Trend and Seasonality in Fatal Road Accidents in the U.S. States in 2006–2016

Jiří Procházka [1] | *University of Economics, Prague, Czech Republic*
Milan Bašta [2] | *University of Economics, Prague, Czech Republic*
Matej Čamaj [3] | *University of Economics, Prague, Czech Republic*
Samuel Flimmel [4] | *University of Economics, Prague, Czech Republic*
Milan Jantoš [5] | *University of Economics, Prague, Czech Republic*

Abstract

Understanding the dynamics of the daily number of fatal road traffic accidents is important for local authorities, police departments, healthcare facilities and insurance companies, enabling them to design preventive measures, provide appropriate emergency service and care and reliably estimate traffic accident insurance costs. In the present study, using the Fatality Analysis Reporting System provided by the U.S. National Highway Traffic Safety Administration, we construct a daily time series of the number of accidents for each state of the United States. We model the trend as well as yearly and weekly seasonality present in the time series and provide respective trend and seasonality statistics. Differences in accident rates and yearly seasonality between states were detected, clustering analysis being applied to identify clusters of states with similar yearly seasonality, weekly seasonal patterns for different states proving to be about the same.

Keywords

Road traffic accidents, yearly seasonality, long seasonal period, generalized linear model, model selection, cluster analysis

JEL code

C22, R41

INTRODUCTION

The main aim of this paper is to examine the daily number of motor vehicle accidents on the roads of the Unites States that involve at least one fatality. Special focus will be given on the characteristics

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic.

[2] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. Corresponding author: e-mail: milan.basta@vse.cz.

[3] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic.

[4] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic.

[5] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic.

of the trend and seasonal components, the yearly seasonal component in particular. The findings of the study can be useful for local authorities, police departments and hospitals as well as for insurance companies.

Levine et al. (1995) investigated changes in the daily number of motor vehicle accidents for the City and County of Honolulu in 1990. Their results suggest that more accidents occur during Fridays and Saturdays and during minor holidays. They also identified weather conditions as a relevant factor influencing the number of accidents. Nofal and Saeed (1997) examined monthly variations of the number of road accidents in Riyadh city from 1989 through 1993. Among others, they observed seasonal variations in the number of accidents, the accidents being maximal during the summer season. Edwards (1996) identified an increasing pattern in the number of accidents throughout the calendar year for England and Wales in the period from 1980 to 1990, the first quarter of the year having the lowest level and the last quarter the highest level of accidents. Jones et al. (2008) studied variations in mortality and morbidity from road traffic accidents in England and Wales from 1995 through 2000. Using a geographical approach and district-level data with various explanatory variables (population numbers and characteristics, traffic exposure, road length, curvature and junction density, land use, elevation, hilliness, etc.), they identified risk factors that predicted variations in mortality and morbidity.

In our analysis, we construct daily time series of the number of motor vehicle road accidents involving at least one fatality for each state of the United States from 2006 to 2016. We model the trend and seasonal components of the time series for each state separately and provide summary statistics. We also explore geographical associations.

The data used in the analysis are presented in Section 1. The model for the number of accidents is introduced in Section 2. The results of the analysis are provided in Section 3. The last section concludes.

1 DATA

Fatality Analysis Reporting System (FARS) data provided free of charge by the U.S. National Highway Traffic Safety Administration (NHTSA)[6] have been used. The FARS database offers detailed information on each motor vehicle road accident in the United States which involves at least one fatality. FARS stores each accident as an individual data record with a unique identification number, revealing the details of the accident such as the date and time, geographic location, number of fatalities, weather conditions, etc.

We have removed duplicate records using the identification number of each accident. Yearly, monthly and daily averages and corresponding rates per 1 000 population (in parentheses) for the entire United States[7] are as follows: 32 664.64 (0.1058), 2 722.97 (0.0088) and 89.49 (0.0003).

Then we constructed a daily time series of number of accidents for each U.S. state from the beginning of 2006 to the end of 2016, leap days having been removed to simplify the analysis of yearly seasonality (see below). As a result, we have obtained a time series consisting of eleven times 365 (i.e. 4015) observations for each of the 50 states of the U.S.A. and the District of Columbia.[8]

Average annual accident rates per 1 000 population for each state[9] in the 2006–2016 period are presented in Figure 1. The average rate per state is 0.12, the lowest (0.050) and highest (0.21) ones occurring in the District of Columbia and in Mississippi, respectively. The Pacific coast and north-eastern states generally report lower rates per 1 000 population compared to the rest of the United States.

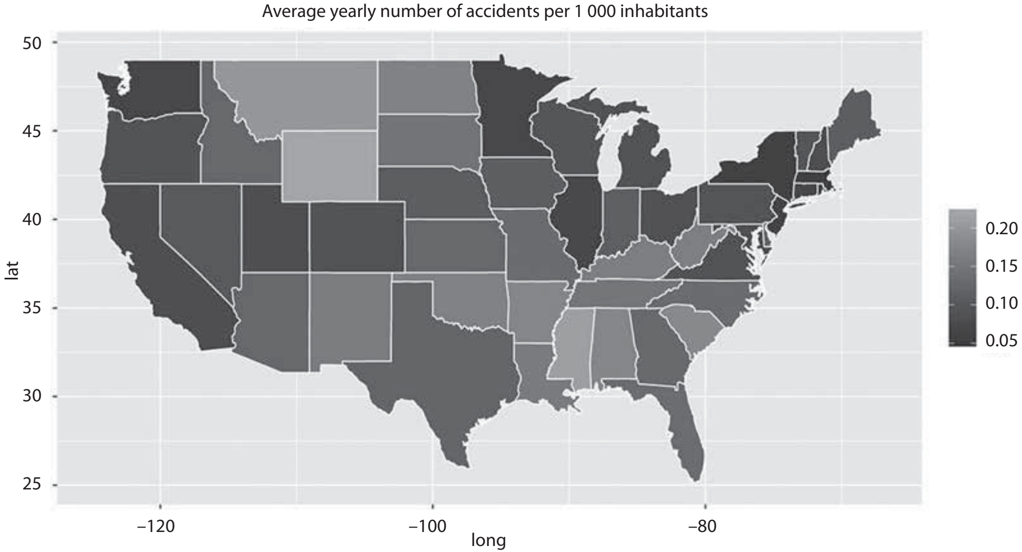
[6] https://www-fars.nhtsa.dot.gov.

[7] The number of inhabitants was obtained from the 2010 United States Census at: https://www.census.gov.

[8] Although the federal District of Columbia (Washington, D.C.) is not a state, it is considered as such (i.e. "the 51st state") in the present analysis.

[9] Data on the population of each state were obtained from the 2010 United States Census, available at: https://www.census.gov.

Figure 1 Average yearly number of accidents per 1 000 population, excluding results for Hawaii (0.078) and Alaska (0.088)



Source: Own construction

2 MODEL FOR THE NUMBER OF ACCIDENTS

Let $\{X_t\}$ be a time series of the number of daily road accidents of length $N = 4\ 015$. Since the number of accidents is a non-negative integer, it can be assumed – providing that the number of accidents is not large enough – that X_t (i.e. the number of accidents at a specific time t) is non-Gaussian. Thus, it may be useful to consider some distribution for X_t which relaxes the normality assumption. Specifically, we can assume that X_t has a density belonging to the exponential family of distributions (Nelder and Baker, 1972; McCullagh and Nelder, 1989):

$$f\left(X_t;W_t,k_t\right)=\exp\left(\frac{X_tW_t-\mathrm{a}\left(W_t\right)}{k_t}\right)\mathrm{c}\left(X_t,k_t\right), \tag{1}$$

where W_t is the canonical parameter, k_t the dispersion parameter and a(.) and c(.) denote some functions. The expected value of a random variable from the exponential family of distributions is equal to the first derivative of $\mathrm{a}(W_t)$ with respect to W_t, while the variance is equal to k_t times the second derivative of $\mathrm{a}(W_t)$ with respect to W_t. Further, the second derivative of $\mathrm{a}(W_t)$ with respect to W_t expressed as a function of the expected value is called the variance function and captures the relationship between the variance of the random variable and its mean.

To be more specific, let us assume that X_t is a Poisson random variable with parameter λ_t. Such a distribution is a special case of an exponential family distribution with $W_t = \log \lambda_t$, $\mathrm{a}(W_t) = e^{W_t}$ and $k_t = 1$. Consequently, we get the following results: the expected value of X_t is given as $\mu_t = E(X_t) = \mathrm{a}^{(1)}(W_t) = \lambda_t$, its variance as $D(X_t) = k_t\mathrm{a}^{(2)}(W_t) = \lambda_t$ and the variance function as $V(\mu_t) = \mu_t$.

We further assume that $\{X_t: t = 1, \ldots, N\}$ is a sequence of N *independent* Poisson random variables with parameters λ_t ($t = 1, \ldots, N$), the means μ_t ($t = 1, \ldots, N$) of the variables being given as:

$$\log\left(\mu_t\right)=T_t+S_t,\ \ t=1,\ldots,N, \tag{2}$$

$$\mu_t=\exp\left(T_t+S_t\right)=\exp\left(T_t\right)\exp(S_t),\ \ t=1,\ldots,N. \tag{3}$$

The model of Formula 2 can be considered as a generalized linear model (GLM), see McCullagh and Nelder (1989). In generalized linear models a monotonic function of the expected value, called a link function, rather than the expected value itself is modeled as a linear combination of regressors. This linear combination is called a linear predictor.

A possible choice (but not the only one) of the link function is the *canonical* link function $g(\mu_t)$ which satisfies $g^{(1)}(\mu_t) = \frac{1}{V(\mu_t)}$. For the case of Poisson distribution this implies that the canonical link function is a logarithmic function. Such a canonical link function is also used on the left-hand side of Formula 2.

The linear predictor on the right-hand side has two parts: a trend component of the linear predictor (T_t) and a seasonal component of the linear predictor (S_t). The two parts will be specified in Sections 2.1 and 2.2. The trend and seasonal component in the mean daily number of accidents are given as $\{\exp(T_t)\}$ and $\{\exp(S_t)\}$, the model for the mean daily number of accidents being multiplicative (see Formula 3).

McCullagh and Nelder (1989) or Dobson and Barnett (2008) present further details on generalized linear models which also include the estimation of the models.

In the generalized linear model we assumed that the Poisson random variables $\{X_t: t = 1,\ldots, N\}$ are independent. This assumption was checked during our analysis and was found to be reasonably satisfied (see Section 3 for details).

If the assumption of the independence of the N Poisson random variables $\{X_t: t = 1,\ldots, N\}$ was not satisfied, the generalized linear model could be extended to capture the dependence among the variables by assuming Poisson generalized ARMA models (GARMA) which can be formulated as:

$$\log(\mu_t) = R_t + \sum_{j=1}^{p}\phi_j\left[\log\left(X'_{t-j}\right) - R_{t-j}\right] + \sum_{j=1}^{q}\theta_j\left[\log\left(\frac{X'_{t-j}}{\mu_{t-j}}\right)\right], \tag{4}$$

where μ_t is the expected value of $\{X_t\}$ conditional on all the information available at time t, $R_t = T_t + S_t$, and ϕ_j, for $j = 1, 2, ..., p$, and θ_j, for $j = 1, 2, ..., q$ are parameters and X'_t is a modified time series defined as $X'_t = \max(X_t, c)$, where $c \in (0, 1)$ (Dunsmuir and Scott, 2015; De Andrade, 2016).

2.1 Trend and yearly seasonal component representation

Four different models for $\{T_t\}$ will be examined: no trend (i.e. intercept only), linear, quadratic and cubic deterministic trend.

$\{S_t\}$ will be decomposed into two parts $\{S_{1t}\}$ and $\{S_{2t}\}$, the former representing yearly and the latter weekly seasonality. Specifically, we write:

$$S_t = S_{1,t} + S_{2,t}, \qquad t = 1,\ldots,N. \tag{5}$$

Yearly seasonality has a long seasonal period, namely $L = 365$. On the other hand, the period of weekly seasonality cannot be considered as a long seasonal period since it is equal to 7.

Yearly seasonality can be modeled using cubic splines. Specifically, let us assume the following cubic spline function (Ramsay and Silverman, 2002; or Ramsay and Silverman, 2005) defined in the interval $[0, L]$, where $L = 365$,

$$C(t_*) = \beta_0 + \beta_1 t_* + \beta_2 t_*^2 + \beta_3 t_*^3 + \sum_{k=1}^{K}\theta_k\left(t_* - \xi_k\right)_+^3,\ 0 \le t_* \le L, \tag{6}$$

where $0 \le t_* \le L$ is a real-valued variable, $(.)_+$ denotes the positive part of the expression in brackets, β_0, β_1, β_2, β_3 and θ_k, for $k = 1, 2, ..., K$, are parameters and ξ_k, for $k = 1, 2, \ldots, K$, called the *knots*, are integer- or real-valued constants which satisfy $0 \le \xi_1 < \xi_2 < \cdots < \xi_K < L$. $C(t_*)$ is a piecewise cubic polynomial and is continuous in the interval $[0, L]$. At the knots, the first and second derivatives of $C(t_*)$ exist, whereas the third ones do not.

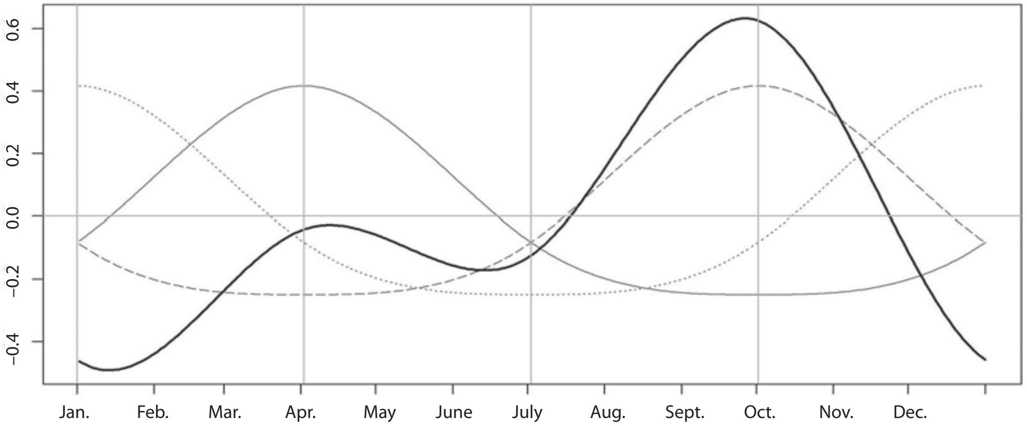
Further, the following constraints are applied to the function $C(t_*)$:

$$\begin{aligned} &C(0) = C(L), \\ &C(0)^{(1)+} = C(L)^{(1)-}, \\ &C(0)^{(2)+} = C(L)^{(2)-}, \\ &C(0)^{(3)+} = C(L)^{(3)-}, \\ &\int_0^L C(l)\,dl = 0, \end{aligned} \tag{7}$$

where $C(0)^{(1)+}$, $C(0)^{(2)+}$, $C(0)^{(3)+}$, are the first, second and third right derivatives at the point 0, and $C(L)^{(1)-}$, $C(L)^{(2)-}$, $C(L)^{(3)-}$, the three left ones at the point L. The five constraints of Formula 7 effectively reduce the number of the cubic spline parameters by 5, ensuring that both ends of the function "connect smoothly" and that the seasonal deviations sum up to zero.

Even though we do not provide an explicit formula for the constrained cubic spline since it would be too complex, it is important to emphasize that the constrained cubic spline can be written as a *linear combination* of K – 1 basis functions. An illustrative example is presented in Figure 2.

Figure 2 An example of a constrained cubic spline function (solid black curve) with four knots in the interval [0, 365]. The position of the knots is denoted by four gray vertical lines. The spline function is obtained as a linear combination of three basis functions denoted by the gray solid, gray dotted and gray dashed curve. The weights of the linear combination are chosen as 1 (solid), –0.5 (dotted) and 2 (dashed).



Source: Own construction

If the constrained cubic spline function is periodically extended with a period equal to L, we obtain a *periodic cubic spline* function. If the periodic cubic spline function is sampled at discrete values 1, 2, …, N, it can serve as a representation of the yearly seasonal component $\{S_{1,t}\}$.

Generally, a large number of knots K leads to a less-biased and high-variance estimate of the yearly seasonal pattern, whereas a low number of knots results in a highly biased and low-variance estimate. The time positioning of knots along the calendar year, though subjective to some extent, is also crucial. More densely positioned knots in a given period of the year lead to a less-biased but more variable estimate of the cycle in that period. The positioning of the knots in our paper is described in the next paragraph and the selection of the K value is given in Section 2.3.

In total, eleven different models for the yearly seasonal cycle will be examined in our analysis: no yearly seasonal cycle and ten models based on periodic cubic splines differing in the number of parameters ($K - 1$): 3, 5, 7, 9, 11, 13, 15, 17, 20 and 30. The knots will be placed equidistantly throughout the year so that the distance between two neighboring knots is constant, the first knot being placed at the beginning of the calendar year. The equidistant placement of the knots is a common choice which often works well (see e.g. Ramsay and Silverman, 2002; or Ramsay and Silverman, 2005). An alternative to the equidistant placement of the knots is to place more knots in those regions where the estimated function exhibits the most complex variations (see e.g. Ramsay and Silverman, 2005) – this approach will not be used in our analysis.

2.2 Weekly seasonal component representation

Weekly seasonality will be modeled as follows:

$$S_{2,t} = \sum_{m=1}^{6} \psi_m Z_{m,t}, \qquad t = 1, \ldots, N, \tag{8}$$

where ψ_m, for $m = 1, 2, ..., 6$, are parameters and $\{Z_{m,t}: t = 1, ..., N\}$, for $m = 1, 2, ..., 6$, are effect coding variables.

In total, two different models for weekly seasonality will be considered: a model with no weekly seasonality and the model of Formula 8.

2.3 Best model selection

88 different models will be examined for each time series, differing in the number of parameters used for the deterministic trend (four alternatives), yearly (eleven alternatives) and weekly seasonality (two alternatives) in the linear predictor. The model with the lowest value of Akaike information criterion (AIC) among these 88 models will be selected as the best, AIC being defined as:

$$AIC = 2P - 2\hat{l}, \tag{9}$$

where P is the number of parameters to be estimated and $\hat{l}$ is the natural logarithm of the maximized likelihood function.

The best model selected by AIC for each state was further checked whether it conforms to the assumptions of the GLM approach (see Section 3).

R software (R Core Team, 2017) has been employed in the analysis. Namely, the glm() function from the R stats package has been used to perform Poisson regression. The part of the model matrix corresponding to yearly seasonality has been created making use of the pbs() function from the pbs R package (Wang, 2013).

3 RESULTS

As explained above, 88 different models have been considered for each of the 51 time series, Table 1 displaying the frequencies of the models that were selected as the most appropriate.

Let $\{\widehat{X_t}: t = 1, ..., N\}$ be the fitted values from the best model and

$$\left\{ \frac{X_t - \widehat{X_t}}{\sqrt{\widehat{X_t}}} : t = 1, \ldots, N \right\} \tag{10}$$

the corresponding Pearson residuals. Based on Cameron and Trivedi (2013, Sec. 7.3.2) we have used the sample autocorrelation function of Pearson residuals from the best model as well as the related significance tests (test for autocorrelation at an individual lag based on a test statistic following normal distribution and portmanteau test based on the Box-Pierce test statistic, see Cameron and Trivedi, 2013, Sec. 7.3.2) to assess the assumption of independence of $\{X_t: t = 1,\dots, N\}$. This assumption seems to be reasonably satisfied for all the 51 time series.[10]

Table 1 Frequencies of the models selected for the 51 time series (WS stands for "weekly seasonality")

			Number of parameters of $\{T_t\}$				Row sums
			1	2	3	4	
Number of parameters of $\{S_{1t}\}$	0	no WS	0	0	0	1	1
		WS	0	0	1	0	1
	3	no WS	0	0	0	0	0
		WS	0	2	8	3	13
	5	no WS	0	0	1	0	1
		WS	0	2	9	7	18
	7	no WS	0	0	0	0	0
		WS	0	0	5	2	7
	9	no WS	0	0	0	0	0
		WS	0	0	1	2	3
	11	no WS	0	0	0	0	0
		WS	0	0	1	1	2
	13	no WS	0	0	0	0	0
		WS	0	0	1	0	1
	15	no WS	0	0	0	0	0
		WS	0	0	0	1	1
	17	no WS	0	0	0	0	0
		WS	0	0	0	1	1
	20	no WS	0	0	0	0	0
		WS	0	0	1	0	1
	30	no WS	0	0	0	0	0
		WS	0	0	0	1	1
Column sums			0	4	28	19	

Source: Own construction

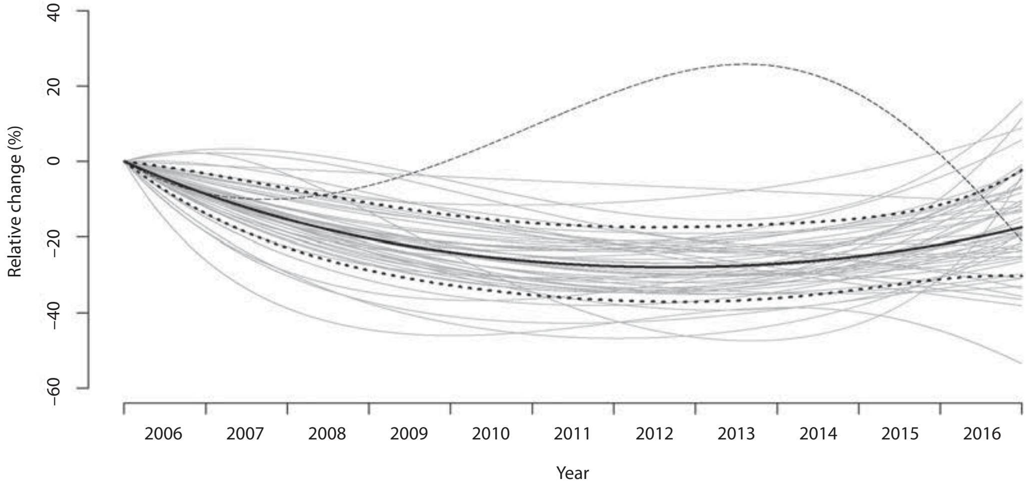
3.1 Examining the trend component

It follows from Table 1 that a quadratic (three-parameter) trend is often the best choice for $\{T_t\}$. If $\{\widehat{T_t}\}$ is an estimate of $\{T_t\}$, then the estimated trend in the mean daily number of accidents is $\{\exp \widehat{T_t}\}$. In Figure 3, the relative change of $\{\exp \widehat{T_t}\}$ (in percentage terms) is indicated for each state with respect to the beginning of the year 2006. There has been a decrease in the number of accidents in most states since 2006 – the least accidents occurring around the year 2012 –, followed by a slight increase in most states.

The relative change of the level of $\{\exp \widehat{T_t}\}$ from the beginning of 2006 to the end of 2016 is shown in percentage points in Figure 4, the average relative change being –16.4 per cent, the minimum and maximum reaching –53.6 and 16.0 per cent in South Dakota and New Hampshire, respectively. There is no clear geographical pattern in the relative change of $\{\exp \widehat{T_t}\}$.

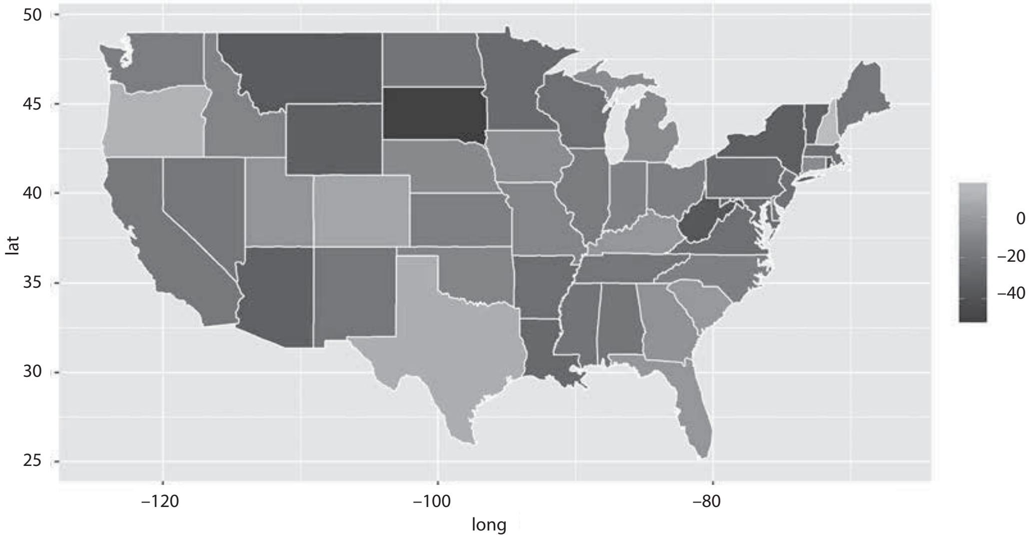
[10] To check the results, we have also performed GARMA modeling, using the glarma() function from the glarma R package (Dunsmuir and Scott, 2015), with $p = 1$, $q = 2$ and $c = 0.01$ (default value). The estimated shapes of seasonal cycles were highly similar to those obtained from GLM modeling.

Figure 3 Relative change (in percentage terms) of $\{\exp \widehat{T}_t\}$ with respect to the beginning of the year 2006 for the 51 states. North Dakota is depicted in thin dashed black, the other states in gray. The geometric mean of the corresponding growth rates (translated into a relative change) is represented by the solid black curve, the two dotted black curves denoting the distance of one geometric standard deviation from the geometric mean.



Source: Own construction

Figure 4 Relative change (in percentage terms) of $\{\exp \widehat{T}_t\}$ from the beginning of 2006 to the end of 2016, excluding results for Hawaii (–29.7) and Alaska (–2.4)



Source: Own construction

3.2 Examining the yearly seasonal component

Table 1 shows that yearly seasonality is present in most of the time series, the most common number of parameters for $\{S_{1,t}\}$ being five. The only two states that do not exhibit yearly seasonality are Hawaii and the District of Columbia. If $\{\widehat{S_{1,t}}\}$ is the estimate of $\{S_{1,t}\}$, then the estimated yearly multiplicative seasonal component in the mean daily number of accidents is $\{\exp \widehat{S_{1,t}}\}$. There is a relatively wide diversity among the 51 states in the temporal variability of $\{\exp \widehat{S_{1,t}}\}$ along the year, some of them showing minor and others considerable variability of the seasonal component.[11]

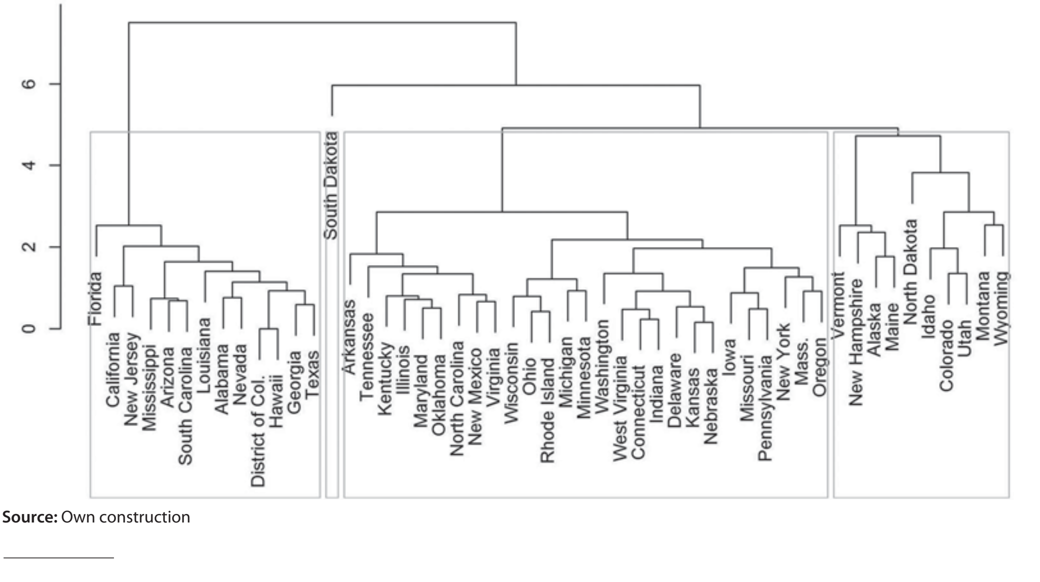
We conduct cluster analysis and form four groups of states with similar yearly seasonalities. Specifically, agglomerative hierarchical clustering with the Euclidean distance and complete linkage (see Everitt and Hothorn, 2011) is used,[12] with each state represented by 365 values[13] $\{\widehat{S_{1,t}} : t = 1, ..., 365\}$. The clustering dendrogram is presented in Figure 5, with the four clusters denoted by gray rectangles.[14] The clusters are also shown in Figures 6 and 7, yearly seasonal cycles $\{\exp \widehat{S_{1,t}}\}$ being illustrated separately for the respective clusters in the latter figure.

It is obvious that the clusters are closely related to the geographical location of each state. The first cluster comprises the Sun Belt southern states such as Florida, Texas and California. Yearly seasonality is not too variable in these states. The geometric mean of geometric standard deviations of $\{\exp \widehat{S_{1,t}}\}$ is 1.05 in the whole cluster, the geometric standard deviation for a single state being defined as:

$$\sigma_g = \exp\left(\sqrt{\frac{1}{365}\sum_{m=1}^{365}\widehat{S_{1,m}}^{2}}\right), \tag{11}$$

where $\widehat{S_{1,m}}$, for $m = 1, ..., 365$, is the estimated yearly seasonal component of the linear predictor for the mth day of the year.

Figure 5 Hierarchical clustering dendrogram. The four clusters are denoted by the gray rectangles.



Source: Own construction

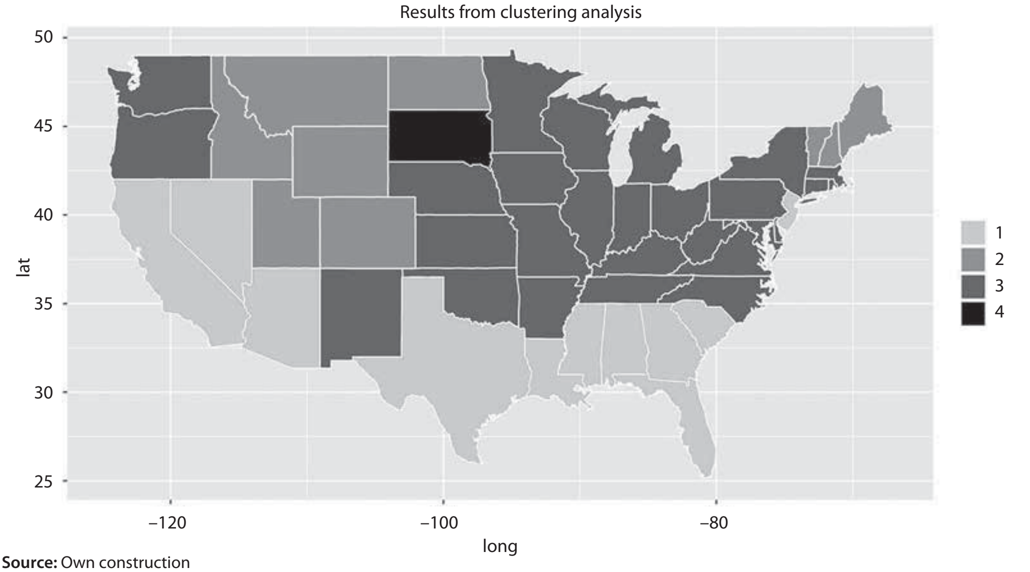
[11] Kirkwood (1979) argues that the geometric mean and geometric standard deviation are reasonable measures of location and spread for the variables which are subject to multiplicative rather than additive variations. Consequently, we use the geometric mean and geometric standard deviation as the measures of location and spread of multiplicative seasonal patterns $\{\exp \widehat{S_{1,t}}\}$ throughout the text.

[12] We use dist(), hclust() and cutree() functions from the R base package to perform the analysis in R.

[13] Since the Euclidean distance is applied, it seems more sensible to do the clustering on $\{\widehat{S_{1,t}}\}$ rather than $\{\exp \widehat{S_{1,t}}\}$.

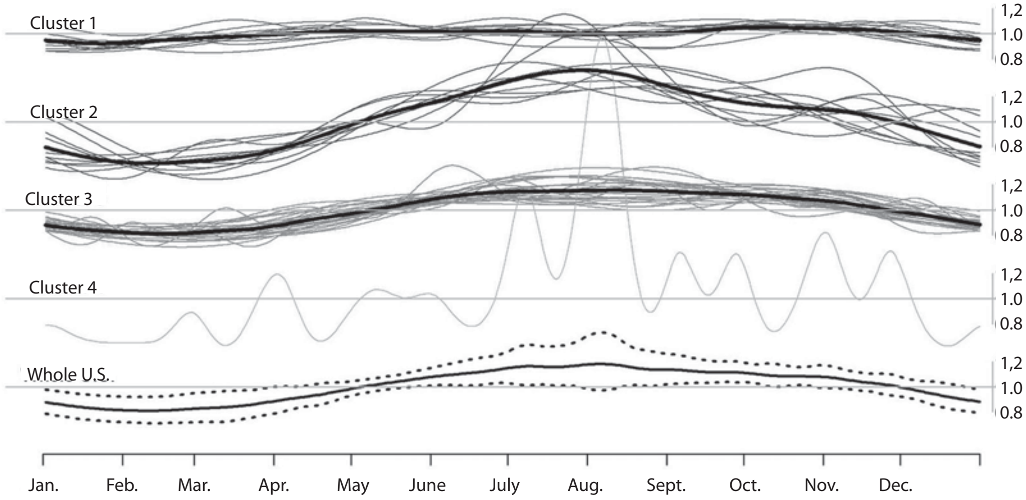
[14] Increasing the number of clusters above five does not lead to a pronounced decrease in the within-cluster sum of squares, while decreasing the number of clusters below three results in a marked increase in the within-cluster sum of squares. Thus, the choice of three, four or five clusters seems to be reasonable. In this analysis, we have opted for four clusters.

Figure 6 Members of the four clusters from hierarchical clustering, except Hawaii (part of the 1st cluster) and Alaska (2nd cluster)



Source: Own construction

Figure 7 Estimated multiplicative yearly seasonal components. The bottom plot (the entire U.S.A.) represents the geometric mean of the 51 seasonal components (thick solid black curve) and the distance of one geometric standard deviation from the geometric mean (two thick dotted curves). Individual seasonal components are presented in the upper plots being grouped into four separate clusters. The geometric mean of each cluster is shown by the thick solid black curve, except for cluster 4 which consists of only one state (South Dakota). Each cluster has its own y axis on the right-hand side of the plot.



Source: Own construction

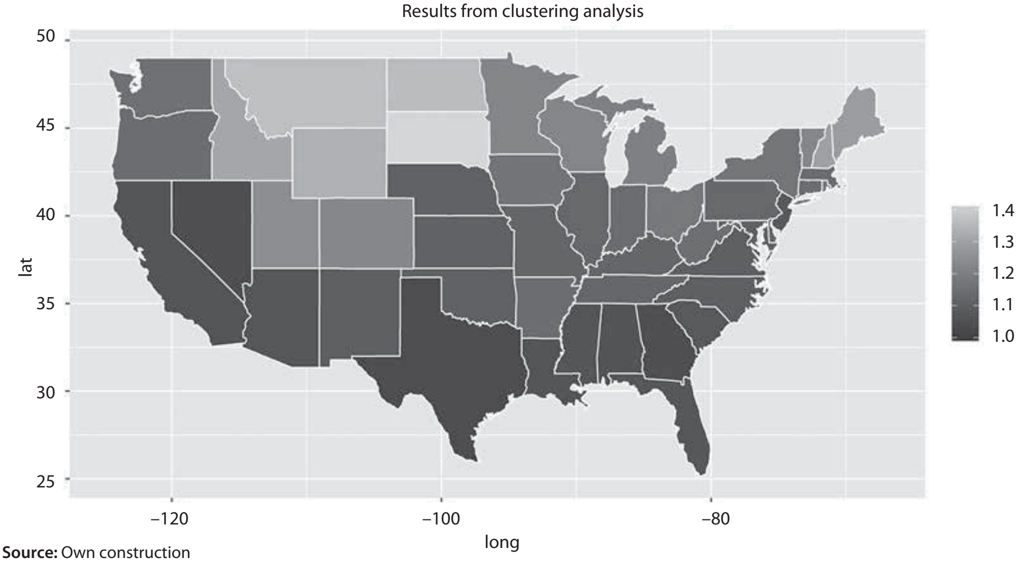
The second cluster consists of the Mountain West states, e.g. Idaho, Montana and Wyoming. The yearly seasonal cycle exhibits significant variations, the geometric mean of geometric standard deviations of $\{\exp \widehat{S_{1,t}}\}$ being 1.30.

The remaining states, except for South Dakota, form the third cluster. They report medium yearly seasonal variability and the geometric mean of geometric standard deviations of $\{\exp \widehat{S_{1,t}}\}$ is 1.14.

South Dakota is the only member of the fourth cluster. It displays extreme yearly seasonal variations, the geometric standard deviation of $\{\exp \widehat{S_{1,t}}\}$ being 1.42.

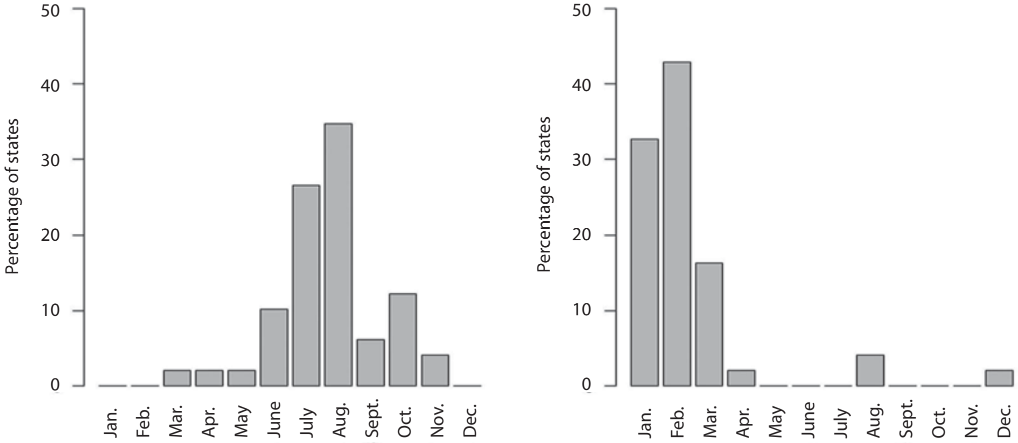
σ_g is plotted for individual states in Figure 8. A clear geographical pattern emerges, related to the above clusters.

Figure 8 Geometric standard deviation of the multiplicative yearly seasonal cycle for each state, except Alaska (1.345) and Hawaii (1). The light and dark colors indicate large and small standard deviations, respectively.



Source: Own construction

Figure 9 Distribution of the maxima (left plot) and minima (right plot) of the yearly multiplicative seasonal component among the states along the year. Percentage of states having a maximum or minimum in a given month is presented on the vertical axis. Hawaii and the District of Columbia are excluded from the plot since they do not exhibit yearly seasonality.



Source: Own construction

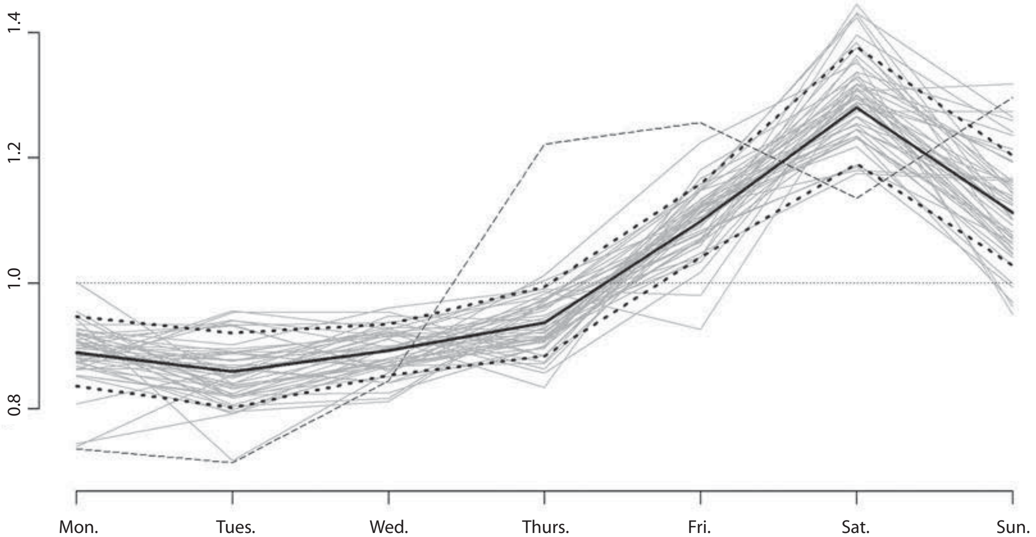
In general, accidents tend to occur more frequently during summer and autumn months as shown in Figures 7 and 9. The latter figure illustrates the distribution of the maxima (left plot) and minima (right plot) of the yearly multiplicative seasonal cycle among the 51 states along the year, the maxima occurring mostly in the summer (July and August) and the minima during winter months (January, February, March).

3.3 Examining the weekly seasonal component

It follows from Table 1 that weekly seasonality is often present in the time series. The only two states that do not exhibit weekly seasonality are Alaska and the District of Columbia. If $\{\widehat{S_{2,t}}\}$ is an estimate of $\{S_{2,t}\}$, then the estimated multiplicative weekly seasonal component in the mean daily number of accidents is $\{\exp \widehat{S_{2,t}}\}$.

Multiplicative weekly seasonal components for the 51 states plotted in Figure 10 are largely similar, except for Alaska and the District of Columbia (see the thin dotted gray line in the figure), which do not show a weekly seasonal pattern, and Rhode Island (thin dashed gray curve), whose pattern slightly differs from that of the other states. The typical pattern has a minimum on Tuesdays and a maximum on Saturdays.

Figure 10 Multiplicative weekly seasonal component for the 51 states. Alaska and the District of Columbia are represented by the thin dotted gray line, Rhode Island by the thin dashed gray curve and the other states by the gray curves. The solid black curve is the geometric mean of individual curves, the two dotted black curves showing the distance of one geometric standard deviation from the geometric mean.



Source: Own construction

DISCUSSION AND CONCLUSIONS

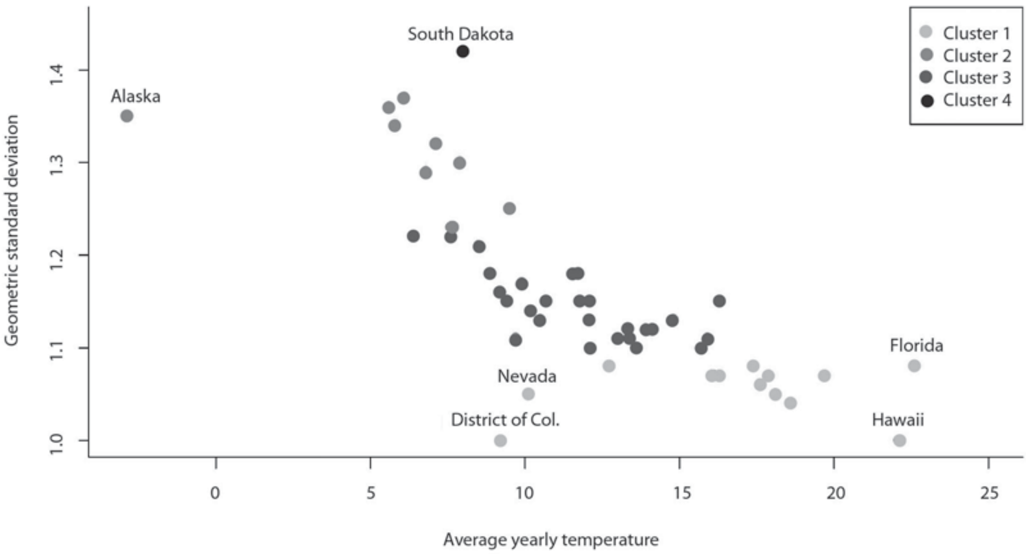
We have constructed daily time series of the number of motor vehicle road accidents with at least one fatality for the individual states of the U.S.A. and the District of Columbia in the period from the beginning of 2006 to the end of 2016. Poisson generalized linear models were used to examine the trend component as well as yearly and weekly seasonal cycles in the time series. Given the long period analyzed, the yearly seasonal component was represented by periodic cubic splines.

Summary statistics on annual averages, the trend component, and yearly and weekly seasonality are presented.

Despite having no intention of exploring causal mechanisms by which the fluctuations of the number of traffic accidents occur, we have some ideas to think about in this context. Specifically, the differences in annual accident rates per 1 000 population between the U.S. states (see Section 1) can be explained by different numbers of motor vehicles and diverse travel motivations and behavior, the above factors resulting in different distances covered by car in a year per 1 000 population. Also, the road network range and quality as well as decisions of the local authorities regarding transport can help to explain the differences (see Peden et al., 2004).

A similar pattern of trend components in the states observed in Section 3.1 suggests that the long-run dynamics of the number of accidents may be governed by some common factors (such as the improved vehicle safety features, the U.S. economic performance and gasoline prices). Longthorne et al. (2010) argue that the decline in the number of accidents is largely due to a decrease in the number of young drivers' car crashes, implying that it might be caused by rising unemployment among the youth. It is obvious that yearly seasonal patterns are related to the geographical location of the states (Section 3.2). Moreover, Figure 11 reveals the relationship between the geometric standard deviation of the multiplicative yearly seasonal pattern $\{\exp \widehat{S_{1,t}}\}$ and the average annual temperature of each state.[15] The states with higher average temperatures tend to have less variable yearly seasonal patterns, whereas those with lower average temperatures report higher variability. This may indicate lower winter traffic volumes (compared to summer ones) in the states with low annual average temperatures since some local roadways are not safe enough to drive on. Summer driving, on the other hand, seems to be more comfortable, which is reflected in higher traffic density levels and accident rates.

Figure 11 Geometric standard deviation of the yearly seasonal cycle plotted against the average annual temperature (in degrees Celsius) for the 51 states



Source: Own construction

[15] Average annual temperatures were obtained from the National Centers for Environmental Information website at: https://www.ncdc.noaa.gov.

The similarity between weekly seasonal patterns (Section 3.3) is explicable by comparable lifestyles in different states, people traveling more by car to see relatives and friends or go on trips, particularly on weekends.

In further research, we will focus on a causal explanation for different accident rates.

ACKNOWLEDGMENTS

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

We thank Karel Helman, PhD, (Private University College of Economic Studies) for kind assistance in proofreading the text.

We thank the anonymous reviewers whose comments helped improve the content of the paper.

References

CAMERON, A. C. AND TRIVEDI, P. K. *Regression analysis of count data*. Cambridge university press, 2013.

DE ANDRADE, B. S., ANDRADE, M. G., EHLERS, R. S. Bayesian GARMA models for count data. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 2016, 1(4), pp. 192–205.

DOBSON, A. AND BARNETT, A. *An Introduction To Generalized Linear Models.* 3rd Ed. Chapman & Hall/CRC, 2008.

DUNSMUIR, W. AND SCOTT, D. The glarma package for observation driven time series regression of counts. *Journal of Statistical Software*, 2015, 67(7), pp. 1–36.

EDWARDS, J. B. Weather-related road accidents in England and Wales: a spatial analysis. *Journal of Transport Geography*, 1996, 4(3), pp. 201–212.

EVERITT, B. AND HOTHORN, T. *An introduction to applied multivariate analysis with R*. Springer Science & Business Media, 2011.

JONES, A. P., HAYNES, R., KENNEDY, V., HARVEY, I. M., JEWELL, T., LEA, D. Geographical variations in mortality and morbidity from road traffic accidents in England and Wales. *Health & place*, 2008, 14(3), pp. 519–535.

KIRKWOOD, T. Geometric means and measures of dispersion. *Biometrics*, 1979, 35, pp. 908–909.

LONGTHORNE, A., SUMBRAMANIAN, R., CHOU-LIN, C. *An Analysis of the Significant Decline in Motor Vehicle Traffic Crashes in 2008* [online]. Technical report No. DOT HS 811 346, U.S. Department of Transportation, National Highway Traffic Safety Administration, 2010. <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/811346>.

LEVINE, N., KIM, K. E., NITZ, L. H. Daily fluctuations in Honolulu motor vehicle accidents. *Accident Analysis & Prevention*, 1995, 27(6), pp. 785–796.

MCCULLAGH, P. AND NELDER, J. A. *Generalized linear models*. Chapman and Hall/CRC, 1989.

NELDER, J. A. AND BAKER, R. J. *Generalized linear models*. John Wiley & Sons, Inc., 1972.

NOFAL, F. H. AND SAEED, A. A. W. Seasonal variation and weather effects on road traffic accidents in Riyadh city. *Public health*, 1997, 111(1), pp. 51–55.

PEDEN, M. et al., eds. *World report on road traffic injury prevention*. Geneva: World Health Organization, 2004.

R CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. Vienna, Austria: R Foundation for Statistical Computing, 2017. <https://www.r-project.org>.

RAMSAY, J. O. AND SILVERMAN, B. W. *Applied functional data analysis: methods and case studies*. New York: Springer, 2002.

RAMSAY, J. O. AND SILVERMAN, B. W. *Functional Data Analysis.* 2nd Ed. Springer, 2005.

WANG, S. *pbs: Periodic B Splines*. R package version 1.1, 2013.