is dummy variable

defined as

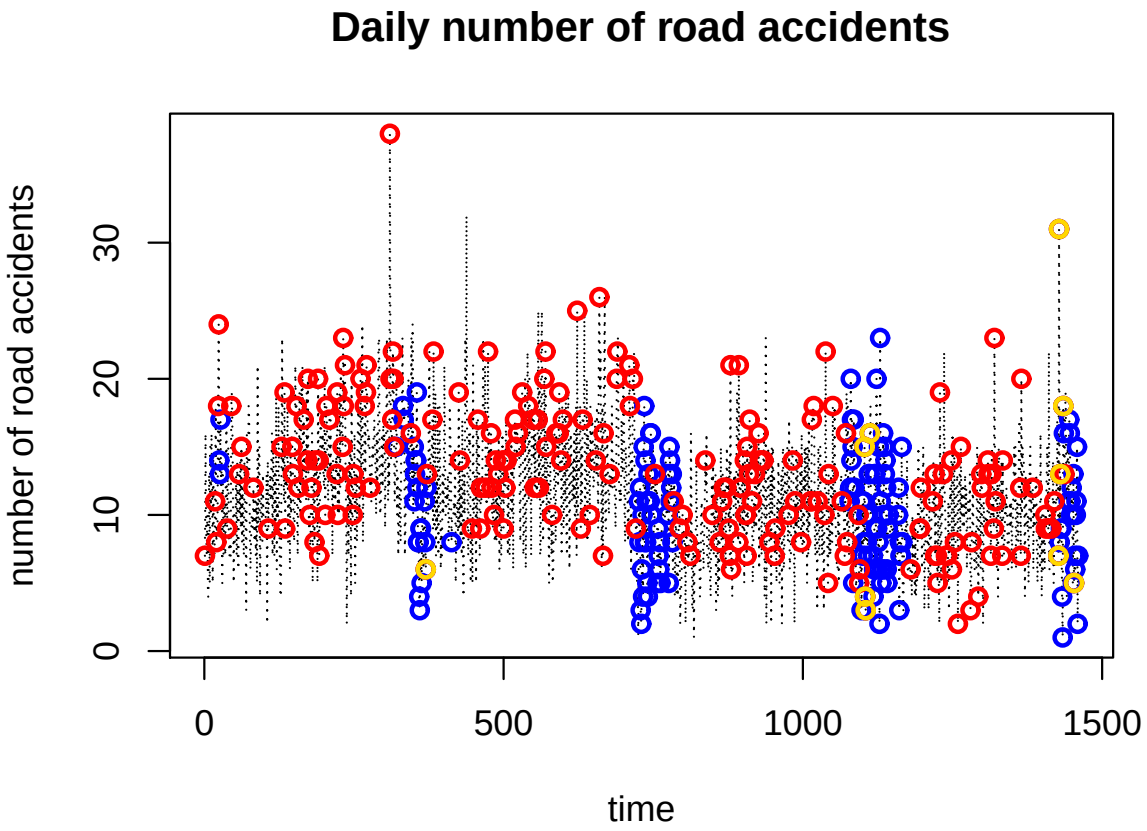
$$\delta_t = \begin{cases} 1 & t < 810 \\ 0 & t \geq 810 \end{cases}. \tag{9}$$

O_t of Equation (7), will be given as follows

$$O_t = \gamma_0 + \gamma_1 F_t + \gamma_2 R_t + \gamma_3 FR_t + \gamma_4 UE_t, \tag{10}$$

where $\{F_t : t = 1,\ldots,N\}$ indicates frost and is equal to 1 if the average daily temperature is below zero at time t, $t-1$ and $t-2$, $\{R_t : t = 1,\ldots,N\}$ indicates increased rainfall and is equal to 1 if the average daily rainfall exceeds 2.6 mm/hour, $\{FR_t : t = 1,\ldots,N\}$ indicates combination of increased rainfall and frost and $\{UE_t : t = 1,\ldots,N\}$ indicates unexpected events such as first frost etc. The time series of daily road accident with marked indicators of frost, rainfall and their combination can be seen in Figure 3.

Figure 3: Black dotted line represents number of daily road accidents. Blue dots represents the frost, red dots represents increased rainfall and gold dots represents combination of frost and increased rainfall.



Source: Own elaboration.

4. Building of prediction model

Putting all parts of the model together we must determine: K (see Equation (4)) and the orders p and q in E_t (see Equation (2)). For the determination of the orders p and q, we will use an automated procedure auto.arima from the forecast package (Hyndman and Khandakar, 2008). In order to determine K, we will use the BIC criterion. A minimal value of K equal to 5 will be assumed. We will build three models. The first model will contain the whole seasonal part S_t, the error term E_t, the deterministic trend T_t and indicator variables F_t, R_t and FR_t. The second model will contain S_t, E_t and T_t (and will thus not contain the other external variables). The third model will contain the same variables as the first model plus a variable indicating a drop of the temperature below zero in November, and the average monthly search volume of word "doprava"(traffic) on Google.

Table 1: First model: Bayesian information criterion for various values of K.

Number of basis (K)	Bayesian information criterion
$K = 5$	8332.632
$K = 6$	8346.251
$K = 7$	8359.252
$K = 8$	8372.738
$K = 9$	8385.655
$K = 10$	8395.332
$K = 20$	8498.722
$K = 10$	8619.361

Source: Own elaboration.

The results for the first model are presented in Table 1. With an increasing number of basis functions (K), BIC is increasing. As a result, we will set $K = 5$, which corresponds to 10 basis functions of Equation (4). Further, using the auto.arima procedure, ARMA($p = 1$,$q = 2$) is selected as an optimal model for E_t. After fitting of the model on the full time series we will compare our fitted values with the original data (see Figure 4).

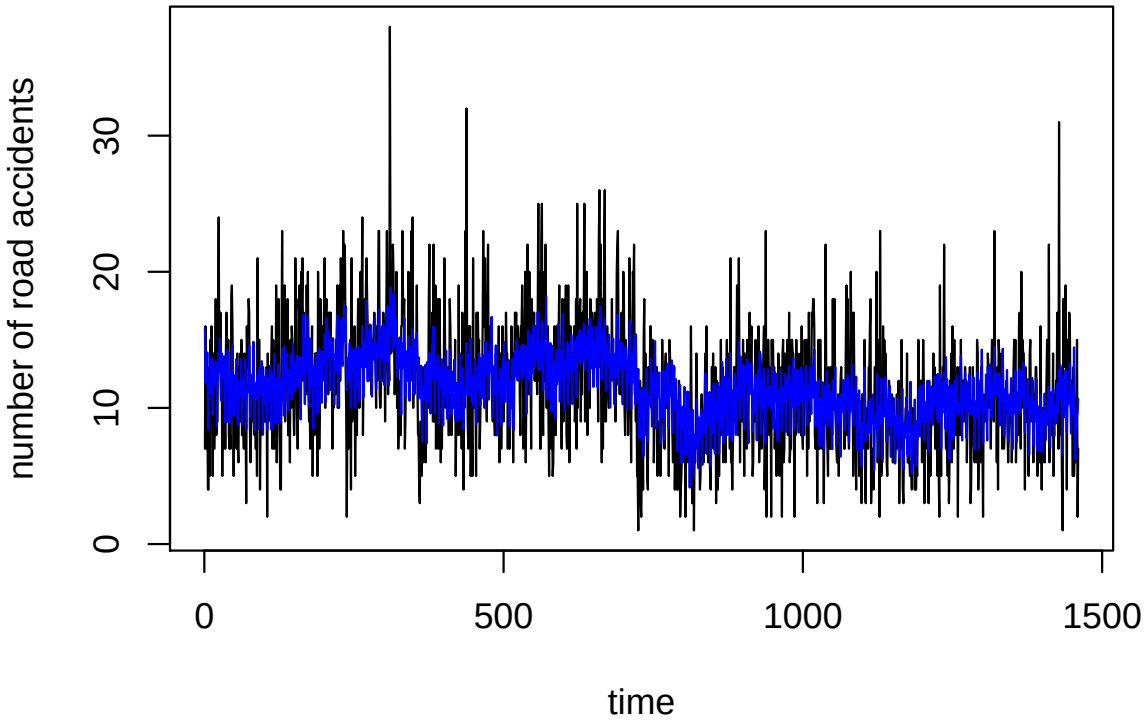
As we can see, the in sample fit seems to be very good. The mean absolute error (MAE) is equal to 3.065 and the root mean squared error (RMSE) is equal to 3.925. We will also backtest our model. Specifically, we exclude the last 365 observations and estimate the model using only the first 1095 observations, which includes, among other things, the determination of the orders of the ARMA model for E_t and the value of K.

Table 2: First model (shorter time series): BIC criterion for various values of K.

Number of basis (K)	Bayesian information criterion
$K = 5$	6333.874
$K = 10$	6391.730
$K = 15$	6440.056
$K = 20$	6504.617

Source: Own elaboration.

Figure 4: Black line represents observed number of daily road accidents. Blue line represents fitted values.



Source: Own elaboration.

Table 3: Second model (whole time series): Bayesian information criterion for various values of K

Number of basis (K)	Bayesian information criterion
$K = 5$	8354.381
$K = 10$	8401.750
$K = 15$	8449.840
$K = 20$	8513.029

Source: Own elaboration.

As we can see from Table 2, using BIC as a decision criterion, we also opt for $K = 5$, which implies the use of 10 basis functions in Equation (4). Also in this case, the auto.arima procedure leads to the same orders as in the case with the complete data. After fitting the model on the training set (the first 1095 observations), we predict the last 365 observations (the test set). The RMSE of the prediction is equal to 4.558 and MAE is equal to 3.593. Based on the backtesting we can conclude that the model is reasonably stable over time. In the following parts

of the chapter we will estimate the second and third model (see above). We follow analogous procedures to those described above.

For the second model, K is selected as 5 (see Table 3) and the auto.arima procedure leads to an ARMA($p = 4$,$q = 3$) model, which is a more complex ARMA model than in the first case. So even if we removed 3 parameters from the model, the auto.arima brought back 4 extra parameters compared to the first model. So in the end we have one more parameter in the model. In sample RMSE is higher in the second model than in the original one. Namely, RMSE is equal to 3.943 and MAE is equal to 3.063. However, following the same backtesting procedure as in the first model, we get the out of sample RMSE equal to 4.864 and the out sample MAE equal to 4.018, which are higher values than in the original situation.

In the third model, we add two external variables into the O_t part compared to the first model (see above). The first variable is the unexpected frost represented by a dummy variable equal to 1 if temperature drops below zero in November. This dummy variable is supposed to capture the common situation when drivers underestimate the arrival of winter season and they do not change summer tires in time. The second external variable is a numeric variable representing the average monthly search volume of the keyword "doprava" (traffic) on Google. Increased volume of such search can indicate that something wrong is going on on the roads. We will repeat the same procedure as in the previous models, which leads to the same value of K. For this value of K an ARMA($p = 1$,$q = 1$) was chosen as an optimal model. The in sample RMSE is equal to 4.134 and MAE to 3.192. The out of sample RMSE is equal to 4.546 and MAE to 3.484. So within the third model, we have achieved slightly better results compared to the original model, but the difference between RMSE and MAE of the first and third model is very small. In Table 4 estimates of regression coefficients for selected explanatory variables can be found.

Table 4: Point estimates of regression parameters and estimates of their standard error of selected variables for model number 1 and 3.

	Model 1		Model 3	
Variable	Point estimate	Standard error	Point estimate	Standard error
Rainfall	1.646	0.302	1.662	0.302
Frost	0.0747	0.502	-0.268	0.509
Frost:Rainfall	0.577	1.375	0.328	1.371
Google hits	-	-	0.032	0.015
November frost	-	-	2.977	1.082

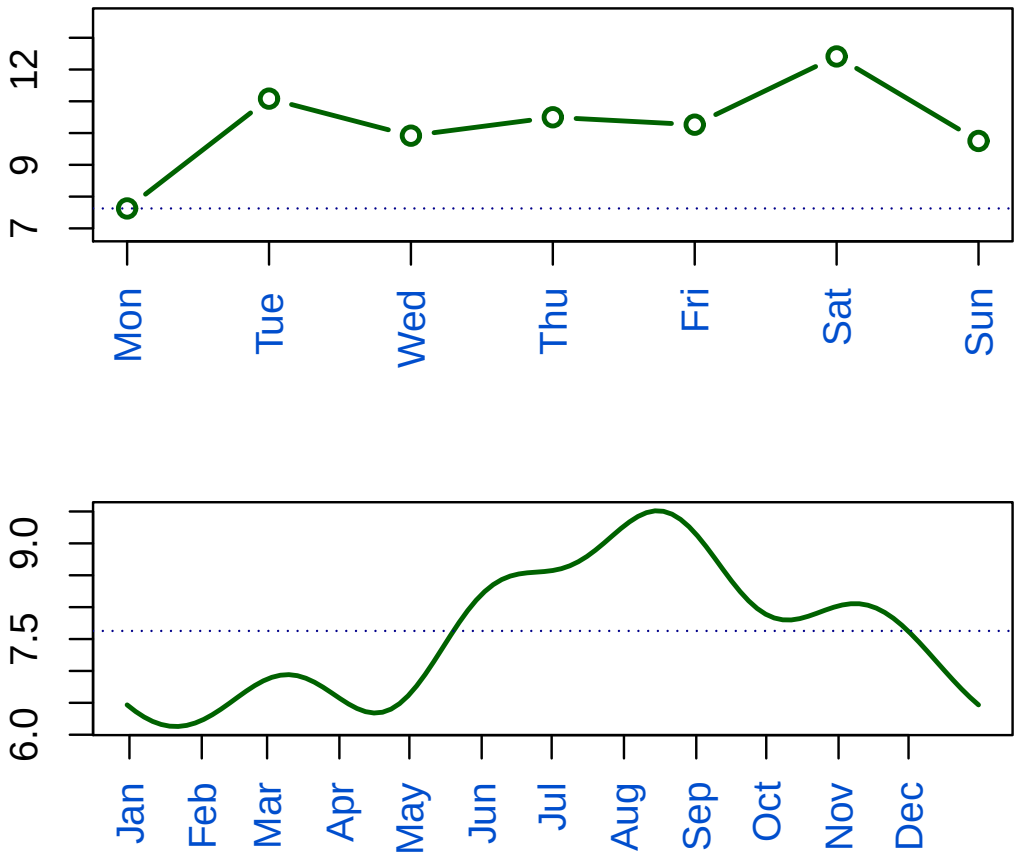
Source: Own elaboration.

As we can see in Table 4, increased rainfall has large positive effect on average number of daily road accidents. Combination of frost and rainfall also contributes to an increase of average number of daily road accidents. On the other had, from the first model one can conclude, that the frost by itself does not bring such a contribution. But from the Model 3 we can see, that frost in November has high positive effect on the average number of daily road accidents. This effect can be caused by the fact, that frost in November is not expected by drivers and therefore many drivers can be using summer tires.

The estimate of the both seasonal cycle derived from Model 3 is demonstrated in Figure 5. Regarding days of the week, the lowest average number of road accidents is registered on Mon-

day and the highest number on Saturday. Such a phenomenon can be caused by the fact that uninsured drivers usually do not use their cars on daily basis and Saturday can be the day when they use their car for short distance trips to the mall etc. Regarding days of the year, the highest number of road accidents is registered in summer months.

Figure 5: The estimate of seasonal cycle. First chart displays seasonality due to different days of the week. Second chart displays seasonality due to days of the year. Dashed blue line in both charts represents intercept of the model.



Source: Own elaboration.

5. Conclusion

After having constructed three models for our data set, we can conclude that including external variables improved the performance of our models in terms of out of sample RMSE and MAE. Not only the temperature and rainfall, but also the indicator of unexpected events such as frost in November seems to be useful for prediction. Also the average monthly search volume of the word "doprava" (traffic) on Google seems to be beneficial for the prediction model. For further research, it could be interesting to explore a larger number of time series of a similar type, so that the true importance of external variables for prediction modeling can be

evaluated. On the other hand, it is obvious that such a large set of time series is very hard to get.

Acknowledgements

This paper is supported by the grant F4/67/2016 (Modelování sezonních časových řad s velkou délkou sezónnosti) which has been provided by the Interní grantová agentura Vysoké školy ekonomické v Praze (Internal grant agency of the University of Economics in Prague).

References

[1] Gould, P. G., Koehler, A. B., Ord, J. K., Snyder, R. D., Hyndman, R. J.,Vahid-Araghi, F. 2008. Forecasting time series with multiple seasonal patterns. European Journal of Operational Research, 191(1),pp. 207-222.

[2] Hyndman, R. J., Athanasopoulos, G. 2014. Forecasting: principles and practice. OTexts

[3] Hyndman, R. J., Khandakar, Y. 2008 Automatic time series forecasting: the forecast package for R. Innsbruck: Journal of Statistical Software. 1-22pp., http://www.jstatsoft.org/article/view/v027i03.

[4] Prochazka, J. 2017 Modelování sezónnosti s dlouhou délkou periody – aplikace na modelování počtu denních dopravních nehod nepojištěných vozidel. Praha: Den doktorandů 2017. 152–158pp., http://fis.vse.cz/wp-content/uploads/2016/10/dd_fis_2017_clanky_cely_sbornik.pdf.

[5] Prochazka, J., Flimmel, S., Jantos, M., Basta, M. 2016 Long seasonal periods modeling. Banská Štiavnica: AMSE 2016. 292–301pp., http://amse.umb.sk/proceedings/ProchazkaFlimmelJantosBasta.pdf.

[6] Ramsay, J. O., Silverman, B. W. 2002 Applied functional data analysis: methods and case studies. New York: Springer.