Accident Analysis and Prevention 125 (2019) 336–343

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Forecasting German crash numbers: The effect of meteorological variables

Kevin Diependaele[a,*], Heike Martensen[a], Markus Lerner[b], Andreas Schepers[b], Frits Bijleveld[c,d], Jacques J.F. Commandeur[c,d]

[a] VIAS institute, Brussels, Belgium
[b] Bundesanstalt für Straßenwesen, Bergisch Gladbach, Germany
[c] Stichting Wetenschappelijk Onderzoek Verkeersveiligheid, Den Haag, The Netherlands
[d] Vrije Universiteit, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Road safety
Meteorological effects
Structural time-series model

ABSTRACT

At the end of each year, the German Federal Highway Research Institute (BASt) publishes the road safety balance of the closing year. They describe the development of accident and casualty numbers disaggregated by road user types, age groups, type of road and the consequences of the accidents. However, at the time of publishing, these series are only available for the first eight or nine months of the year. To make the balance for the whole year, the last three or four months are forecasted. The objective of this study was to improve the accuracy of these forecasts through structural time-series models that include effects of meteorological conditions. The results show that, compared to the earlier heuristic approach, root mean squared errors are reduced by up to 55% and only two out of the 27 different data series yield a modest rise of prediction errors. With the exception of four data series, prediction accuracies also clearly improve incorporating meteorological data in the analysis. We conclude that our approach provides a valid alternative to provide input to policy makers in Germany.

1. Introduction

Reliable accident numbers are essential for monitoring road safety. Obtaining such numbers and analyzing them with respect to short- and long-term trends is, however, not a trivial task. The difficulties that road safety agencies are facing in many countries are both practical and theoretical in nature. Apart from various causes of underreporting, police records are often difficult to obtain directly and/or without serious administrative bottlenecks. Publication of official accident numbers thus typically involves considerable time-delays, which is, of course, detrimental to effective policy making.

In Germany, the Federal Highway Research Institute (Bundesanstalt für Straßenwesen; BASt) collects accident numbers from 1991 (for the whole of Germany) and publishes a yearly report on the balance of national accident numbers. The evolution is considered on a monthly basis for several subcategories (27 in total), based on road, user and accident variables (e.g., type of road, age, cyclist, fatal, damage only, influenced by alcohol, etc.). The report is prepared and published in December. At that time, however, preliminary accident numbers have only been released up to August of the running year for most subcategories and maximally up to September. At the same time, these numbers are still subject to up- or downward corrections and final numbers are only released in the next year (typically around June).

To deal with this particular situation, BASt has developed heuristics to predict (a) adjustments between provisional and final data from January to August/September and (b) the evolution of accident numbers in the different subcategories from August/September to December. Throughout the years, these heuristics have been refined as the result of a continuous learning process. In particular, the experience with typical responses of accident numbers to weather conditions have had an important impact. This is not surprising, given the variety of countries where correlations between accidents and weather variables have been demonstrated (see e.g., Bergel-Hayat et al., 2013). These correlations arise from complex underlying dynamics. There are, however, two main sources. The first concerns the well-known fact that certain weather conditions have an impact on the risk of accidents (e.g., Theofilatos and Yannis, 2014). The second type of correlation arises from the impact of weather on risk exposure (e.g., Sabir, 2011; Liu et al., 2017; Theofilatos and Yannis, 2014).

Precipitation, like snow and rain, increases accident risk because of reduced friction between vehicles and the road surface and because visibility is reduced (precipitation itself, splashing water, frozen/fogged windscreens, etc.) (e.g., Brodsky and Hakkert, 1988). Research shows that, certainly in the case of rain, this increase is often not overruled by

[*] Corresponding author.
E-mail address: kevin.diependaele@vias.be (K. Diependaele).

https://doi.org/10.1016/j.aap.2018.07.022
Received 6 February 2018; Received in revised form 22 June 2018; Accepted 15 July 2018
Available online 18 August 2018

compensation behaviour such as reduced speed, less frequent overtaking, etc. (Focant and Martensen, 2014). The correlation with risk exposure arises from relationships between weather conditions and mobility patterns (e.g., Sabir, 2011; Liu et al., 2017; Theofilatos and Yannis, 2014).

When a substantial amount of snow has fallen, all road users may avoid unnecessary trips, while rain, frost and small amounts of snow are thought to reduce mostly the mobility of two-wheelers and pedestrians. The lowering effect on the traffic volume can be so strong, that even with increased accident risk, fewer accidents are observed in bad weather. How specifically the risk- and exposure-related effects combine into a net effect on accident numbers, varies strongly across different weather conditions and road users (e.g., Focant and Martensen, 2014; Sabir, 2011).

In the German scenario, it remains hard to make valid predictions about the evolution of accident numbers, even though the pattern of weather conditions is almost entirely known at the time of the analysis. As mentioned above, the weather influence on risk and exposure generates different and sometimes even opposite effects for different accident types, or even for one and the same accident type. Given the inherent correlation of weather variables (e.g., snowfall and temperature) it is also not straightforward to identify the critical variables and their critical values with respect to accident occurrence. Some variables also clearly interact. ‘Warm and dry’ will have a different impact than ‘warm and wet’ or ‘cold and dry’ or ‘cold and wet’. Weather variables also correlate and/or interact with other sources of variation, such as daylight hours, school/public holidays, alcohol consumption, etc.

In the present work we developed a time-series modelling strategy that quantifies the impact of weather conditions on accident numbers in Germany. It must be emphasized, however, that it was not the goal of this strategy to disentangle the effect of meteorological conditions on accident risks and occurrences. The sole objective was to develop a tool which would improve the accuracy of the year-end forecasts by BASt, concerning various types of national accident/injury numbers. The main ingredients of our approach are (a) a decomposition of historical data according to structural time-series models with seasonal, trend and weather-regression components and (b) a projection of these models together with known weather values to impute missing accident numbers from August/September to December.

This approach allows to disentangle long-term trends (e.g. casualty reduction due to better occupant protection) from seasonal patterns (e.g. variation in crash-occurrence due to changes in day-light patterns or school holidays). For meteorological variables it moreover allows to specifically include deviations from the typical seasonal pattern and interactions between weather variables. Regional variation of meteorological conditions was taken into account to a limited extent. While the accident data were only available at the national level, weather data were considered at a regional level and weighted by population numbers so as to ensure that the conditions in densely populated areas would have more weight on the predictions.

The results are validated against the consolidated accident numbers of the last years and the predictions based on the earlier BASt heuristics.

2. Data and data preparation

2.1. Accident numbers

The accident numbers are divided into 27 different series with monthly count data since 1991 (see Table 1).

For each series there exist a provisional and a final version. The provisional data provide the preliminary counts as they were available for BASt in December of each year. In the majority of these series, the values are systematically missing from August to December. For the first five series in Table 1, the provisional data are available up to September. The final data series provide the completed and adjusted accident counts for the whole year. For any given year, these data are only released during the next year. Hence, for the current study provisional data were available up to August/September 2016 and final data up to December 2015. This is illustrated in Fig. 1 for the total number of police recorded accidents.

Table 1
Overview of the different monthly accident count series since 1991, including descriptive statistics.

Type of crash/victim	Mean	St.Dev.	Q1	Median	Q3
Inj. acc. urban	19155	3735	16284.5	19599.5	22151
All severely injured	7761	2400	5964	7453.5	9451
Killed rural	342.72	139	223	332	451
All injury acc.	29192.41	5520	25228	29251	33614
Killed urban	147.55	60	103	136	177
Killed moto	69	49	17	71.5	108
Inj. acc. alcohol	2167	834	1514.5	2054	2684
All victims	38583.48	7348	33699	38481	44499
Killed 65 +	106.43	29	85	102	122
Killed 25-64	286.21	117	189	277	*381*
Killed cyclists	49	23	33	45	64
Killed 18-24	119.7	59	69	114	163
Inj. acc. rural	8150	1729	6941	8112.5	9672
Inj. acc. BS	2795	636	2238	2809	3363
Killed all	554.86	218	368	533	*718*
Killed caroccup.	321	144	189	317.5	437
Killed pedstrians	80	46	45	68	100
Killed motorveh.	344	149	211	344	462
Killed motorway	64.59	28	43	62	*82*
Killed BS	138.47	61	85	132	190
Killed < 15	19.47	13	8	16	30
Killed goodsveh.	18	6	14	18	*22*
Killed moped	12	7	7	10	15
Inj. acc. motorway	1888	357	1613.75	1848.5	2164
Property damage	8556.76	1589	7530	8491	9549
Killed 15-17	22.66	12	11	21	33
All accidents	193350.15	13714	184994	193025	202089

2.2. Weather data

Raw weather data was downloaded from the Deutscher Wetterdienst (DWD). The data represent daily values since January 1st 1991 for the variables shown in Table 2 (see DWD for a more detailed description).

The data were downloaded from eight weather stations. These stations we identified from a (maximum) surface mapping between (1) fifteen climate zones and their reference weather station (DWD Test Reference Years, TRY, 2004) and (2) the German NUTS 2 regions. This mapping is shown in Table 3. The selected weather stations appear in italics.

The weather data were prepared for the analysis in four steps. First, a number of transformations were applied to reduce the marked skewness in a number of the raw variable distributions (thus, to reduce the influence of atypical values during aggregation). After visual inspection, we decided to apply a square root transformation for variables whose distributions are bounded by zero, i.e., Average wind speed, Maximum wind speed, Precipitation height and Sunshine duration. For percentage data, the natural correlation between mean and variance was countered by applying the arcsine square root transformation (i.e., $\sin^{-1}\sqrt{x/max(x)}$). Specifically, it was applied to Cloud coverage, Relative humidity and the binary precipitation type variables (after calculating the monthly averages $[\sin^{-1}\sqrt{\bar{x}}]$ in the latter cases).

In the second step, an additional set of variables was created, by expressing the original variables as deviations from the monthly station-level averages. For a given weather variable x, the deviation on day i from the mean for month j at station k was expressed as a z -score, i.e., $z_{x_{ijk}} = (x_{ijk} - \bar{x}_{jk})/\sqrt{Var(x_{jk})}$, with $\bar{x}_{jk}$ the sample mean of variable x in month j at station k since 1991 and $Var(x_{jk})$ the sample variance. The variables that were obtained in this way served as an additional set of input data, quantifying how usual or unusual the daily values are for a

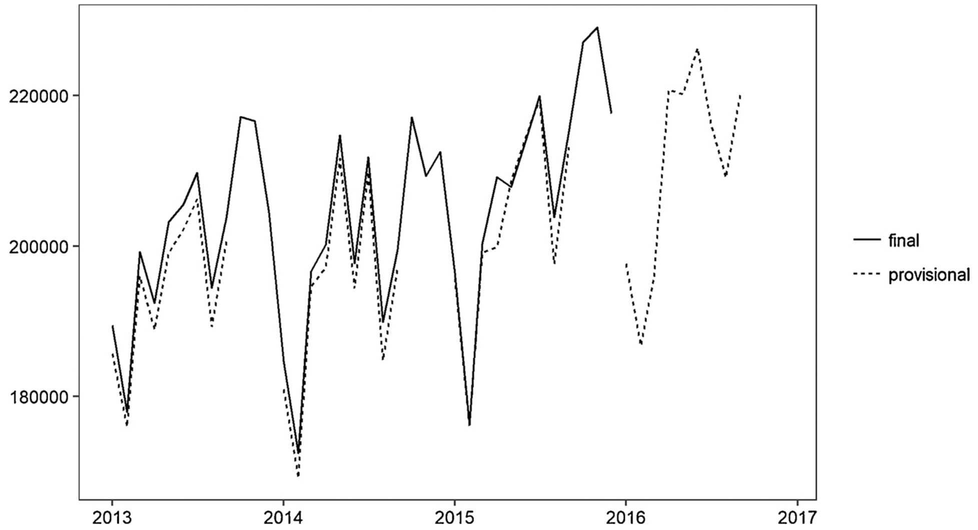

Fig. 1. Illustration of the available data for a given data series. The full series starts in January 1991.

Table 2
Overview of the different weather parameters.

Average temperature (°C; 2 m above ground level)	LUFTTEMPERATUR
Maximum temperature (°C)	LUFTTEMPERATUR_MAXIMUM
Minimun temperature (°C)	LUFTTEMPERATUR_MINIMUM
Minimum temperature at ground level (°C)	LUFTTEMP_AM_ERDB_MINIMUM
Vapor pressure (hPa)	DAMPFDRUCK
Cloud coverage (1/8)	BEDECKUNGSGRAD
Air pressure (hPa)	LUFTDRUCK_STATIONSHOEHE
Relative humidity (%)	REL_FEUCHTE
Average wind speed (m/s)	WINDGESCHWINDIGKEIT
Maximum wind speed (m/s)	WINDSPITZE_MAXIMUM
Precipitation height (mm)	NIEDERSCHLAGSHOEHE
Precipitation types (yes/no)	NIEDERSCHLAGSHOEHE_IND
- no precipitation	
- rain only	
- snow only	
- rain and snow	
Sunshine duration (h)	SONNENSCHEINDAUER
Snow height (cm)	SCHNEEHOEHE

given variable at a given weather station, in the light of all other observations since 1991 in the same month and at the same station.

Thirdly, the daily station-level values for each variable were aggregated into monthly national values. The arithmetic mean of each variable was calculated per station, year and month. This mean was then combined into a national value by calculating a weighted mean across the weather stations. The weights in this calculation were proportional to the yearly population size since 1991 in the NUTS regions that were associated with the weather stations (source: Eurostat).

Finally, the matrix of aggregated weather data was transformed into its principal components (PCs). The matrix is characterized by a naturally high degree of multicollinearity, which implies that a small change in the data can drastically change the pattern of estimated coefficients and the associated error distributions. This is not a problem per se, since it was not within the scope of the current study to link specific patterns in accident counts to individual weather variables. However, to avoid overfitting and losing predictive power, the objective is to predict future accident counts, using only those weather variables that have a proven statistically significant relationship with accident counts. Multicollinearity compromises straightforward choices in this respect. The PCs, by definition, provide orthogonal (uncorrelated) dimensions across all weather variables. For the number of PCs extracted from the matrix of aggregated weather data and the number of PCs retained in the analyses of the crash number series we refer to Section 3.1.

3. Methods of analysis

3.1. Weather regression models

Predictions were based on independent models of the 27 fin. l accident data series. These models took the form of a log-normal

Table 3
Overview of the surface mapping between climate zones and NUTS 2 regions.

Climate zone	Reference station	NUTS 2 regions (DE)
Nordseeküste	Bremerhaven	
Ostseeküste	Rostock-Warnemünde	
Nordwestdeutsches Tiefland	*Hamburg-Fuhlsbüttel*	50, 60, 91, 92, 93, 94, F0
Nordostdeutsches Tiefland	*Potsdam*	30, 40, 80, D5, E0
Niederrheinisch-westfälische Bucht & Emsland	*Essen*	A1, A2, A3, A4
Nördliche & westliche Mittelgebirge - Randgebirge	*Bad Marienberg*	A5, B1, B2, C0
Nördliche & westliche Mittelgebirge - Zentrale Bereiche	*Göttingen*	72, 73
Oberharz & Schwarzwald (mittlere Lagen)	Braunlage	
Thüringer Becken & Saechsisches Huegelland	*Chemnitz*	D2, D4, G0
Südöstliche Mittelgebirge < = 1000 m	Hof	
Erzgebirge, Boehmer, & Schwarzwald > 1000 m	Fichtelberg	
Oberrheingraben & unteres Neckartal	*Mannheim*	12, 13, 71, B3
Schwäbisch-fränkisches Stufenland & Alpenvorland	*Passau*	11, 14, 21, 22, 23, 24, 25, 26, 27
Schwäbische Alb & Baar	Stötten	
Alpenrand & -täler	Garmisch-Partenkirchen	

structural time series model with weather PC regression components. Specifically, the linear predictor included a random level, slope, (monthly) seasonal and irregular effect, but also fixed weather PC effects. The generic definition was as follows:

$$
\begin{array}{l|l}
y_t = \exp(\mu_t + \sum_{j=1}^{[s/2]} \gamma_{jt} + \sum_{k=1}^{r} \beta_k x_{kt} + \varepsilon_t), & \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2) \\
\mu_{t+1} = \mu_t + \nu_t + \eta_t, & \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2) \\
\nu_{t+1} = \nu_t + \xi_t, & \xi_t \sim \mathcal{NID}(0, \sigma_\xi^2) \\
\gamma_{j,t+1} = \gamma_{jt} \cos\lambda_j + \gamma_{jt}^* \sin\lambda_j + \omega_{jt}\,, & \\
\gamma_{j,t+1}^* = -\gamma_{jt} \sin\lambda_j + \gamma_{jt}^* \cos\lambda_j + \omega_{jt}^*\,, & \omega_{jt}, \omega_{jt}^* \sim \mathcal{NID}(0, \sigma_\omega^2) \\
\lambda_j = 2\pi j/s &
\end{array}
$$

with

y_t the observed count at time t for a given series of accident data,

μ_t the trend component, with random disturbance η_{t-1}, slope ν_{t-1} and slope disturbance ξ_{t-1},

γ_{jt} the trigonometric seasonal components with periodicity $s = 12$ and disturbances ω_{jt} and ω_{jt}^*,

[s/2] the largest integer ≤ s/2 (for monthly data [s/2] = 6),

x_{kt} the value of the k -th weather regressor at time t,

β_k the time-invariant coefficient of the k -th weather regressor and

ε_t the additive observation level noise term

For further details concerning this formulation of what is known as the basic structural time series model we refer to (Durbin and Koopman, 2012; Harvey, 1989) as this is beyond the scope of the present paper. All disturbances were modelled as independent Gaussian processes. Models were fit separately for each accident series according to the following procedure. This procedure critically relies on Kalman filtering, mode estimation and importance sampling, as implemented in the R-package KFAS (Helske, 2017; R Core Team, 2017).

Initially, observed counts were log-transformed and models were fit using the regular Kalman filter. Importantly, in this step, any zero-counts were replaced with missing values. Models were fit through the Kalman filter with exact diffuse initialization of state vector a_1 and its associated variance-covariance matrix P_1 (Koopman and Durbin, 2003).

$$
\begin{array}{l|l}
a_1 = (\mu_1\ \nu_1\ \gamma_{1,1}\ \gamma_{1,1}^*\ \gamma_{2,1}\ \gamma_{2,1}^* \cdots \gamma_{6,1}\ \beta_1 \cdots \beta_r)', & \beta_{k,1} = \beta_{k,2} = \cdots = \beta_{k,T} \\
P_1 = \begin{pmatrix} \sigma_{\mu_1}^2 & \cdots & \sigma_{\mu_1\beta_r} \\ \vdots & \ddots & \vdots \\ \sigma_{\mu_1\beta_r} & \cdots & \sigma_{\beta_r}^2 \end{pmatrix} &
\end{array}
$$

The Kalman filter provides the equations to calculate

$$a_{t|t} = E(\alpha_t|y_1,\ \ldots,y_t),$$

$$a_{t+1} = E(\alpha_{t+1}|y_1,\ldots,y_t),$$

$$P_{t|t} = Var(\alpha_t|y_1,\ldots,y_t),$$

$$P_{t+1} = Var(\alpha_{t+1}|y_1,\ldots,y_t)$$

recursively from

$$a_t = E(\alpha_t|y_1,\ldots,y_{t-1}),$$

$$P_t = Var(\alpha_t|y_1,\ldots,y_{t-1})$$

given the values of the disturbance variance parameters σ_ε^2, σ_η^2, σ_ξ^2 and σ_ω^2. The results of the Kalman filter were used in a backward recursion (the so-called state smoothing recursion) to obtain the expected value of α_t, given the entire series, i.e., $\hat{\alpha}_t = E(\alpha_t|y_1,\ldots,y_n)$ and the corresponding variance $V_t = Var(\alpha_t|y_1,\ldots,y_n)$ (see Durbin and Koopman, 2012, for details). Models were optimized with respect to the unknown variance disturbance parameters σ_ε^2, σ_η^2, σ_ξ^2 and σ_ω^2 (see Helske, 2017; Koopman, 2003 for details). For each model, optimization was done ten times with a different random set of starting values for the unknown disturbance variance parameters. During the development of our (automated) procedure, we found that ten replications were sufficient to obtain stable results. The fitted model with the highest likelihood value was retained.

For each series, the model was calibrated with respect to the matrix of weather PC regressors. Initially, for each data series, all weather PCs that explained at least 3% of the variance in the aggregated weather data according to the principal components analysis (i.e., the first seven PCs), were included in the model together with their first order interactions. After fitting this model, all interaction terms with a p-value ≥.01 were removed. The model was refit and subsequently, all simple effects of weather PCs with a p-value ≥.10 and which did not occur in any of the interactions were also excluded. The final model thus contains PCs that yield (a) significant (α = .01) first order interaction(s) or a significant (α = .10) main effect.

Finally, to accommodate zero-counts, the calibrated weather regression models were refit in a generalized linear approach using the untransformed counts with a logarithmic link function. The models were fit through the methods of (1) mode estimation, which finds an approximating linear Gaussian state space model, and (2) importance sampling, which corrects for approximation errors (see Durbin and Koopman, 2012, for details). Starting values for the disturbance variance parameters were taken directly from the best fitting model of the log-transformed counts. The number of simulation runs in the importance sampling procedure was set to 1000, as optimized during the development of our procedure.

3.2. Final-provisional models

The weather regression models cannot be applied directly to predict accident counts for the last three/four months of the year, since the accident counts that are available for January-August/September are only provisional. The weather regression models are fit to the final count data. Feeding provisional counts to these models for the first eight/nine months would yield biased results since there are the marked differences between provisional and final counts (see e.g., Fig. 1). Before applying the weather regression models to make predictions for September/October-December, the provisional data in each series are therefore adjusted through simple models of the following form:

$$
\begin{array}{l|l}
d_t = \mu_t + \sum_{j=1}^{[s/2]} \gamma_{jt} + \varepsilon_t, & \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2) \\
\mu_{t+1} = \mu_t + \eta_t, & \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2) \\
\gamma_{j,t+1} = \gamma_{jt} \cos\lambda_j + \gamma_{jt}^* \sin\lambda_j & \\
\qquad\qquad + \omega_{jt}\,, & \omega_{jt}, \omega_{jt}^* \sim \mathcal{NID}(0, \sigma_\omega^2) \\
\gamma_{j,t+1}^* = -\gamma_{jt} \sin\lambda_j + \gamma_{jt}^* \cos\lambda_j & \\
\qquad\qquad + \omega_{jt}^*\,, & \\
\lambda_j = 2\pi j/s &
\end{array}
$$

with

d_t the observed difference between final and provisional count for a given data series at time t, with random disturbance ε_t,

μ_t the level component, with random disturbance η_{t-1} and

γ_{jt} the j -th trigonometric seasonal component with periodicity s

and disturbances ω_{jt} and ω_{jt}^{*}

To enable the estimation of a seasonal component, the months for which provisional data are systematically missing were removed from each series, resulting in an adjusted periodicity $s < 12$. Theoretically, this implies, however, that we ignore the increased uncertainty due to these missing values. The parameters were estimated following the same procedure as for log-counts described above. The smoothed state vectors $(\hat{\mu}_t\, \hat{\gamma}_{1,t}\, \hat{\gamma}_{1,t}^{*} \cdots \hat{\gamma}_{s/2,t})'$ were used to calculate $\hat{d}_t$ for $t = n + 1, \ldots, n + s$. In the example shown in Fig. 1, $t = n + 1$ corresponds to January 2016 and $t = n + s$ corresponds to September 2016. The predicted difference $\hat{d}_t$ is added to the observed provisional data for $t = n + 1, \ldots, n + s$, which yields the prediction for the final data in this period.

3.3. Weather PC models

After adjusting the provisional data of each series for the running year, they were merged with the corresponding final data since 1991. These series could be fed into the fitted weather regression models to obtain smoothed predictions for the last three/four months, taking into account the available weather PC values. However, before this was done, one last modelling step took place. The reason was that at the time of reporting in December, when accident predictions are needed, months of weather data are only complete up to November. To accommodate this, we decided to impute the missing values through a classical structural time series model, fitted independently to each weather PC. Specifically, these models had the following form:

$$
\begin{array}{l|l}
y_t = \mu_t + \sum_{j=1}^{[s/2]} \gamma_{jt} + \varepsilon_t, & \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2) \\
\mu_{t+1} = \mu_t + \nu_t + \eta_t, & \eta_t \sim \mathcal{NID}\left(0, \sigma_\eta^2\right) \\
\nu_{t+1} = \nu_t + \xi_t, & \xi_t \sim \mathcal{NID}\left(0, \sigma_\xi^2\right) \\
\gamma_{j,t+1} = \gamma_{jt} \cos \lambda_j + \gamma_{jt}^{*} \sin \lambda_j & \\
\qquad + \omega_{jt}, & \omega_{jt}, \omega_{jt}^{*} \sim \mathcal{NID}(0, \sigma_\omega^2) \\
\gamma_{j,t+1}^{*} = -\gamma_{jt} \sin \lambda_j + \gamma_{jt}^{*} \cos \lambda_j & \\
\qquad + \omega_{jt}^{*}, & \\
\lambda_j = 2\pi j / s &
\end{array}
$$

with

y_t the observed value for a given weather PC at time t, with random disturbance ε_t

μ_t the trend component, with random disturbance η_{t-1}, slope ν_{t-1} and slope disturbance ξ_{t-1}

γ_{jt} the j -th trigonometric seasonal component with periodicity $s = 12$ and disturbances ω_{jt} and ω_{jt}^{*}

The parameters were estimated following the same procedure as for log-counts described above. The smoothed state vectors $(\hat{\mu}_t\, \hat{\nu}_t\, \hat{\gamma}_{1,t}\, \hat{\gamma}_{1,t}^{*} \cdots \hat{\gamma}_{6,t})'$ were used to calculate $\hat{y}_t$, for $t = n + 1, \ldots, T$ where n corresponds to November and T to December of the running year.

3.4. Validation

After feeding the fitted weather regression model (see 3.1) with (a) the final and adjusted provisional data (see 3.2) and (b) the observed and imputed weather PC values (see 3.3), the smoothed state vectors $(\hat{\mu}_t\, \hat{\nu}_t\, \hat{\gamma}_{1,t}\, \hat{\gamma}_{1,t}^{*} \cdots \hat{\gamma}_{6,t}\, \hat{\beta}_1 \ldots \hat{\beta}_r)'$ were used to calculate $\hat{y}_t$, for $t = n + 1, \ldots, T$, where n corresponds to the last month for which provisional data were available (i.e., August/September).

The main objective of the present study was to investigate whether these predictions would provide a better approximation of the final accident counts as compared to the predictions from the established BASt heuristics. Hence, the following validation method was applied. For each year between 2000 and 2015 and each accident data series, predictions were generated based on (a) the final data of all previous years, (b) adjusted provisional data for the running year, (c) the observed weather PC values up to November of the running year and (d) the weather PC imputations for December of the running year. The estimated variances of all disturbance parameters were kept constant across all validation years and were taken from each model as fitted to the whole data set. Eventually, the predictions for each month of a given year were summed and compared with the sum of the eventual final data for that year. The same was done for the predictions that were published by BASt at the time. The accuracy of both methods was evaluated by comparing the root mean squared error (RMSE) for each accident data series.

In order to investigate the added value of the weather regression in our approach, we repeated the validation procedure now using the predictions of a log-normal structural time series model without weather PC regression components for the final data.

4. Analysis and results

In this section we report the results of the validation procedure, as described in the previous section. The results of our validation study are presented in Fig. 2. The bars give the RMSEs for each of the accident series, as obtained with (1) the present model-based approach (white bars), (2) the same model-based approach, but without weather PC regression components (grey bars), and (3) the earlier heuristic approach of BASt (black bars). The number in parentheses behind each accident series is the number of weather effects (main effects and interactions) that were retained in the model fitting stage of the final weather regression model for that series. It is beyond the scope of this paper to provide a detailed overview of each of the models and the exact nature of the weather effects therein.

In comparison with the year totals that were previously published by BASt, and in terms of the criterion that was optimized in this study in order to improve upon the predictions obtained by BASt, i.e. RMSE, we see that predictions become more accurate in 23 of the 27 accident data series, with accuracies increasing up to 55 percent, see Fig. 2. In four cases, there was either no difference or performance dropped slightly. For the number of fatalities on motorways and the fatally injured users of good vehicles, the accuracy did not change. For the total number of killed road users and the number of killed road users between 25 and 64 years old, there were slight drops in performance (i.e., -8% and -6%, respectively).

With respect to the weather PCs, we see that by including these, we obtain higher prediction accuracies in 20 of the 27 data series. The gain in accuracy is largest for injury accidents in an urban environment (42%). The performance did not change with or without weather regressors for (1) the fatalities under 15 years, (2) fatally injured users of goods vehicles and (3) fatally injured users of mopeds. But we also see that for these series, few weather effects were retained during model fitting (1, 3 and 1, respectively). Predictions appear to be somewhat better without weather variables in the case of (1) injury accidents on motorways, (2) property damage only accidents, (3) fatalities between 15 and 17 years and (4) the total number of accidents. Despite this, a reasonable number of weather effects were retained during model fitting.

Inspection of the Mean Percentage Errors (MPE) for the 27 series in all models displayed in Fig. 3 shows that the final analyses including weather variables tend to underfit the data more often than they overfit: 17 as opposed to 10 times respectively. We also see that – in terms of MPE – the final model with weather variables only outperforms

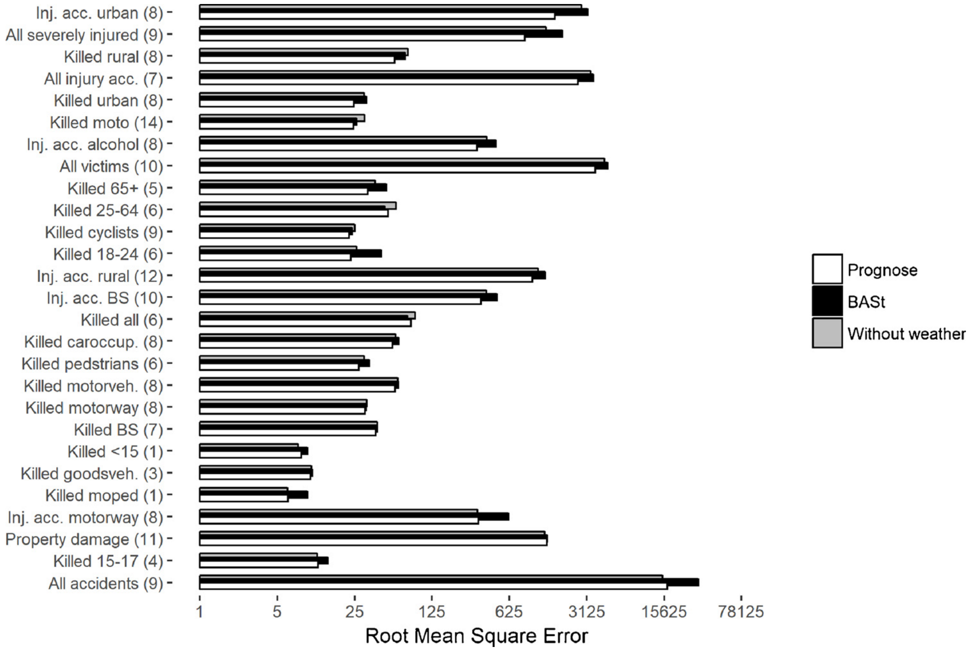

Fig. 2. Summary of the validation results: RMSE.

the BASt results in 16 out of 27 cases. It should be borne in mind however that the aim of the present study was to optimize the RMSE, not the MPE.

Finally, in Fig. 4 we show the predictions for the total number of accidents according to the three scenarios in 2015, and the actual observed count for this series in 2015. Clearly, the weather regression model yields the most accurate prediction for this series, followed by the model without weather variables, while the prediction of the heuristic BASt model is the least accurate of the three.

5. Discussion

Many road safety agencies are confronted with the situation where there are important structural delays in the publication of official national accident counts. Nevertheless, analyses based on recent data are critical for efficient policy making. In this study we have dealt with the task of predicting year totals based on preliminary monthly counts of different accident/injury types, available from January to August/September for the particular case of Germany. The main ingredient of our approach was a structural time series model, trained on all monthly count data since 1991 and including meteorological predictors. In comparison with the earlier heuristic approach used by BASt, the results of our validation study indicate important gains in prediction accuracy, up to 55% in the case of severe injury counts. Only two out of 27 data series showed a modest (-6%, -8%) drop in prediction accuracy. The main conclusion is therefore that our model-based approach is a valid alternative to provide input to policy makers in Germany when only preliminary accident data are available for the running year.

It is important to keep in mind that the present modelling approach was specifically designed with a practical purpose in mind, i.e., a tool to assist BASt in their yearly reporting on the evolutions in accidents and injuries. All modelling and prediction steps were implemented in a

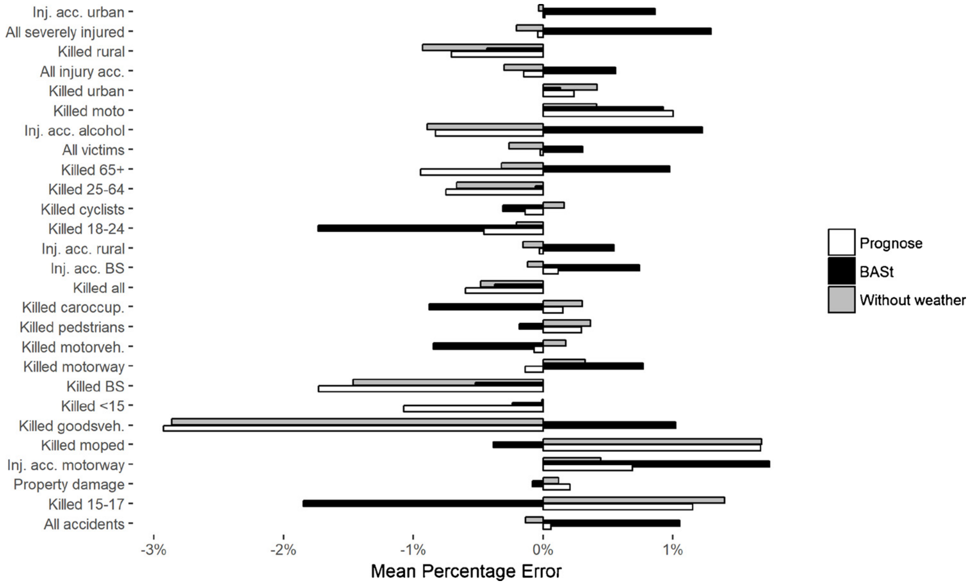

Fig. 3. Summary of the validation results: MPE.

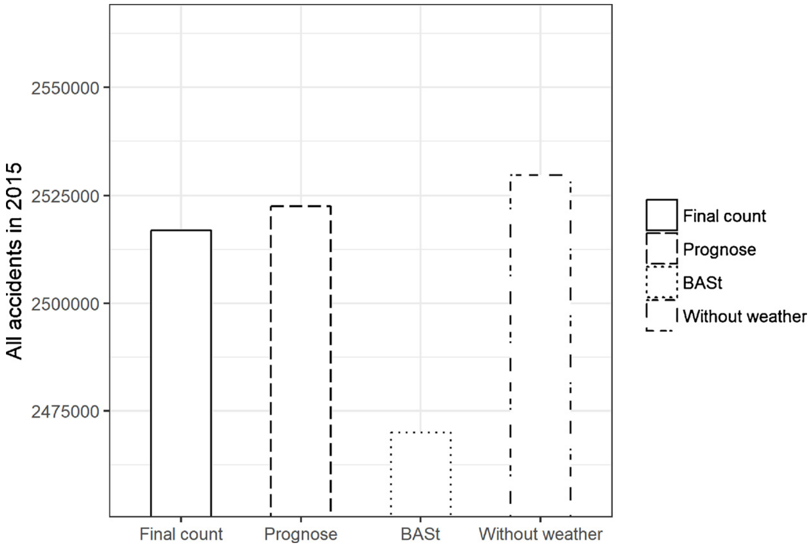

Fig. 4. Predictions for all accidents in 2015, and actual counts (first bar on the left).

stand-alone application, which guided several practical decisions. The treatment of weather effects had to be automated and harmonized across the 27 different data series. It is clear that the performance of each individual model could be improved by a thorough manual checking and analysis of the specific nature of the weather effects.

Nevertheless, our results confirm that the number of accidents has an important association with weather conditions (see, e.g., Bergel-Hayat et al., 2013). Moreover, the results demonstrate that using known weather values as regressors increases the prediction accuracy of accident/injury counts. When weather variables were omitted from the current models, we see an overall drop in prediction performance, with only few exceptions. In four cases we see that there was a slightly better performance without weather variables, despite the fact that several significant relationships with weather variables were identified. The latter seems to indicate that there is still room for improvement in order to capture more complex dynamics in the relationship between weather conditions and road safety. We discuss this in more detail below. Compared to the earlier heuristic approach, however, we see that there is a gain in prediction accuracy, even for those cases where weather regression did not improve the performance.

Two major methodological issues in the present study were how to resolve 1) the temporal discrepancy between the weather data (available on a daily basis) and the crash/victim data (available on a monthly basis), and 2) the spatial discrepancy between the weather data (available for eight different regional weather stations) and the crash/victim data only available at the national level.

Unlike in earlier studies, the present methodology does not provide direct insights in the actual nature of weather effects on road safety. Daily weather data from individual weather stations were transformed and scaled. Next, they were aggregated up to the level of monthly national values. Regional population densities were used as weights during this aggregation to exploit the regional variations in weather conditions in predicting accident occurrence at the national level. Through these transformations, the resulting values cannot easily be traced back to familiar meteorological patterns. Moreover, the current model terms concern the principal components of the transformed weather data. Given the inherent strong correlation between weather variables, principal components allow to obtain coefficients that are robust to new data or small changes in the input data. At the same time, it is straightforward to exclude components which yield no significant relationship with the accident/injury counts and thus avoid over-fitting and poor generalizability. Similar approaches have been applied in regular regression modelling with large quantities of correlated covariates such as in genetics (e.g., Bair et al., 2006).

The use of a principal components regression approach by no means excludes the possibility of interpreting the weather effects. In fact, we believe it is more insightful to identify latent dimensions in the relationship between weather and road safety, instead of focusing on direct physical measures, as the former are better candidates to reveal patterns of meteorological values or ‘weather scenarios’ that have critical implications for road safety. Performing such an analysis was beyond the scope of the present study, but could provide new theoretical insights and a basis for fine tuning the current approach. With respect to further developments, we believe there is ample room for improvements to capture more complex dynamics in the relationship between weather conditions and road safety. In the current approach we have gone as far as including two-way interactions and, next to absolute values, including expressions of values as month- and station-specific deviations. In this way, it was theoretically possible to capture, for instance, differences in the effect between exceptional rainfall with low and high temperatures.

Although the ‘exceptionality’ transformation introduced a specific form of non-linearity, prediction accuracies might gain significantly from other forms of non-linear weather effects. A number of authors has pointed out the importance of specific temporal patterns. For instance, after a long dry period, rain can disproportionally increase risk because dust and oil form a slippery film on the road surface (e.g., Eisenberg, 2004). It remains challenging, however, to define any non-linear patterns a priori, especially since different types of accidents/injuries might be sensitive to different types of patterns. In addition, the patterns should be relevant at the scale of monthly nationwide aggregations (a heavy local summer thunderstorm might not be).

It is also important to consider the true nature of weather effects. As already mentioned in the introduction, there are two main sources of correlation with road safety. The first concerns the fact that certain weather conditions have an impact on the risk of accident occurrence, e.g., by reducing visibility. The second type of correlation arises from the impact on risk exposure, i.e., by altering the traffic volume for given type(s) of road users (e.g., motorcyclists). In the current framework, there is no distinction between these two types of correlation. This is clearly suboptimal for the quality of predictions when the same weather conditions generate opposite effects on risk and exposure for a given type of accidents/injuries. Such difficulty could be addressed by incorporating weather effects in the latent risk modelling framework (Bijleveld et al., 2008) where risk and exposure are explicitly treated as different concepts. However, this multivariate approach critically depends on the availability of relevant historical exposure data (e.g., pedestrian counts) which can be linked to historical weather data.

6. Conclusions

Based on the clear gains in prediction accuracies, we conclude that the present structural time-series modelling approach provides a valid alternative to predict the evolution of accident/injury numbers at the end of the year in Germany. Our study also confirms that it is important to view short-term evolutions in crash statistics in the light of weather conditions. Structural time series models with meteorological predictors are a powerful and convenient tool to capture these relationships. The current models could still gain in prediction accuracy by elaborating on more complex dynamics of weather conditions and by including different sets of predictor variables, with proven relationships with road safety, such as economic variables. However, in the present context, additional variables can only be of practical use if their availability does not suffer from serious structural delays.

References

Bair, E., Hastie, T., Debashis, P., Tibshirani, R., 2006. Prediction by supervised principal components. J. Am. Stat. Assoc. 101, 119–137.

Bergel-Hayat, R., Debbarha, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: weather effects. Accid. Anal. Prev. 60, 456–465.

Bijleveld, F.D., Commandeur, J.J.F., Gould, P.G., Koopman, S.J., 2008. Model-based measurement of latent risk in time series with applications. J. R. Stat. Soc. A 171, 265–277.

Brodsky, H., Hakkert, A.S., 1988. Risk of a road accident in rainy weather. Accid. Anal. Prev. 20 (3), 161–176.

Durbin, J., Koopman, S.J., 2012. Time Series Analysis by State Space Methods, 2nd ed. Oxford University Press.

Eisenberg, D., 2004. The mixed effects of precipitation on traffic crashes. Accid. Anal. Prev. 36, 637–647.

Focant, N., Martensen, H., 2014. Are There More Accidents in the Rain? Exploratory Analysis of the Influence of Weather Conditions on the Number of Road Accidents in Belgium. Belgian Road Safety Institute, Brussels.

Harvey, C., 1989. Forecasting, Structural Time Series Models and the Kalman Filter. Cambridge University Press.

Helske, J., 2017. KFAS: exponential family state space models in R. J. Stat. Softw. 78 (10).

Koopman, S.J., 2003. Disturbance smoother for state space models. Biometrika 80, 117–126.

Koopman, S.J., Durbin, J., 2003. J. Filtering and smoothing of state vector for diffuse state space models. J. Time Ser. Anal. 24, 85–98.

Liu, C., Susilo, Y.O., Karlstr¨om, A., 2017. Weather variability and travel behaviour – what we know and what we do not know. Transp. Rev. 37 (6), 715–741.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sabir, M., 2011. Weather and Travel Behaviour. Vrije Universiteit Amsterdam, Amsterdam.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256.