Prediction of Daily Traffic Accident Counts and Related Economic Damage in the Czech Republic

Dana Nejedlová[1]

Abstract. Free datasets describing weather and traffic accidents in the Czech Republic have been used for the training of neural network that would predict the number of traffic accidents and the level of economic damage for a given day. The aim of the research is to find out whether there are enough statistical dependencies in the available data so that a practically usable predictor could be trained from them. The Pearson's chi-squared test was used to select input attributes for the neural network. The selected attributes are month, day of week, temperature in two selected preceding days and in the current day, precipitation, and snow. The neural network has been trained on the daily data of the years 2009 till 2014 divided into training and development test sets. The accuracy of the network after this training on more recent days is higher than majority voting, which can motivate a future research.

Keywords: Pearson's chi-squared test, feed-forward neural network, traffic accidents.

JEL Classification: C32, C45, R41
AMS Classification: 68T, 62H

1 Introduction

Predicting traffic accidents is a challenging task because they are not planned and apart from e.g. insurance frauds everyone wishes to avoid them. Scientific studies of traffic accidents focus on identifying factors that influence their incidence. Article [6] models by logistic regression the influence of rainfall on single-vehicle crashes using the information about crash severity, roadway geometries, driver demographics, collision types, vehicle types, pavement conditions, and temporal and weather conditions from databases available in Wisconsin, United States of America. Article [8] models by linear regression traffic volume in Melbourne, Australia. It identifies weather as a significant factor influencing the traffic volume as well as daytime and nighttime periods, day of the week, month, and holidays. The traffic volume is then used to predict the count of accidents by a nonlinear model that takes into account the observation that the number of accidents is directly proportional to the traffic volume and to the severity of weather conditions, but at the same time the inclement weather deters motorists from venturing onto the road and thus reduces the traffic volume, which justifies the used nonlinear regression model. Article [2] presents autoregressive time-series model of daily crash counts for three large cities in the Netherlands. This study uses data about daily crash counts, accurate daily weather conditions (wind, temperature, sunshine, precipitation, air pressure, and visibility), and daily vehicle counts for each studied area. This article also shows that if information on daily traffic exposure is not available, the information about day of the week is a good substitute. Article [15] uses autoregressive time-series model of annual road traffic fatalities in Great Britain and the monthly car casualties within the London congestion charging (CC) zone to investigate the impact of various road safety measures and legislations (seat-belt safety law, penalty points for careless driving, driving with insurance, and seat-belt wearing for child passengers). This study uses also the data about the traffic volume in the form of annual vehicle-kilometers travelled in Great Britain. Article [1] suggests a negative binomial model as the most appropriate for predicting traffic accidents using various details about studied location of traffic and about the drivers (age and gender). Available databases allowed assigning accidents to particular segments of State Road 50 in Central Florida, United States of America. Each segment was characterized by roadway geometries, traffic volumes, and speed limits. Weather was not included in the variables that contribute to accident occurrence. Results of this research identify combinations of qualities of drivers and roads which are most likely to result in an accident, which can be used as a source for arranging preventive interventions.

This contribution identifies attributes significantly influencing traffic accident counts in the Czech Republic. It draws these attributes from free publicly available datasets about daily traffic accident counts, related economic damage, and weather. These datasets do not contain as much details as those that were available to the above-mentioned studies. It is not surprising that the precision of prediction of traffic accidents using these datasets is low but when this precision is higher than the precision of a mere guessing then it indicates the existence of

[1] Technical University of Liberec, Faculty of Economics, Department of Informatics, Studentská 1402/2, 461 17 Liberec 1, Czech Republic, e-mail: dana.nejedlova@tul.cz.

statistical dependencies that hold true in the whole studied time period. The existence of prevailing statistical dependencies can motivate future research using longer data series, additional data sources that could possibly become available, and various data models, the result of which could be a practically usable predictor of the number of accidents and the economic damage caused by these accidents for a given time period of the future. The data modeling technique used in this study is Pearson's chi-squared test to identify the input data for the feed-forward neural network.

2 Data sources

Datasets used in this study are sources of information about traffic accidents, weather, and geomagnetic activity as some studies, e.g. [17, 18], indicate a possible effect of solar and geomagnetic activity on traffic accidents.

2.1 Traffic accidents

Dataset of traffic accidents is published by the Police of the Czech Republic on their web page [14]. The earliest record in this source is for February 10, 2009. The dataset is updated on a daily basis, and the data can be downloaded separately for each day. The data are structured by date and by 14 Czech regions, each pair of day and region having the following attributes: number of accidents, fatalities, severe injuries, minor injuries, and economic damage in thousands of Czech crowns (CZK). The accidents are classified also by their causes (speeding, not giving way, inappropriate overtaking, inappropriate driving, other cause, alcohol) in such a way that the sum of the number of accidents having a certain cause is equal or higher than the number of accidents, suggesting that some accidents have been assigned to more than one cause. There is an error in this database: on August 26, 2012 there are no data.

2.2 Weather

Weather dataset [11, 12] was acquired from the National Centers for Environmental Information. The earliest record in this source is for January 1, 1775. The dataset is updated on a daily basis with a several day delay. Data for the Czech Republic can be ordered from [13] by clicking the "Add to Cart" button, "View All Items" button, selecting the "Custom GHCN-Daily CSV" option, selecting the Date Range, clicking the "Continue" button, checking all Station Detail & Data Flag Options flags, selecting Metric Units, clicking the "Continue" button, entering email address, and clicking the "Submit Order" button. The data are structured by meteorological station and date of observation. For each of this pair the following attributes are available: precipitation in mm, snow depth in mm, and maximum and minimum temperature in degrees C. There are some errors in this dataset: records for March 18, 2007 and September 9, 2013 are missing. Some other days exhibited missing measurements in all meteorological stations. For these days the missing values have been added to the dataset from internet weather service [4] that presents measurements of a number of Czech meteorological stations for a selected day. Of all available meteorological stations the data from station Ruzyně were used.

2.3 Geomagnetic activity

Dataset of geomagnetic activity can be downloaded for a submitted time period from web page [5]. The earliest available date of observation is January 1, 2000. The data are structured by days.

3 Data preprocessing

Available time series of days have been divided into 3 parts defined by Table 1. The first two parts called Training Set and Development Test Set have been used for selecting useful attributes for the feed-forward neural network that would predict the number of accidents and related economic damage for a given day and for the estimation of parameters of this neural network. The third part called Test Set is used for the presentation of results in Section 5. This methodology of dividing the data into three sets is recommended e.g. in [7] and also in [3] where the Development Test Set is called Validation Set.

Set	From Date	To Date	Number of Days
Training	2009.02.10	2014.02.09	1826
Development Test	2014.02.10	2015.02.09	365
Test	2015.02.10	2016.04.19	435

Table 1 Division of available data

In order to obtain meaningful results of analysis the acquired data have been transformed into values as described below.

3.1 Traffic accidents

Of all available information in [14] the total number of accidents and the total economic damage per day for all Czech regions have been selected as the values that will be predicted.

The number of accidents

There is a rising tendency in this time series that demanded transforming the original values into their differences from quadratic trend of the period spanning the Training and Development Test sets. The values of these differences have been transformed into three approximately equifrequent categories defined by Table 2. Without applying the quadratic trend the count of the "Low" category would be inadequately low in Development Test and Test sets containing late days, see Table 1, which would make the prediction of the Test Set from the Training and Development Test sets impossible.

Economic damage

This time series has been processed the same way as the number of accidents, see Table 2, but before this step the original values have been deprived of outliers. The outliers have been defined as the values above 50,000 thousand CZK. The outliers have been substituted by the mean value of the joint Training and Development Test sets equal to 16,876 thousand CZK, which resulted in classifying them into the "High" category.

Time Series Quadratic Trend	**The Number of Accidents** $y = x - 6.37E\text{-}06 \cdot t^2 - 0.00471 \cdot t - 196$			**Economic Damage** $y = x - 6.69E\text{-}04 \cdot t^2 + 1.5025 \cdot t - 13297$		
Border Values	**Lower than –20**	**>= –20 and < 22**	**>= 22**	**Lower than –2000**	**>= –2000 and < 1300**	**>= 1300**
Set / Category	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**
Training	602	628	596	611	611	604
Development Test	132	100	133	119	126	120
Test	141	123	171	131	139	165

Table 2 Division of daily values in the traffic accidents dataset into categories

In Table 2 the x symbol stands for the original number of accidents and the number of thousand CZK with outliers substituted by the mean value, t stands for the order of the day in the time series where the date 2009.02.10 has t equal to 1, and y is the resulting value which is assigned the category defined by border values.

3.2 Weather

From weather dataset [11, 12], where each row is an observation of a single meteorological station for a single day, all available information has been utilized. To predict traffic accidents for a given day, weather for this day had been transformed into single values of attributes listed in Section 2.2.

Precipitation and snow depth

Two columns for a binary value of either precipitation or snow have been added to the weather dataset [11, 12] with value equal to zero when the observed value of precipitation or snow in a meteorological station in this row was either zero or missing. When the observed value was above zero, its binary value was equal to one. A single value of either precipitation or snow depth for a single day has been computed as arithmetic mean of these binary values for all meteorological stations respectively for precipitation and snow depth.

Temperature

For each meteorological station a single temperature for a single day has been computed as arithmetic mean of its maximum and minimum temperature. Missing values have been ignored. Missing values of both maximum and minimum temperature resulted in a missing value of arithmetic mean. A single value of temperature for a single day has been computed as arithmetic mean of these mean values for all meteorological stations ignoring missing values. The resulting temperature has been classified into categories Low, Medium, and High using border values 5 and 14 degrees C in a similar way to the border values –20 and 22 or –2000 and 1300 in Table 2.

3.3 Geomagnetic activity

Each row of the dataset obtainable from web page [5] contains values of 8 K-indices measured in a three-hour interval in a given day characterizing geomagnetic activity of this day. Missing measurements denoted as "N" have been replaced by a preceding measurement. The underlying theory is explained in [16]. Each day has been classified into one of possible categories Quiet, Unsettled, Active, Minor storm, Major storm, and Severe storm according to rules stated in [9].

4 Pearson's chi-squared tests

Pearson's chi-squared tests [19] employ categorical values obtained in data preprocessing described in Section 3. Data involved in their computation come from merged Training and Development Test sets. The results are summarized in Table 3.

Row	Test	Degrees of Freedom	0.995 Quantile	X^2
1	Temperature and the Number of Accidents	4	14.8643	48.567
2	Temperature and the Level of Economic Damage	4	14.8643	40.686
3	Geomagnetic Activity and the Number of Accidents	6	18.5490	6.0727
4	Precipitation and the Number of Accidents	2	10.5987	0.2950
5	Precipitation and the Level of Economic Damage	2	10.5987	6.9586
6	Snow and the Number of Accidents	2	10.5987	41.0788
7	Snow and the Level of Economic Damage	2	10.5987	44.3082
8	Month and the Number of Accidents	22	42.7958	229.361
9	Month and the Level of Economic Damage	22	42.7958	130.746
10	Day of the Week and the Number of Accidents	12	28.2995	1094.896
11	Day of the Week and the Level of Economic Damage	12	28.2995	745.122
12	Week after the Transition to or from Daylight Saving Time (DST) and the Number of Accidents	4	14.8643	14.263

Table 3 Tested dependencies

Remarks to the rows of Table 3 are in Table 4.

Row	Commentary to the Statistical Tests for a Given Row in Table 3
1, 2	There is a low number of accidents and a low level of economic damage on days with low temperature and vice versa.
3	The influence of geomagnetic activity on the number of accidents is not significant. Only 4 categories of the lowest levels of geomagnetic activity listed in Section 3.3 have been observed, that is why there are 6 degrees of freedom: $(4 - 1) \cdot (3 - 1)$, 3 standing for Low, Medium, and High number of accidents.
4, 5	The influence of precipitation on and the number of accidents and on the level of economic damage is not significant. Precipitation was treated as a binary value. The first category was for days with zero precipitation and the second category was for days with the precipitation above zero. The computation of precipitation is described in Section 3.2.
6, 7	There is a low number of accidents and a low level of economic damage on days with non-zero snow depth and vice versa. Snow depth was treated the same way as precipitation in a contingency table.
8-11	Month and day of the week are strongly correlated with the number of accidents and the level of economic damage, suggesting that the primary cause of accidents and economic damage is the level of traffic.
12	Week after the transition to or from daylight saving time and the number of accidents might be somewhat correlated. Days have been classified into three categories: days of the week after the transition to the DST, days of the week after the transition from DST, and the other days. The week begins on Sunday.

Table 4 Details about tested dependencies in **Table 3**

Tested dependencies with X^2 very close to a 0.995 Quantile (i.e. 1 – 0.995 is the probability that the tested pairs of attributes are independent) have been examined more closely in a series of contingency tables differing in a shift of the attributes Temperature, Week after the Transition to or from DST, Precipitation, and Geomagnetic Activity relatively to the other attributes The Number of Accidents and The Level of Economic Damage. The

result can be seen in Figure 1. The horizontal axis shows the number of days by which the attributes have been shifted. Had there been a peak around the zero shift, a significant dependency could have been detected in this way. Peaks in the negative part of the horizontal axis mean that the shifted attributes have a delayed effect on accidents and related economic damage. Peaks in the positive part of the horizontal axis mean that the categories of accidents and related economic damage have counts different from random distribution on days preceding the change in the shifted attributes.

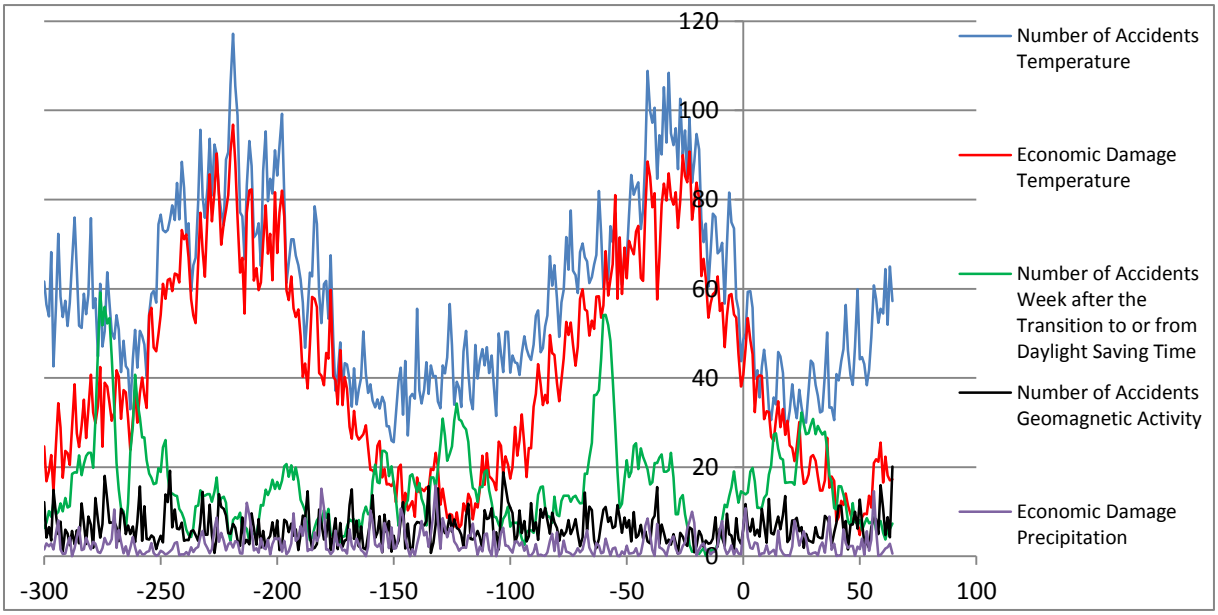

Figure 1 Pearson's chi-squared tests with a time shift in days

It can be seen from Figure 1 that Precipitation and Geomagnetic Activity do not have a significant influence at least in the used categorical representation.

The Transition to or from DST has a significant peak distant approximately a month from the zero shift in the positive part of the horizontal axis. Approximately 2 weeks after the time shift to or from DST the Number of Accidents becomes randomly distributed. Examining the related contingency table reveals that there is a relatively low number of accidents before the spring time shift to DST and a relatively high number of accidents before the autumn time shift from DST. Two biggest peaks at –60 and –276 are probably a manifestation of a yearly temperature and traffic cycle described in the next paragraph. It is probable that all significant peaks on the Transition to or from DST dataset in Figure 1 are caused by this cycle.

The data series involving Temperature in Figure 1 suggest that accidents and related economic damage have a cyclic character and that they react to the change in temperature with approximately a month's delay. This delay suggests that the primary cause of accidents and economic damage is the level of traffic. There is a low number of accidents and low level of economic damage in days with low temperature and vice versa in the contingency table for a minus-32-day shift. And there is a high number of accidents and high level of economic damage in days with low temperature and vice versa in the contingency table for a minus-219-day shift. It means that these two peaks in Figure 1 are anticorrelated. Although having correlated (or anticorrelated) attributes in the input of a predictor is not recommended, see e.g. [3], the neural network described in Section 5 has been usually more accurate when it used both of them.

5 Prediction using the feed-forward neural network

A feed-forward neural network has been reported in paper [10] dealing with traffic crashes occurred at intersections as a better solution than the Negative Binominal Regression employed also in articles [1, 2, 15].

A feed-forward neural network has been programmed in C language according to the algorithm described in [20] and used to predict separately the category of the number of accidents and the level of economic damage from month, day of week, temperature 219 days ago, temperature 32 days ago, temperature for the current day, precipitation and snow depth. All categorical attributes have been encoded in the form of the so-called *dummy variables* described in [3]. Precipitation and snow depth have been represented respectively as a single value described in Section 3.2. The network has been trained according to the Training Set and the training has been stopped when its accuracy of prediction of the Development Test Set was maximal. Neural network with final weights after this training has predicted correctly the category of the Number of Accidents in the range of 53% to

64% of days and the category of the Level of Economic Damage in the range of 43% to 50% of days in the Test Set, while guessing the most frequent category has accuracy of 39% and 38% respectively.

6 Conclusion

Artificial neural network can predict the number of accidents and the level of economic damage with accuracy slightly above majority voting. Its low accuracy is most likely caused by imprecise data. It would be possible, however, to create training data for a specific region and use the neural network trained with the similar methodology as described in this contribution for e.g. assessing the impact of local preventive police interventions.

References

[1] Abdel-Aty, M. A., and Radwan, A. S.: Modeling traffic accident occurrence and involvement. *Accident Analysis and Prevention*. **32** (2000), 633–642.

[2] Brijs, T., Karlis, D., and Wets, G.: Studying the effect of weather conditions on daily crash counts using a discrete time-series model. *Accident Analysis and Prevention*. **40** (2008), 1180–1190. DOI: 10.1016/j.aap.2008.01.001.

[3] Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning*. 2nd edition. Springer, New York, 2009. [accessed 2016-03-01]. Available from: http://statweb.stanford.edu/~tibs/ElemStatLearn/.

[4] In-počasí: Archiv. [online]. (2016). [accessed 2016-04-04]. Available from WWW: http://www.in-pocasi.cz/archiv/.

[5] Institute of Geophysics of the Czech Academy of Sciences, v. v. i.: Data archive - K indices. [online]. (2016). [accessed 2016-04-04]. Available from WWW: http://www.ig.cas.cz/en/structure/observatories/geomagnetic-observatory-budkov/archive.

[6] Junga, S., Qinb, X., and Noyce, D. A.: Rainfall effect on single-vehicle crash severities using polychotomous response models. *Accident Analysis and Prevention*. **42** (2010), 213–224. DOI: 10.1016/j.aap.2009.07.020.

[7] Jurafsky, D, and Martin, J. H.: *Speech and Language Processing*. Prentice Hall, Inc., Englewood Cliffs, New Jersey 07632, 2000.

[8] Keay, K., and Simmonds, I.: The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. *Accident Analysis and Prevention*. **37** (2005), 109–124. DOI: 10.1016/j.aap.2004.07.005.

[9] Kubašta, P.: *Výpočet geomagnetické aktivity*. [online]. (2011). [accessed 2016-01-28]. Available from WWW: https://www.ig.cas.cz/kubasta/PersonalPages/zpravy/noaaVypocet.pdf.

[10] Liu, P., Chen, S.-H., and Yang, M.-D.: Study of Signalized Intersection Crashes Using Artificial Intelligence Methods. In: *Proceedings of the 7th Mexican International Conference on Artificial Intelligence* (Gelbukh, A., Morales, E. F., eds.). Springer-Verlag, Berlin Heidelberg, 2008, 987–997.

[11] Menne, M. J., Durre, I., Korzeniewski, B., McNeal, S., Thomas, K., Yin, X., Anthony, S., Ray, R., Vose, R. S., Gleason, B. E., and Houston, T. G.: *Global Historical Climatology Network - Daily (GHCN-Daily), Version 3*. [FIPS:EZ - Czech Republic - 2000-01-01--2016-03-31]. NOAA National Climatic Data Center. (2012), DOI: 10.7289/V5D21VHZ. [accessed 2016-04-14].

[12] Menne, M. J., Durre, I., Vose, R. S., Gleason, B. E., and Houston, T. G.: An Overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Oceanic Technol.* **29** (2012), 897-910. DOI: 10.1175/JTECH-D-11-00103.1.

[13] National Centers for Environmental Information: Daily Summaries Location Details. [online]. 2016. [accessed 2016-04-04]. Available from WWW: http://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/FIPS:EZ/detail.

[14] Policie České republiky (Police of the Czech Republic): *Statistiky dopravních nehod* (*Traffic Accidents Statistics*). [online]. (2015). [accessed 2016-04-14]. Available from WWW: http://aplikace.policie.cz/statistiky-dopravnich-nehod/default.aspx.

[15] Quddus, M. A.: Time series count data models: An empirical application to traffic accidents. *Accident Analysis and Prevention*. **40** (2008), 1732–1741. DOI: 10.1016/j.aap.2008.06.011.

[16] Reeve, W. D.: *Geomagnetism Tutorial*. [online]. (2010). [accessed 2016-01-28]. Available from WWW: http://www.reeve.com/Documents/SAM/GeomagnetismTutorial.pdf.

[17] Střeštík, J., and Prigancová, A.: On the possible effect of environmental factors on the occurrence of traffic accidents. *Acta geodaet., geophys. et montanist. Acad. Sci. Hung*. **21** (1986), 155-166.

[18] Verma, P. L.: Traffic accident in India in relation with solar and geomagnetic activity parameters and cosmic ray intensity (1989 to 2010). *International Journal of Physical Sciences*. **8**.10 (2013), 388-394. DOI: 10.5897/IJPS12.733.

[19] Wilcox, R. R.: *Basic Statistics*. Oxford University Press, Inc., New York, 2009.

[20] Winston, P. H.: *Artificial intelligence*. 3rd edition. Addison Wesley, Reading, MA, 1992.