

# **Zeitreihenmodelle mit meteorologischen Variablen zur Prognose von Unfallzahlen**

**Berichte der  
Bundesanstalt für Straßenwesen**

**Mensch und Sicherheit Heft M 276**

The logo consists of the word "bast" in a bold, lowercase, green sans-serif font. The letters are slightly shadowed, giving it a 3D appearance. The logo is positioned in the bottom right corner of the page, within a white vertical bar that runs along the right edge of the cover.

# **Zeitreihenmodelle mit meteorologischen Variablen zur Prognose von Unfallzahlen**

von

Heike Martensen  
Kevin Diependaele

Belgian Road Safety Institute (BRSI)  
Brüssel

**Berichte der  
Bundesanstalt für Straßenwesen**

Mensch und Sicherheit Heft M 276

**bast**

Die Bundesanstalt für Straßenwesen veröffentlicht ihre Arbeits- und Forschungsergebnisse in der Schriftenreihe **Berichte der Bundesanstalt für Straßenwesen**. Die Reihe besteht aus folgenden Unterreihen:

A - Allgemeines  
B - Brücken- und Ingenieurbau  
F - Fahrzeugtechnik  
M - Mensch und Sicherheit  
S - Straßenbau  
V - Verkehrstechnik

Es wird darauf hingewiesen, dass die unter dem Namen der Verfasser veröffentlichten Berichte nicht in jedem Fall die Ansicht des Herausgebers wiedergeben.

Nachdruck und photomechanische Wiedergabe, auch auszugsweise, nur mit Genehmigung der Bundesanstalt für Straßenwesen, Stabsstelle Presse und Öffentlichkeitsarbeit.

Die Hefte der Schriftenreihe **Berichte der Bundesanstalt für Straßenwesen** können direkt bei der Carl Schünemann Verlag GmbH, Zweite Schlachtpforte 7, D-28195 Bremen, Telefon: (04 21) 3 69 03 - 53, bezogen werden.

Über die Forschungsergebnisse und ihre Veröffentlichungen wird in der Regel in Kurzform im Informationsdienst **Forschung kompakt** berichtet. Dieser Dienst wird kostenlos angeboten; Interessenten wenden sich bitte an die Bundesanstalt für Straßenwesen, Stabsstelle Presse und Öffentlichkeitsarbeit.

Ab dem Jahrgang 2003 stehen die **Berichte der Bundesanstalt für Straßenwesen (BAST)** zum Teil als kostenfreier Download im elektronischen BAST-Archiv ELBA zur Verfügung.  
<http://bast.opus.hbz-nrw.de>

## Impressum

**Bericht zum Forschungsprojekt FE 82.0633/2015:**  
Unfallprognose

**Fachbetreuung**  
Markus Lerner

**Herausgeber**  
Bundesanstalt für Straßenwesen  
Brüderstraße 53, D-51427 Bergisch Gladbach  
Telefon: (0 22 04) 43 - 0

**Redaktion**  
Stabsstelle Presse und Öffentlichkeitsarbeit

**Druck und Verlag**  
Fachverlag NW in der  
Carl Schünemann Verlag GmbH  
Zweite Schlachtpforte 7, D-28195 Bremen  
Telefon: (04 21) 3 69 03 - 53  
Telefax: (04 21) 3 69 03 - 48  
[www.schuenemann-verlag.de](http://www.schuenemann-verlag.de)

ISSN 0943-9315  
ISBN 978-3-95606-361-9

Bergisch Gladbach, November 2017

## Kurzfassung – Abstract

### Zeitreihenmodelle mit meteorologischen Variablen zur Prognose von Unfallzahlen

Die Bundesanstalt für Straßenwesen (BASt) bringt zum Ende jeden Jahres eine Prognose der Unfall- und Verunglücktenzahlen des noch laufenden Jahres heraus, um so über die Entwicklung der Verkehrssicherheit in Deutschland Bilanz ziehen zu können. Dabei wird das Unfallgeschehen nach dem Schweregrad der Konsequenzen, der Ortslage sowie Alter und Art der Verkehrsbeteiligung der Verunglückten in 27 Zeitreihen unterteilt. Zu diesem Zeitpunkt sind die Daten lediglich für die ersten acht oder neun Monate erhältlich. Um Bilanz zu ziehen, werden die Anzahlen der letzten drei oder vier Monate prognostiziert. Gesamtziel des hier beschriebenen Forschungsvorhabens ist die Optimierung der jährlichen Unfallprognosen durch Anwendung von strukturellen Zeitreihenmodellen, bei denen die Vorhersagen aus dem Trend der vorliegenden Monate, und der Dynamik der vorhergehenden Jahre abgeleitet werden. Um dem Einfluss der Witterungsverhältnisse Rechnung zu tragen, werden dabei meteorologische Variablen in das Vorhersagemodell aufgenommen.

Um die Modelle zu testen, werden die endgültigen Daten der letzten 15 Jahre jeweils aus den vorläufigen Daten der ersten Monate vorhergesagt und mit den tatsächlich beobachteten endgültigen Unfall- und Verunglücktenzahlen verglichen. Die Resultate zeigen, dass im Vergleich zu den bisherigen Vorhersagen mithilfe der hier vorgestellten Modelle die Vorhersagen für 25 der 27 Reihen präziser werden. Lediglich zwei Reihen zeigen einen leichten Anstieg des Vorhersagefehlers.

Beim Vergleich von Modellen mit und ohne meteorologischen Variablen zeigt sich, dass 23 der 27 Reihen besser vorhergesagt werden können, wenn man das Wetter berücksichtigt. Neben der verbesserten Vorhersage ermöglicht die Aufnahme der Wettervariablen auch eine Einschätzung, wie groß der Einfluss der Witterungsgegebenheiten auf das Unfallgeschehen ist.

Es zeigt sich also, dass die Anwendung von strukturellen Zeitreihenmodellen und die Berücksichtigung von meteorologischen Variablen zu einer deutlichen Verbesserung der Vorhersagegenauigkeit führen.

Die Verbesserung der Vorhersagen durch die Aufnahme von Wettervariablen bestätigt nochmals den Einfluss der Witterungsumstände auf das Unfallgeschehen.

### Time-series models with meteorological variables to forecast accident figures

At the end of each year, the German Federal Highway Research Institute (BASt) draws the balance of the road-safety development by forecasting the accident and casualty numbers of the closing year. They describe the development of 27 time-series 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. 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 which derive the forecasts for the remaining months from the trend observed in the first months of the year and the dynamics of the earlier development. To take the weather conditions into account, meteorological variables are included into the forecasting model.

To test the models, the final data of the last 15 years were predicted from the preliminary data of the first months for each year. These predictions were compared to the actually observed final numbers of accidents and casualties. The results show that, compared to the earlier heuristic approach, 25 out of the 27 time-series are forecasted more precisely by the models presented. Only two series show a slightly increasing prediction error.

When comparing models with and without meteorological variables, 23 out of 27 series were predicted more accurately when taking the weather into account. Apart from increasing the forecasting precision, the inclusion of meteorological variables also allows estimating to what extent changes in the observed numbers of accidents and casualties can be attributed to the specific weather condition in a particular month.

We conclude that structural time series modelling and the inclusion of meteorological variables clearly

improves the forecasting precision in the year-end prognosis of the German accident and casualty numbers. The improved prediction due to the inclusion of meteorological variables confirms the dependence of these numbers on the weather condition throughout the year.

## Summary

### Time-series models with meteorological variables to forecast accident figures

### Objective

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 crashes. This results in 27 series of monthly accident or casualty numbers which are at the time of publishing available only as preliminary numbers for the first 8 or 9 months of the year. To make the balance for the whole year, the last 3 or 4 months are forecasted. The objective of this study is to improve these forecasts by using Structural Time Series Models and by introducing meteorological variables into the model.

### Background

Meteorological variables like rain and snow but also warm temperatures and sunshine can influence the number of accidents in two ways. 1) For the accident risk, factors like vision and visibility and a possibly reduced friction can lead to an increased accident risk, but road users' compensation for these problems can also lead to a reduction of risk. 2) The weather also has a strong effect on the exposition to risk, because it influences the number of trips undertaken as well as the choice of traffic mode (e. g., walking or cycling in fine weather, using the car with rain or cold weather). In particular the traffic volume of vulnerable road users is strongly affected by the weather condition. In the present study, however, we did not differentiate between risk and exposure effects, as the focus was not so much on understanding how the weather effects come about, but rather to use them to improve the end-of-year forecasts.

Such an improvement can take place in two ways: 1) Exceptional weather conditions in the first months of the year can affect the observed number of accidents or casualties. The inclusion of these conditions into the model helps to disentangle true

trend changes from weather effects. As an example, a reduced number of motorcycle casualties due to cold and rainy weather will most likely go back up as soon as the weather goes back to normal. It should therefore not be interpreted as a permanently decreasing trend that would also affect the future development of the casualties. 2) Knowing the weather in the period that has to be forecasted can help to make more precise forecasts.

In their heuristic end-of-year prognosis, BAST experts have always taken into account textual description of the weather for each month. The objective of the present project was to formalize the inclusion of meteorological variables into the forecasting process.

### Methodology

For 27 time series of monthly accident or casualty numbers, three modelling steps were undertaken. (1) The final numbers for the first 8 or 9 months of the current year are estimated from the preliminary ones, modelling the number of not yet registered accidents from the differences between final and preliminary numbers in previous years. (2) The trend, the typical seasonal variation, and the relation between meteorological variables and accident or casualty numbers was modelled on the basis of the final data since 1991 (for the first months of the current years, these were the estimated "final" numbers from Step 1. (3) The trend and seasonal variation estimated in Step 2, as well as the value of the weather variables in the last months of the current year are used to forecast the accident and casualty numbers of these months.

The meteorological predictors are derived from 8 German Weather stations; each considered representative for a climate zone (BBR, 2011). In Aggregation the data from each station is weighted by the population of its climate zone. 16 meteorological variables – concerning temperature, precipitation (type and amount), sunshine duration and cloud coverage, air pressure, humidity and wind – were pre-processed in several steps. Additionally to the observed daily values, their Z-scores (their deviations from the monthly station-level averages) were included to quantify how usual or unusual the observed values were. Z-scores and original values are aggregated and modelled in a principal component analysis to get independent factors which are then included as predictors into

the 27 time-series models. To calibrate the models, the resulting components and their first order interactions are at first all included and subsequently removed if they do not show a significant contribution.

## Model evaluation

The procedure is evaluated over the last 15 years. For each year the forecast that the model would have made on the basis of the preliminary data for the first 9 months is compared to the actually observed final number. The forecasting errors were aggregated over 15 years by calculating RMSEs (Root Means Squared Errors). The RMSEs produced by the time-series models were compared to those achieved by the forecasts of the BAST team. The increase of the forecasting precision was calculated by

$$\%Improvement = (RMSE_{BAST} - RMSE_{Model}) / RMSE_{BAST}$$

The results show that, compared to the earlier heuristic approach, 25 out of the 27 time-series are forecasted more precisely by the models presented. The prediction error has been reduced up to 55%. Only two series show a slightly increasing prediction error.

When comparing models with and without meteorological variables, 23 out of 27 series were predicted more accurately when taking the weather into account.

## Output

The main output of the project is a stand-alone application, in which reading and processing weather data from the 8 Stations, as well as the calibration and estimation of the models is automated and harmonized across the 27 different accident and casualty series. By including the calibration process into this automatic procedure, the models will be up-dated to changing weather- or accident-constellations during the forecasts in the years to come.

Additional to the forecasted values, for each series the estimate of the monthly “weather effect” is provided. To achieve this alternative models are calculated for each series and each month in which all preceding processes are identical, except that the meteorological values of the month in question are replaced by their long-term mean. The ratio

between both estimates is an estimate of the increase or decrease in the observed numbers of accidents and casualties that can be attributed to the specific weather of that month.

Next to the German language report, a Technical Annex (in English) gives a detailed description of all procedures involved.

## Conclusion

Based on the clear gains in prediction accuracies, we can conclude that the present structural time-series modelling approach provides a valid alternative to predict the evolution of accident/casualty numbers at the end of the year in Germany. Our study also confirms that it is important to view evolutions in accident statistics in the light of weather conditions. Structural time series with meteorological predictors are a powerful and convenient tool to capture these relationships.

## Inhalt

<b>Tabellen-/Abbildungsverzeichnis</b> .....	8	<b>3 Resultate</b> .....	22
<b>1 Problemstellung</b> .....	9	3.1 Vorhersagequalität .....	22
1.1 Daten des deutschen Unfallgeschehens .....	9	3.2 Outputs per Zeitreihe .....	23
1.2 Schätzungen der BAST .....	10	3.2.1 Tabellarisch .....	23
1.3 Zeitreihenmodelle zur Vorhersage von Unfallhäufigkeiten .....	11	3.2.2 Grafischer Output .....	24
1.3.1 State Space Analyse .....	11	<b>4 Diskussion</b> .....	25
1.3.2 Vorhersagen .....	12	4.1 Konstruktionsprinzipien .....	25
1.4 Witterungseinflüsse .....	12	4.2 Evaluation der Modelle .....	26
1.4.1 Meteorologische Variablen als Prädiktoren .....	13	4.3 Einschränkungen und Ausblick .....	26
1.4.2 Räumliche Variation .....	15	<b>5 Schlussfolgerung</b> .....	27
<b>2 Methode</b> .....	15	<b>6 Literatur</b> .....	27
2.1 Strukturelle Zeitreihen Modelle .....	15	<b>Anhang</b> .....	29
2.1.1 Allgemeine Struktur .....	15		
2.1.2 Schätzung .....	15		
2.1.3 Extrapolation der vorläufigen Daten .....	16		
2.1.4 Das Wettermodell .....	16		
2.2 Klimadaten .....	16		
2.2.1 Rohdaten .....	16		
2.2.2 Klimazonen .....	17		
2.3 Wettervariablen und ihre Transformation .....	19		
2.3.1 Schiefekorrektur .....	19		
2.3.2 Z-Transformation .....	20		
2.3.3 Aggregation .....	21		
2.3.4 Hauptkomponenten-Analyse .....	21		
2.3.5 Vervollständigung der Wetterdaten (Imputation) .....	21		
2.4 Selektion der Prädiktoren .....	21		
2.5 Wettereffekte .....	22		



## Tabellenverzeichnis

- Tabelle 1: Zeitreihen im deutschen Unfallgeschehen
- Tabelle 2: Deutsche Klimazonen mit Referenzstationen (TRY 2004) und die zugewiesenen NUTS2-Regionen
- Tabelle 3: Meteorologische Variablen aufgenommen in die Modellierung
- Tabelle 4: Vorhersagegenauigkeit (RMSE) per Reihe für das Wettermodell, die Vorhersagen des BAST Teams (BAST) und einem Modell ohne Wettervariablen
- Tabelle 5: Tabellarischer Output am Beispiel der Reihe G00 (Getötete Insgesamt)

## Abbildungsverzeichnis

- Bild 1: Endgültige vs. vorläufige Daten (Detailedaten): mittlere Abweichung (%) 2001-2014
- Bild 2: Getötete außerorts ohne BAB (GO2). Oberer Bildrahmen: Endgültige (rot) und vorläufige (blau) Daten. Unterer Bildrahmen: Differenz zwischen endgültigen und vorläufigen Daten
- Bild 3: Vorhersage bis 2025 der Verletzten pro Quartal in Belgien mit Konfidenzintervall (68%) (MARTENSEN, BERGEL-HAYAT, in Vorbereitung)
- Bild 4: Effekte von Witterungseinflüssen auf die Anzahl der verletzten Unfallopfer in Belgien (2003-2014). MARTENSEN, FOCANT & DIEPENDAELE, 2016
- Bild 5: Unfälle mit Verletzten mit Motorradbeteiligung (Belgien 2010-2014). Modellierung ohne meteorologische Prädiktoren (obere Grafik) und mit meteorologischen Prädiktoren (untere Grafik) (MARTENSEN & BERGEL-HAYAT, IRTAD, 2017)
- Bild 6: TRY-Regionen für Deutschland (BBR, CEC, DWD: Aktualisierte und erweiterte Testreferenzjahre von Deutschland für mittlere, extreme und zukünftige Witterungsverhältnisse. Projektbericht 2011)
- Bild 7: Zuordnung der Deutschen NUTS2 Regionen zu den Klimazonen der TRY-Studie
- Bild 8: Verteilungsform vor und nach der Transformation von rechtsschiefen Variablen
- Bild 9: Verteilungsform vor und nach der Transformation von linksschiefen Variablen
- Bild 10: Verteilungsform vor und nach der Transformation von Indikatorvariablen
- Bild 11: Grafischer Output zur Evaluation der Wittereffekte am Beispiel der Reihe G00 (Getötete Insgesamt)
- Bild 12: Grafischer Output zur Evaluation der Vorhersagegenauigkeit am Beispiel der Reihe G00 (Getötete Insgesamt)

# 1 Problemstellung

Die Entwicklung der Anzahl der Unfälle und der Verunglückten im Straßenverkehr sind wichtige Indikatoren, um den Fortschritt in der Verkehrssicherheit zu überwachen (BERGEL-HAYAT & ZUKOWSKA, 2015; DUPONT et al., 2014; MARTENSEN & DUPONT, 2010; DUPONT & MARTENSEN, 2012). So bringt auch die Bundesanstalt für Straßenwesen (BASt) zu jedem Jahresende eine Bilanz der Entwicklung der Verkehrssicherheit in Deutschland in Form einer Prognose der Unfall- und Getöteten-Zahlen des noch laufenden Jahres (LERNER, SCHEPERS, PÖPPEL-DECKER, LEIPNITZ & FITSCHEN, 2013, 2014) heraus. Zum Zeitpunkt der Erstellung sind dabei lediglich die Daten von Januar bis August erhältlich, und die letzten Monate werden extrapoliert aus dem Trend, der sich im ersten Teil des Jahres abzeichnete.

Die vorliegende Studie dient der Optimierung dieser Prognose. Es muss betont werden, dass die entwickelten Verfahren zwar auf inhaltlichen Konzepten und theoretischen Erwägungen basieren, ihr Ziel ist es aber nicht, das Wissen dazu zu erweitern, sondern eine effiziente Vorhersage der Jahresenddaten zu ermöglichen.

## 1.1 Daten des deutschen Unfallgeschehens

Die zu analysierenden Daten betreffen monatliche Anzahlen von Unfällen und den dabei Verunglückten. Diese Anzahlen werden nach einer Reihe von Kriterien in verschiedene Zeitreihen eingeteilt.

In Bezug auf die Unfallschwere werden vier Kategorien unterschieden:

- alle polizeilich erfassten Unfälle,
- schwerwiegende Unfälle mit Sachschaden,
- Unfälle mit Personenschaden,
- Unfälle mit Getöteten.

Bei Unfällen, durch die Personen zu Schaden kamen, wird auch unterschieden nach Ortslage:

- innerorts,
- außerorts (ohne BAB),
- Bundesautobahnen,
- Bundesstraßen außerorts

und danach, ob einer der Beteiligten alkoholisiert war.

Die bei Verkehrsunfällen Verunglückten werden ebenfalls nach dem Schweregrad eingeteilt:

- Getötete,
- Schwerverletzte.

Die Getöteten werden nach Altersgruppen unterschieden:

- Kinder: 0-14,
- Jugendliche: 15-17,
- Junge Erwachsene: 18-24,
- Erwachsene: 25-64,
- Senioren: 65+

und nach der Art der Verkehrsbeteiligung:

- alle Kraftfahrzeuge (außer Zweiräder),
- Güterkraftfahrzeuge,
- Personenkraftwagen,
- Motorrad,
- Moped/Mofa,
- Fahrrad,
- Fußgänger und

nach der Ortslage des Unfalls. Hieraus ergeben sich 27 Zeitreihen (Tabelle 1).

Für jede Zeitreihe gibt es vorläufige und endgültige Daten. Die endgültigen Daten sind für alle Jahre erhältlich außer für das laufende. Die endgültigen Daten des Vorjahres werden jeweils im Juni des Folgejahres veröffentlicht. Zum Zeitpunkt der Jahresendprognosen (also Anfang Dezember) sind vorläufige Daten bis August bzw. September erhältlich. Die Verfügbarkeit und Qualität der vorläufigen Daten ist abhängig von der Art der Unfallreihe.

Die Eckdaten (Zeitreihen U00, UP0, USS, G00, und V00) betreffen die Gesamtanzahl der Unfälle eines bestimmten Schweregrades pro Monat. Diese Daten basieren auf der Erstregistrierung der Unfälle und sind zum Zeitpunkt der Jahresendprognose erhältlich bis einschließlich September. Sie weichen im Mittel weniger als 1 % von den endgültigen Daten ab.

Die Detaildaten (alle anderen Zeitreihen) basieren auf den ausgefüllten Unfallformularen. Diese Daten

	Beschreibung
U00	Polizeilich registrierte Unfälle insgesamt
UP0	Unfälle mit Personenschaden insgesamt
USS	Schwerwiegende Unfälle mit Sachschaden (i.e.S.)
G00	Getötete insgesamt
V00	Verunglückte insgesamt
GO1	Getötete innerorts
GO2	Getötete außerorts ohne BAB
GO3	Getötete auf BAB
GO4	Getötete auf Bundesstraßen außerorts
GA1	Getötete Kinder unter 15 Jahren
GA2	Getötete im Alter von 15 bis 17 Jahren
GA3	Getötete im Alter von 18 bis 24 Jahren
GA4	Getötete im Alter von 25 bis 64 Jahren
GA5	Getötete im Alter von 65 Jahren und älter
GV1	Getötete Fahrer und Mitfahrer von Kfz. (ohne mot. Zweiräder)
GV2	Getötete Fahrer und Mitfahrer von Pkw
GV3	Getötete Fahrer und Mitfahrer von Güterkraftfahrzeugen
GV4	Getötete Fahrer und Mitfahrer von Krafträdern/Kraftrollern
GV5	Getötete Fahrer und Mitfahrer von Moped/Mofas
GV6	Getötete Fahrer und Mitfahrer von Fahrrädern
GV7	Getötete Fußgänger
UP1	Unfälle mit Personenschaden innerorts
UP2	Unfälle mit Personenschaden außerorts ohne BAB
UP3	Unfälle mit Personenschaden auf BAB
UP4	Unfälle mit Personenschaden auf Bundesstraßen außerorts
UPA	Alkoholunfälle mit Personenschaden
VSV	Schwerverletzte insgesamt

Tab. 1: Zeitreihen im deutschen Unfallgeschehen

sind zum Zeitpunkt der Jahresendprognose bis einschließlich August verfügbar und weichen im Mittel 2-3 % von den endgültigen Daten ab. Sie können die endgültigen Anzahlen unter- aber gegebenenfalls auch überschätzen (dies ist viel seltener und passiert vor allem, wenn ein Unfall falsch kategorisiert wird).

In Bild 2 ist anhand der Reihe „Getötete außerorts ohne BAB“ der Unterschied zwischen vorläufigen und endgültigen Daten dargestellt. Die Abweichungen der vorläufigen von den endgültigen Daten zeigt ein saisonales Muster. In Bild 1 wird der allgemeine Trend dieses Musters dargestellt, indem die prozentualen monatlichen Abweichungen über alle Zeitreihen und alle Jahre gemittelt wurden. Es zeigen sich strukturelle Unterschiede in den Monaten, wobei die Anzahlen im Frühjahr tendenziell unvollständiger sind als die der Sommer- oder Wintermonate.

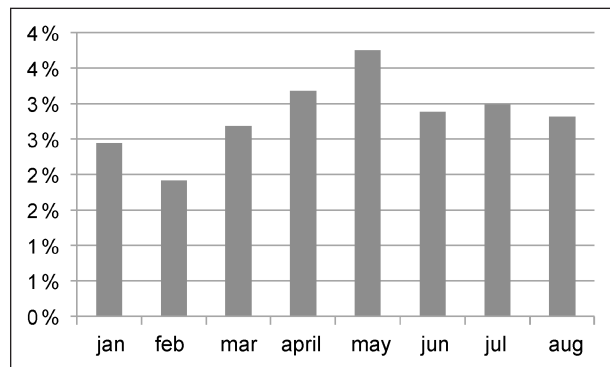


Bild 1: Endgültige vs. vorläufige Daten (Detaildaten): mittlere Abweichung (%) 2001-2014

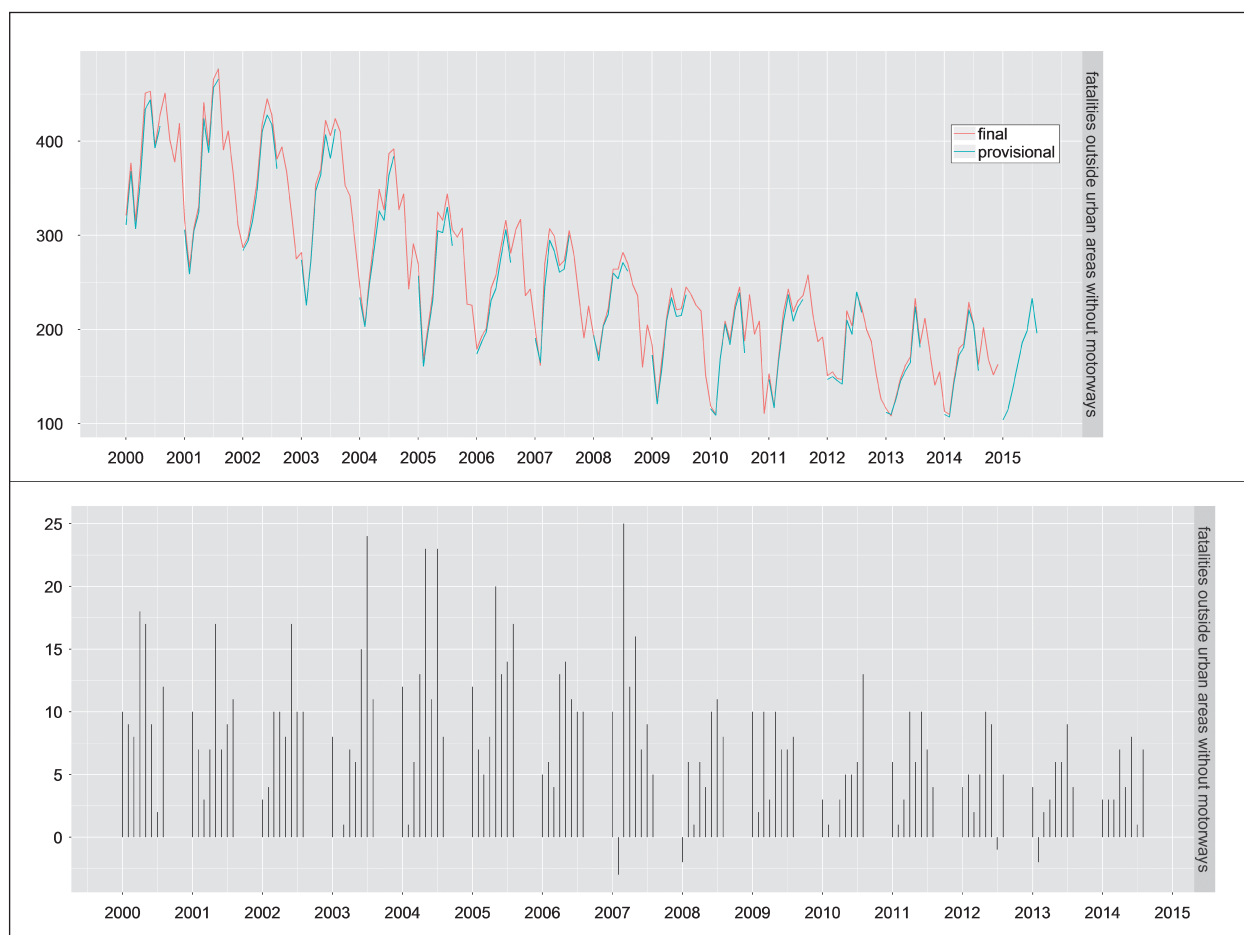
## 1.2 Schätzungen der BAST

Die Bundesanstalt für Straßenwesen bringt zum Ende jeden Jahres eine Prognose der Unfall- und Getötetenzahlen des noch laufenden Jahres heraus, um so Bilanz ziehen zu können über die Entwicklung der Verkehrssicherheit in Deutschland. Zu diesem Zeitpunkt sind die Daten von Januar bis August erhältlich, und die letzten Monate werden extrapoliert. Bei den Prognosen der BAST wird seit Jahren Witterungseinflüssen Rechnung getragen, und das Team der an der Prognose beteiligten Wissenschaftler dokumentiert seit Jahrzehnten systematisch die meteorologische Situation und das Unfallgeschehen eines jeden Monats. Dabei wird das Unfallgeschehen in die in Kapitel 1.1 genannten Zeitreihen eingeteilt. Bei der Jahresendvorhersage geben die Experten unter Berücksichtigung der textlichen Wetterbeschreibungen für jede Zeitreihe, zunächst unabhängig voneinander eine Schätzung für die verbleibenden Monate ab. Durch Diskussion und Plausibilisierung anhand der Summe der Zeitreihen werden die Schätzungen der einzelnen Experten zu einer gemeinsamen Vorhersage kombiniert.

Ziel des hier beschriebenen Forschungsvorhabens war es, den Zusammenhang zwischen Wetter und Unfallgeschehen mathematisch anhand eines Zeitreihenmodells zu beschreiben und mit diesem Modell Schätzungen für die Jahresendprognose zur Verfügung zu stellen.

Um die endgültige Anzahl der Unfälle und Verunglückten für das ganze Jahr vorherzusagen, müssen zwei Schätzungen vorgenommen werden:

- Schätzung der endgültigen Anzahlen für die Monate mit vorläufigen Daten,
- Schätzung der endgültigen Anzahlen für die letzten Monate des Jahres.



**Bild 2:** Getötete außerorts ohne BAB (GO2). Oberer Bildrahmen: Endgültige (rot) und vorläufige (blau) Daten. Unterer Bildrahmen: Differenz zwischen endgültigen und vorläufigen Daten

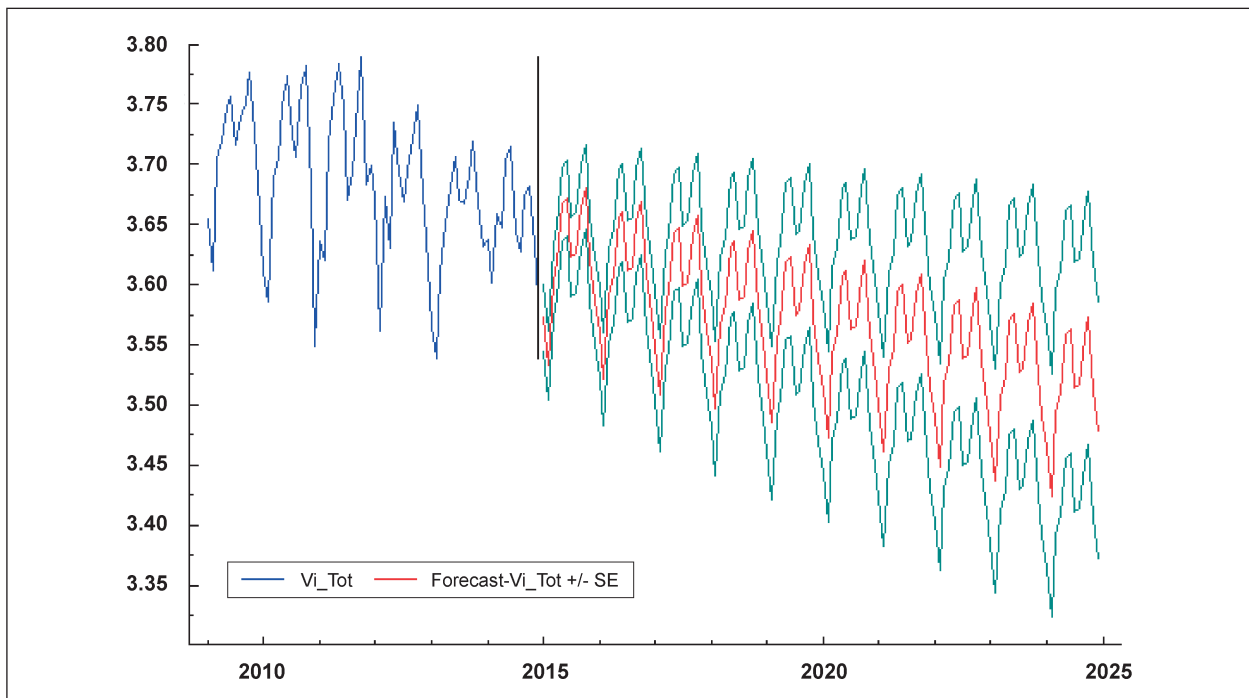
Ziel des hier beschriebenen Projekts waren dabei insbesondere zwei methodische Innovationen: (1) Einsatz von Strukturellen Zeitreihenmodellen. (2) Einbeziehung von meteorologischen Variablen in die Vorhersage.

### 1.3 Zeitreihenmodelle zur Vorhersage von Unfallhäufigkeiten

Bei Prognosen über die zu erwartenden Entwicklungen von Verkehrsunfall- oder Verunglücktenzahlen auf der Basis von Zeitreihenanalysen dominieren zwei analytische Ansätze: Modelle im Rahmen der State Space Technik (HARVEY, 1989; DURBIN, 2001) und das ARIMA System (BOX & JENKINS, 1970). State Space Modelle bilden eine Generalisierung des ARIMA Ansatzes und können eine wesentlich größere Bandbreite von möglichen Dynamiken integrieren. Die hier erstellten Modelle sind darum auch im State Space Ansatz implementiert.

#### 1.3.1 State Space Analyse

Eine Zeitreihe – also zum Beispiel die Unfallzahlen in aufeinanderfolgenden Monaten – wird im Rahmen einer State Space Analyse in verschiedene Komponenten untergliedert, wie zum Beispiel Trend (systematische Ab- oder Zunahme), saisonale Komponente (jahreszeitliche Schwankungen) und Einflüsse von externen Variablen (z. B. Witterungseinflüsse). Jede dieser Komponenten wird einzeln modelliert und kann abhängig vom spezifischen Muster der zu modellierenden Daten entweder deterministisch (gleichbleibend) oder stochastisch (über die Zeit veränderlich) sein. Die Modelle für die einzelnen Komponenten werden schließlich zum Gesamtmodell zusammengefügt. Die Kalibrierung der Modellparameter in State Space Analysen beruht auf dem bekannten Kalman Filter (KALMAN, 1960). Diese Methode wurde ursprünglich im Kontext des NASA Apolloprogramms entwickelt, um die Trajektorien von ballistischen Geschossen vorherzusagen. Mathematisch ist der Filter jedoch ein effizienter Algorithmus, um die



**Bild 3:** Vorhersage bis 2025 der Verletzten pro Quartal in Belgien mit Konfidenzintervall (68%) (MARTENSEN, BERGEL-HAYAT, Vorbereitung)

Parameter eines beliebigen linearen dynamischen Systems zu schätzen, dessen Ergebnis nicht störungsfrei gemessen werden kann. Die entscheidende Eigenschaft des Kalman Filters ist dabei die Optimierung der Parameter im Hinblick auf die Fähigkeit des Modells, neue Daten vorherzusagen. Damit sind State Space Analysen ein effizienter Weg zu optimalen Vorhersagen von zukünftigen Messwerten.

### 1.3.2 Vorhersagen











Die Modelldynamik, die über die vorhergehenden Zeitpunkte beobachtet wurde, kann – unter der Annahme, dass alle Prozesse weiterlaufen wie bisher – genutzt werden, um die Entwicklung für zukünftige Zeitpunkte vorherzusagen. Die Vorhersagen für die zu erwartende Unfallzahl bestehen aus einer Punktschätzung und dem Konfidenzintervall. Als Beispiel ist in Bild 3 die vorhergesagte Anzahl Verletzter pro Quartal für Belgien in den nächsten 10 Jahren dargestellt.

Die Punktschätzung basiert auf der jeweils letzten Schätzung der Trend- und der Saisonkomponente. Wichtiger als die Punktschätzung ist aber das Konfidenzintervall, das aus den Abweichungen zwischen vorhergesagtem und beobachtetem Wert bei den vorhergehenden Messzeitpunkten geschätzt wird. Es gibt den Bereich an, in dem mit

der gewünschten Wahrscheinlichkeit (in Bild 3, z. B. 68 %) der zukünftige Messwert liegen wird.

## 1.4 Witterungseinflüsse

Meteorologische Variablen wie Regen, Frost, aber auch warme Temperaturen und Sonnenschein können das Unfallgeschehen auf zwei Arten beeinflussen: Sie können das Unfallrisiko verändern, aber auch die Risikoexposition, an einem Unfall beteiligt zu sein (FOCANT & MARTENSEN, 2014). Für das Risiko spielen Faktoren wie Sicht, Sichtbarkeit oder auch eine mögliche Verminderung der Fahrbahngriffigkeit eine Rolle, die zu einem erhöhten Risiko führen, aber auch mögliche Kompensationsstrategien der Verkehrsteilnehmer (langsamer fahren, mehr Abstand halten), die das Unfallrisiko vermindern können. Das Endergebnis dieser gegenläufigen Tendenzen ist nicht einfach vorherzusagen und unterscheidet sich auch nach der Schwere der Unfälle und der Art der Verkehrsbeteiligung. Allgemein ergibt sich als Trend, dass Schnee und Frost eher zu einem niedrigeren Unfallrisiko führen und Regen eher zu einem höheren Risiko (siehe Übersicht bei FOCANT & MARTENSEN, 2014). Für die Risikoexposition spielt das Wetter eine wichtige Rolle, weil es die Entscheidungen sowohl über die Art der Verkehrsteilnahme (z. B. Rad oder Auto?) als auch über die Realisierung, Verschiebung oder

Variable	Pedestrian 	Cyclist 	Moped 	Motorcycle 	Car 	All
Maximum Temperature 	+	++	+++	++++		+
Number of Snow Days 		-	--	---	-	-
Number of Frost Days 	-					
Precipitation Height 		-		--	+	
Sun Duration 		+		+	(+)	(+)

**Bild 4:** Effekte von Witterungseinflüssen auf die Anzahl der verletzten Unfallopfer in Belgien (2003-2014). MARTENSEN, FOCANT, & DIEPENDAELE, 2016

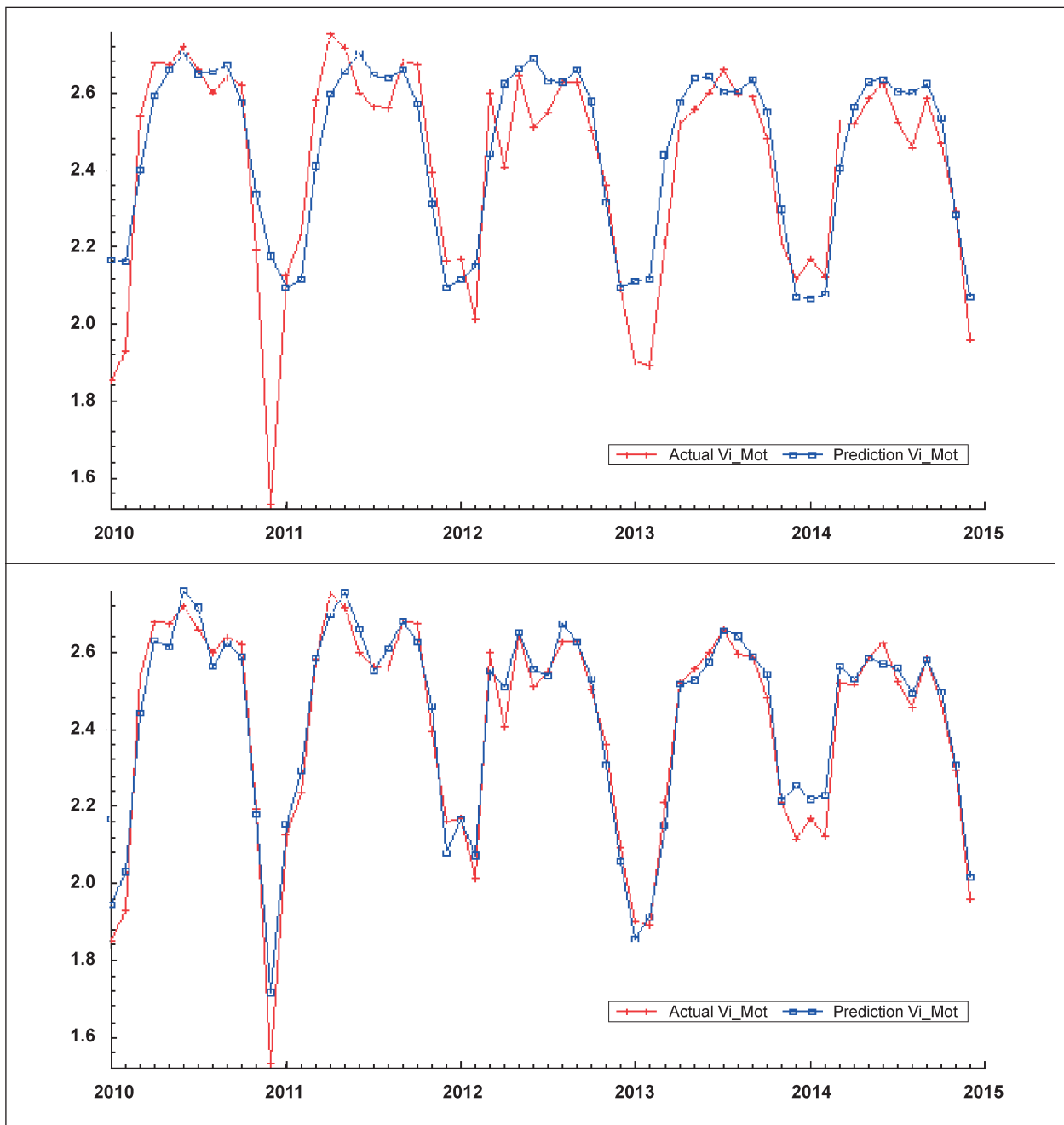
Unterlassung von möglichen Wegen beeinflusst. Hiervon sind besonders die nicht-motorisierten Verkehrsteilnehmer betroffen, die Witterungseinflüssen stärker ausgesetzt sind (SABIR, 2011). Da für diese Gruppe präzise Daten über die tatsächlich zurückgelegten Wege fehlen, können meteorologische Variablen einen Ersatz/Indikator für die Risikoexposition darstellen. Die beobachteten Witterungseffekte auf die Anzahl der Unfälle und Verunglückten sind also eine Kombination der Risiko- und der Expositionseffekte. Es ist wichtig, hierbei nach der Art der Verkehrsteilnahme zu unterscheiden. In Bild 4 werden die Resultate einer Studie über Witterungseinflüsse auf das belgische Unfallgeschehen wiedergegeben. Vor allem was die Niederschläge betrifft, ist der Einfluss auf den motorisierten Verkehr spiegelbildlich zum Effekt auf das Unfallgeschehen bei den ungeschützten Verkehrsteilnehmern. Es sind auch die letzteren, bei denen wir die stärksten Effekte des Wetters auf die Unfallhäufigkeit beobachten.

#### 1.4.1 Meteorologische Variablen als Prädiktoren

In der vorliegenden Studie geht es darum, den Effekt von meteorologischen Variablen auf die Entwicklung der Unfall- und Verunglücktenzahlen im Straßenverkehr in die Zeitreihenmodelle zur Unfallprognose zu integrieren. Es geht dabei nicht in erster Linie darum, die ursächlichen Zusammenhänge zu erklären, denn gerade für die ungeschützten Verkehrsbeteiligten gibt es wenige zeitlich so detaillierte Mobilitätsdaten, dass sie eine Trennung zwischen Risiko- und Expositionseffekten ermöglichen

würden. Es geht vielmehr darum, das Prognosemodell zu präzisieren und dafür zu sorgen, dass Witterungseinflüsse nicht fälschlich als Langzeittrends interpretiert werden. Es geht dabei um zwei mögliche Korrekturen:

1. Zum einen wird korrigiert in Bezug auf Witterungseinflüsse in den ersten 8/9 Monaten, für welche die Unfall- und Verunglücktenzahlen bekannt sind. Im Frühjahr 2014 wurde beispielsweise unter den Rad- und Motorradfahrern ein Anstieg der Getötetenzahlen um 5-10 % registriert und von den Experten der BASt dem frühen Beginn und dem generell milden Wetter des Frühlings zugeordnet. In einem Zeitreihenmodell ohne meteorologische Variablen schlägt eine solche Erhöhung (oder Reduktion) jedoch als Trendänderung zu Buche und würde als solche auch bei der Jahresendprognose auf die letzten 3 Monate mit unbekanntem Unfallgeschehen extrapoliert. Durch die Aufnahme von Prädiktoren wie zum Beispiel der mittleren Temperatur werden diese Effekte vom Langzeittrend getrennt.
2. Witterungseinflüsse können sich auch auf die letzten 3 Monate, für die die Unfall- und Verunglücktenzahlen vorhergesagt werden sollen, auswirken. Hier basiert in einem Modell ohne meteorologische Variablen eine Prognose auf der Annahme von jahreszeitlich typischen Witterungsbedingungen. Bei extremem Wetter (z. B. einem frühen Wintereinbruch, der tendenziell niedrigere Unfallzahlen nach sich zieht) führt dies zu Vorhersagefehlern. Für die Monate, für die die Wetterdaten bereits bekannt sind, kann die Aufnahme von meteorologischen Variablen deswegen die Vorhersage präzisieren.



**Bild 5:** Unfälle mit Verletzten mit Motorradbeteiligung (Belgien 2010-2014). Modellierung ohne meteorologische Prädiktoren (obere Grafik) und mit meteorologischen Prädiktoren (untere Grafik) (MARTENSEN & BERGEL-HAYAT, IRTAD, 2017)

Externe Variablen mit potenziellem Effekt auf das Unfallgeschehen wie Witterungseinflüsse oder ökonomische Variablen können in State Space Modellen als Prädiktoren aufgenommen werden. Als Beispiel zeigt Bild 5 die monatliche Anzahl von Unfällen mit Verletzten, an denen mindestens ein Motorradfahrer beteiligt war (Belgien, 2008-2013). Die obere Teilgrafik zeigt die Modellierung mit einer Saisonalen Komponente, aber ohne Witterungseinflüsse. Obwohl dieses Modell die jahreszeitlich allgemein üblichen Witterungseinflüsse sehr gut erfasst, kann es

bei extremen Witterungsbedingungen, wie zum Beispiel am sehr kalten und schneereichen Jahresende 2010, die Effekte auf das Unfallgeschehen nicht erklären. Im Vergleich dazu zeigt die untere Teilgrafik ein Modell mit drei zusätzlichen Prädiktoren: dem monatlichen Mittelwert der täglichen Maximum Temperatur, dem Niederschlag (in mm) und der monatlichen Anzahl der Tage mit Schneefall. Ein Großteil der Abweichungen vom typischen jahreszeitlichen Verlauf der Motorradunfälle kann mit diesen 3 externen Variablen erklärt werden.

## 1.4.2 Räumliche Variation

Deutschland ist ein flächenmäßig großes Land mit durchaus unterschiedlichen klimatischen Gegebenheiten – man denke zum Beispiel an die norddeutsche Tiefebene im Vergleich zu den Mittelgebirgen und den Hochgebirgsgebieten der Alpen. Die Unfallzahlen, die es zu modellieren gilt, sind jedoch gesamtdeutsche Daten. Wie kann und muss man dennoch der Variation im Wettergeschehen Rechnung tragen? Hierfür sind einige Prinzipien wichtig:

In der Entwicklung des gesamtdeutschen Unfallgeschehens werden lediglich sehr starke Abweichungen vom Normalwetter sichtbar, die einen Großteil der mobilen Bevölkerung betreffen. Heftiger Schneefall im oberen Alpengebiet wird zwar die lokale Bevölkerung in der Wahl der Verkehrsmittel beeinflussen, wird sich aber – solange der Schneefall auf diese recht dünn besiedelte Region begrenzt bleibt – nicht erkennbar auf die gesamtdeutschen Unfallzahlen auswirken. Um eine Schätzung zu bekommen, wie stark der Einfluss von lokalen Wetterphänomenen ist, werden darum die gemessenen Wetterwerte mit der Population der betreffenden Region gewichtet, um so dem Wetter in dicht besiedelten Gebieten einen stärkeren Einfluss zu geben.

## 2 Methode

### 2.1 Strukturelle Zeitreihen Modelle

Wie in der Einleitung beschrieben, gibt es zwei entscheidende Schritte, um die Unfall- und Verunglücktenzahlen eines ganzen Jahres vorherzusagen: (1) Die Schätzung der endgültigen Anzahlen für die Monate mit vorläufigen Daten und (2) die Vorhersage der Anzahlen für die fehlenden Monate unter Einbeziehung der bisherigen Trends und des Wettergeschehens.

Beiden Schätzungen liegt derselbe Modelltyp zugrunde: ein strukturelles Zeitreihenmodell, basierend auf der State Space Methodologie. Für den ersten Schritt (Schätzung der endgültigen Anzahlen aus den vorläufigen) besteht die Zeitreihe dabei aus den historischen beobachteten Differenzen zwischen vorläufigen und endgültigen Daten. Im zweiten Schritt besteht die Zeitreihe aus Unfall- oder Verunglücktenzahlen. Beide Modellierschritte werden jeweils für alle 27 in Kapitel 1.1 beschriebenen Zeitreihen ausgeführt.

### 2.1.1 Allgemeine Struktur

Die allgemeine Form der hier verwendeten Modelle wird in Gleichung 1 wiedergegeben:

$$\text{Log}(y_t) = \mu_t + \sum_{j=1}^{[s/2]} \gamma_{jt} + \varepsilon_t$$

Wobei  $y_t$  der beobachtete Wert in Monat  $t$  ist,  $\mu_t$  der Level des Modells ist,  $\sum_{j=1}^{[s/2]} \gamma_{jt}$  die saisonale Komponente und  $\varepsilon_t$  der Fehlerterm.

Der Level, also die momentane Höhe der Zeitreihe, bestimmt sich zu jedem Zeitpunkt  $t+1$  aus dem Level des vorhergehenden Monats  $\mu_t$ , der Steigung  $v_t$ , und einer Zufallskomponente  $\eta_t$ .

$$\mu_{t+1} = \mu_t + v_t + \eta_t$$

Die Steigung (Slope) ist also die regelmäßige monatliche Zu- oder Abnahme der Werte. In manchen Fällen ist sie tatsächlich über den ganzen modellierten Zeitraum hin konstant. Sie kann aber ihrerseits auch variieren. Man spricht dann von einem smooth drift einer sanften Strömung in der Zeitreihe. In diesem Falle ist die Steigung des Monats  $t+1$  die des Vormonats plus einer Zufallskomponente  $\xi_t$ .

$$v_{t+1} = v_t + \xi_t$$

### 2.1.2 Schätzung

Das R-Paket KFAS enthält Funktionen für Kalman Filter, Smoother und die Simulation von Linearen State Space Modellen mit exakter oder diffuser Initialisierung. Zusätzlich wurde eine Funktionsbibliothek geschrieben, die spezifische Prozeduren zur Modelldefinition, -schätzung und -evaluation für dieses Projekt enthält. Im Prinzip wurde die Implementierung an STAMP orientiert (KOOPMAN, HARVEY, DOORNIK & SHEPARD, 1999; online unter: <http://stamp-software.com/stamp.html>). Die in STAMP verwendeten Techniken zur Parameterschätzung basieren allerdings auf der Annahme, dass die Fehlerterme für alle Komponenten normal verteilt sind. Diese Verteilungsannahme wird im Kontext der Analyse von Unfall- und Verunglücktenzahlen durch eine Logtransformation recht gut approximiert. Bei kleinen Anzahlen (< 20) ist die Annäherung jedoch nicht mehr so gut und Nullzellen können gar nicht modelliert werden. Die Stratifikation der Unfallzahlen in Subgruppen führt in verschiedenen Fällen zu Nullzellen oder zu Anzahlen,



bei denen man der Annäherung an die Normalverteilung nicht mehr vertrauen kann.

Im Gegensatz zu STAMP ermöglicht KFAS auch die Modellierung von nicht normalverteilten Daten (siehe DURBIN & KOOPMAN, 2012). Die Schätzung der Modelle findet in zwei hintereinander geschalteten iterativen Verfahren statt. Mit dem Standardverfahren werden die Modellparameter erst auf der Basis der logarithmierten Unfallzahlen unter Annahme einer Normalverteilung geschätzt. Alle Null-Zellen werden dabei als Missing Values betrachtet. Die Parameterschätzwerte werden dann als Startwerte für ein zweites iteratives Verfahren benutzt, das die nicht logarithmierten Anzahlen unter der Annahme einer Poissonverteilung modelliert.

### 2.1.3 Extrapolation der vorläufigen Daten

Um unter Berücksichtigung der vorläufigen Daten zu einer Schätzung der endgültigen Anzahlen zu kommen, wird der Unterschied zwischen diesen beiden Zeitreihen als Input genommen. Für jede der 27 Unfallreihen wird also genau genommen die Unterschätzung der endgültigen Daten in den vergangenen Jahren modelliert, nämlich die Differenz  $d_t$  zwischen vorläufigen und endgültigen Daten. Aus der festgestellten Struktur (dem Level – also dem allgemeinen Mittelwert, einer möglichen Steigung, und einer möglichen Saisonkomponente) wird der Unterschied für die Monate vorhergesagt, für die noch keine endgültigen Daten erhältlich sind. Die geschätzte Differenz  $\hat{d}_t$  wird dann zu den vorläufigen Anzahlen addiert.

Eine detaillierte Beschreibung der Extrapolation der vorläufigen Daten ist im technischen Anhang gegeben.

### 2.1.4 Das Wettermodell

Die endgültigen Daten aus den vorhergehenden Jahren bilden die Basis, um die bisherigen Trends zu analysieren und das Verhältnis zwischen Wetter- und Unfallgeschehen zu bestimmen, da sie vollständiger und reliabler sind als die vorläufigen Daten. Diese Daten bilden also den Input zum Wettermodell, das den Kern der Vorhersage bildet.

Der Input zu diesem Modell besteht aus logarithmierten Unfallhäufigkeiten. Das heißt im Log-Raum gehen wir davon aus, dass die Unfallhäufig-

keiten eine lineare Entwicklung zeigen und dass die Komponenten, die im allgemeinen Modellrahmen (Kapitel 2.1.1) beschrieben sind, addiert werden können. Für die tatsächlich beobachteten Häufigkeiten heißt dies, dass wir von einem exponentiellen Verlauf ausgehen und dass die verschiedenen Komponenten multiplikativ miteinander verbunden sind.

Für das Wettermodell wird der allgemeine Modellrahmen (Kapitel 2.1.1) um eine Regressionskomponente erweitert, in der der Einfluss der meteorologischen Variablen geschätzt wird.

$$\text{Log}(y_t) = \mu_t + \sum_{j=1}^{[s/2]} \gamma_{jt} + \sum_{k=1}^r \beta_k x_{kt} + \varepsilon_t$$

Die Regression  $\sum_{k=1}^r \beta_k x_{kt}$  ist hier der allgemeinen Modellgleichung hinzugefügt. Sie enthält  $k$  Prädiktorvariablen  $x_{kt}$  und ihre Regressionsgewichte  $\beta_k$ . Die Gewichte  $\beta_1 - \beta_k$  geben den Einfluss der Prädiktorvariablen auf die Unfallhäufigkeit an.

Eine detaillierte Beschreibung des Wettermodells ist im technischen Anhang gegeben.

## 2.2 Klimadaten

### 2.2.1 Rohdaten

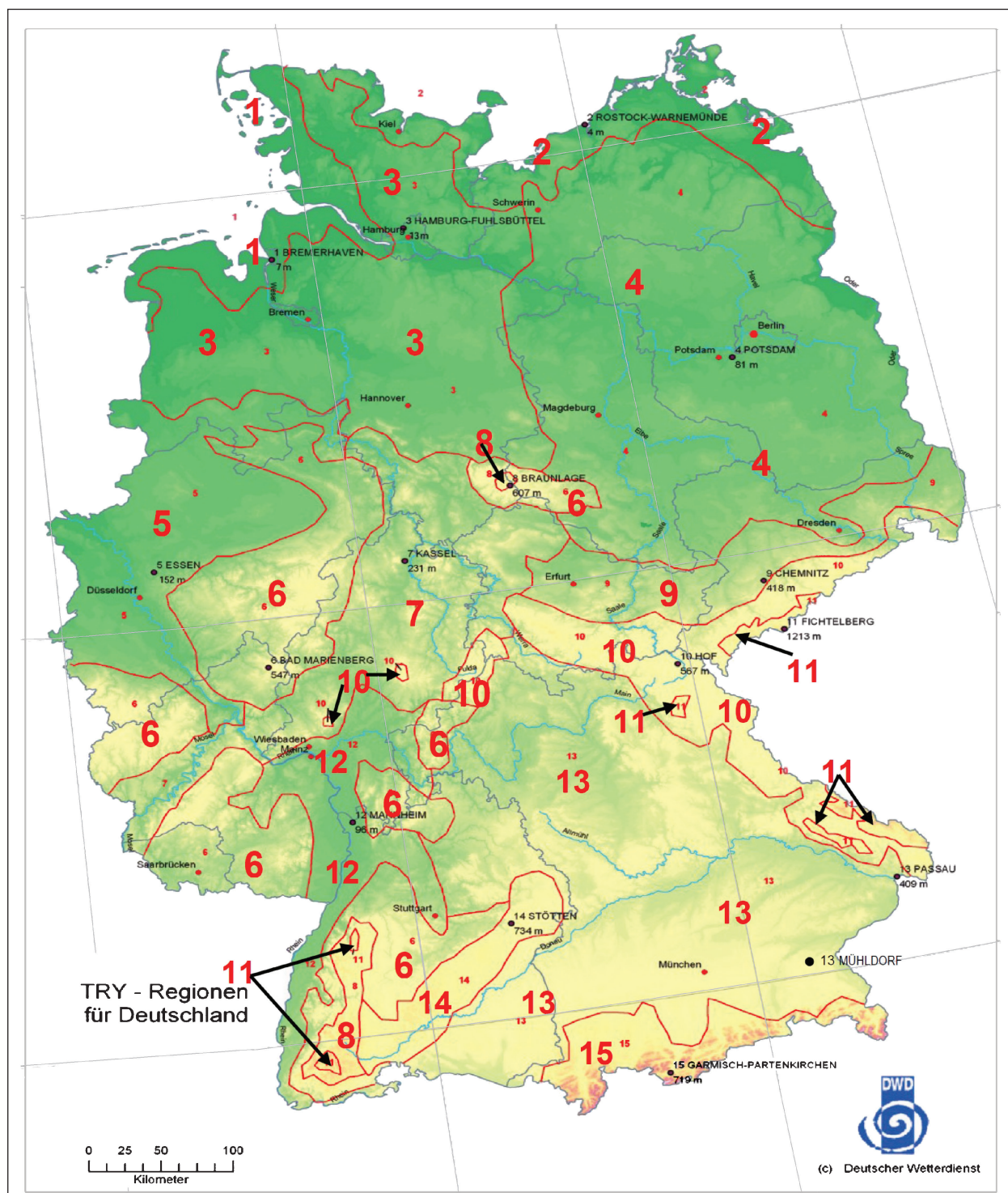
Der Deutsche Wetterdienst (DWD) – und hierbei speziell das German Climate Center – stellt Daten in verschiedenen Aufarbeitungen zur Verfügung (z. B. Rasterdaten oder Mittelwerte für jedes Bundesland). Für das vorliegende Projekt ist die Aufarbeitung weniger relevant als die schnelle Zugänglichkeit der Daten. Wenn Anfang Dezember die Jahresendprognose durchgeführt wird, ist es wünschenswert, die Wettergegebenheiten bis einschließlich November in die Vorhersage aufnehmen zu können. Darum wird hier auf die Stationsdaten zurückgegriffen, die meist täglich aktualisiert werden.

Die Art und Weise, in der die meteorologischen Variablen verarbeitet werden, ist durch die Ziele des hier beschriebenen Projekts motiviert. Alle Verarbeitungsschritte dienen einer optimalen Vorhersage der Monate mit noch fehlenden Unfallzahlen. Die im weiteren Verlauf beschriebenen Transformationen und Aggregationen dienen diesem Ziel, wobei allerdings in Kauf genommen werden muss, dass die ursprüngliche Phänomenologie der gemessenen

Wetterwerte in den Modellierungsergebnissen nur noch bedingt zu erkennen ist. Unter dem Gesichtspunkt der Interpretierbarkeit der Resultate wäre das hier gewählte Verfahren als suboptimal zu beurteilen. Alle Verarbeitungsschritte haben jedoch die Präzision der Modellierung und damit der Vorhersagen erhöht.

## 2.2.2 Klimazonen

Die Selektion der Stationen aus 1090 möglichen basierte auf der Test-reference-year (TRY) Studie (CHRISTOFFER et al., 2004; BBR, 2011), in der 15 verschiedene Klimazonen identifiziert wurden (siehe Bild 6).



**Bild 6:** TRY-Regionen für Deutschland (BBR, CEC, DWD: Aktualisierte und erweiterte Testreferenzjahre von Deutschland für mittlere, extreme und zukünftige Witterungsverhältnisse. Projektbericht 2011)

	TRY Zone	Referenzstation	Zugewiesene NUTS Regionen
1	Nordseeküste	Bremerhaven	
2	Ostseeküste	Rostock-Warnemünde	
3	<i>Nordwestdeutsches Tiefland</i>	<i>Hamburg-Fuhlsbüttel</i>	DE50, DE60, DE91, DE92, DE93, DE94, DEF0
4	<i>Nordostdeutsches Tiefland</i>	<i>Potsdam</i>	DE30, DE40, DE80, DED5, DEE0
5	<i>Niederrheinisch-westfälische Bucht und Emsland</i>	<i>Essen</i>	DEA1, DEA2, DEA3, DEA4
6	<i>Nördliche und westliche Mittelgebirge - Randgebirge</i>	<i>Bad Marienberg</i>	DEA5, DEB1, DEB2, DEC0
7	<i>Nördliche und westliche Mittelgebirge - Zentrale Bereiche</i>	<i>Göttingen*</i>	DE72, DE73
8	Oberharz und Schwarzwald (mittlere Lagen)	Braunlage	
9	<i>Thüringer Becken und Saechsisches Huegelland</i>	<i>Chemnitz</i>	DED2, DED4, DEG0
10	Südöstliche Mittelgebirge bis 1000 m	Hof	
11	Erzgebirge, Boehmer, und Schwarzwald > 1000 m	Fichtelberg	
12	<i>Oberrheingraben und unteres Neckartal</i>	<i>Mannheim</i>	DE12, DE13, DE71, DEB3
13	<i>Schwäbisch-fränkisches Stufenland und Alpenvorland</i>	<i>Passau</i>	DE11, DE14, DE21, DE22, DE23, DE24, DE25, DE26, DE27
14	Schwäbische Alb und Baar	Stötten	
15	Alpenrand und -täler	Garmisch-Partenkirchen	

\* Für ursprüngliche Referenzstation Kassel (Zone 7) keine aktuellen Daten erhältlich

Tab. 2: Deutsche Klimazonen mit Referenzstationen (TRY 2004) und die zugewiesenen NUTS2-Regionen

Diese Zonen formen nicht unbedingt homogene Gebiete. Z. B. bilden die oberen Lagen verschiedener Mittelgebirge eine Zone, während die dazwischen liegenden Gebiete zu anderen Zonen gehören. Zu jeder Zone wurde in der TRY-Studie außerdem eine Referenzstation angegeben, deren Werte als repräsentativ für die ganze Zone angesehen werden können. In Tabelle 2 sind die 15 Deutschen Klimazonen mit der Referenzstation aufgeführt.

Um zu einer Schätzung der Bevölkerungsanzahl für jede Klimazone zu kommen, müssen die Klimazonen der TRY-Studie den statistischen räumlichen Einheiten zugewiesen werden, für die Populationsdaten verfügbar sind. Dies sind die sogenannten NUTS-Regionen (Nomenclatur for territorial units for statistics) die auf 3 Niveaus erhältlich sind. NUTS2 entspricht bei den kleineren Bundesländern dem gesamten Bundesland, während größere Bundesländer in 2- bis 5-Regionen unterteilt sind. Um zu einer Schätzung der Bevölkerung zu kommen, die in jeder Klimazone lebt, sind jeder NUTS2-Region diejenige Klimazone zugewiesen, die die größte flächenmäßige Überlappung aufweist. Dabei sind 6 Klimazonen keiner Region zugewiesen (Nord und Ostseeküste und die verschiedenen Hochlagen). Zum Beispiel befinden sich die Gebiete der Zone 15 Alpenrand und -täler in Oberbayern (NUTS DE21), dennoch dominiert Zone 13 Schwäbisch-fränkisches Stufenland und Alpenvorland diese Region. In der Spalte ganz rechts von Tabelle 2 werden alle Regionen angegeben die einer Klimazone zugeord-

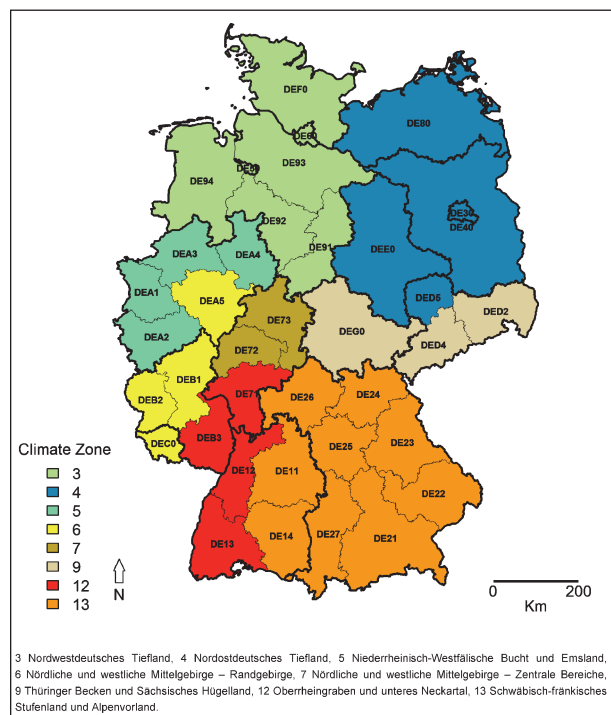


Bild 7: Zuordnung der Deutschen NUTS2-Regionen zu den Klimazonen der TRY-Studie

net sind. In Bild 7 ist grafisch wiedergegeben, welcher Klimazone jede NUTS2-Region zugewiesen ist. Jede Klimazone (bzw. die Daten ihrer Referenzstation) wird mit der Summe der Bevölkerung in der ihr zugeordneten NUTS2-Regionen gewichtet.

## 2.3 Wettervariablen und ihre Transformation

Die Klimadaten werden als Tageswerte vom DWD importiert. Es wird dabei auf die direkt an den Stationen registrierten Tageswerte zugegriffen. Alle quantitativen Messwerte werden in die Modellierung aufgenommen. Zusätzlich werden aus der Variablen Niederschlagshöhe\_Ind noch für jeden Tag vier Indikatorvariablen über die Art des Niederschlags geformt:

- Dry: kein Niederschlag,
- Rain: (nur) Regen,
- Snow: (nur) Schnee,
- RainAndSnow: Regen und Schnee oder Schneeregen.

Alle Variablen sind in Tabelle 3 wiedergegeben.

### 2.3.1 Schiefekorrektur

Im Gegensatz zu normal verteilten Daten haben viele meteorologische Daten einen absoluten Nullpunkt (z. B. Windgeschwindigkeit oder Niederschlagshöhe) oder ein absolutes Maximum (z. B. der Bedeckungsgrad), wodurch die Verteilung vieler

Variablen schief ist. Um beim Aggregieren eine Verzerrung der resultierenden Werte durch extreme Messungen zu vermeiden und um die ganze Bandbreite der Werte für die Schätzung des Regresionsteils im Wettermodell optimal nutzen zu können, sollten die Variablen jedoch möglichst symmetrisch verteilt sein. Die Werte werden darum teilweise transformiert, um die Schiefe der Verteilungen zu korrigieren.

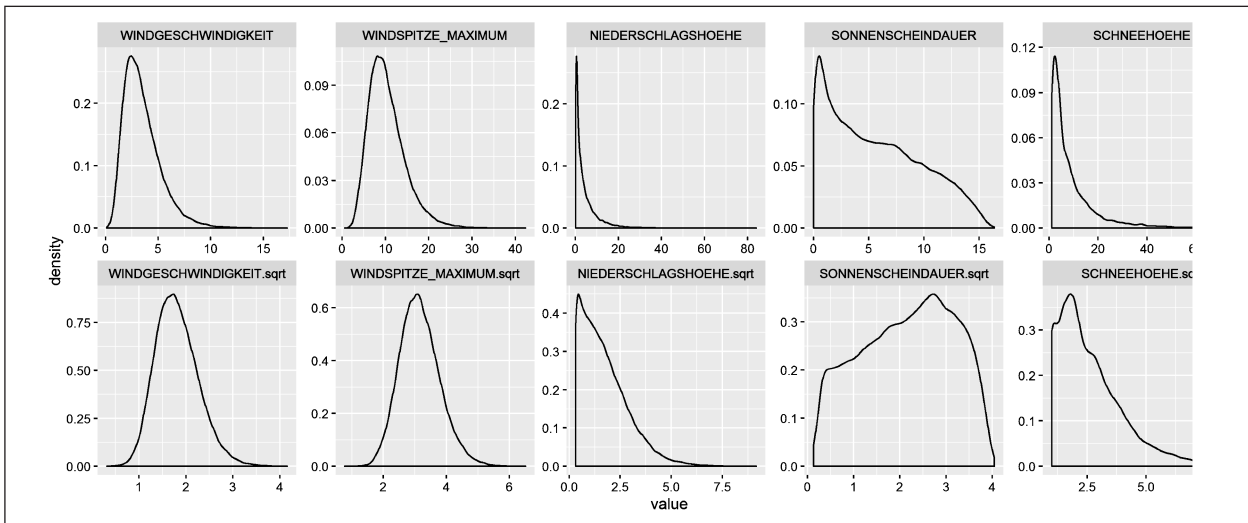
Kontinuierliche Variablen mit absolutem Nullpunkt (Dampfdruck, Luftdruck, Windgeschwindigkeit, Tageswindspitze und Niederschlagshöhe) sind rechts-schief. Um dies zu korrigieren, wird die Quadratwurzel gezogen. In Bild 8 ist die Wahrscheinlichkeitsdichte dieser Variablen vor und nach der Wurzel Transformation abgebildet.

Die Variablen Bedeckungsgrad und relative Luftfeuchtigkeit haben ein absolutes Maximum (nämlich 1) und sind darum links-schief. Sie werden durch die Arkussinus-Wurzel transformiert ( $\sin^{-1}\sqrt{x/\max(x)}$ ). Bild 10 zeigt die Verteilungen vor und nach der Transformation.

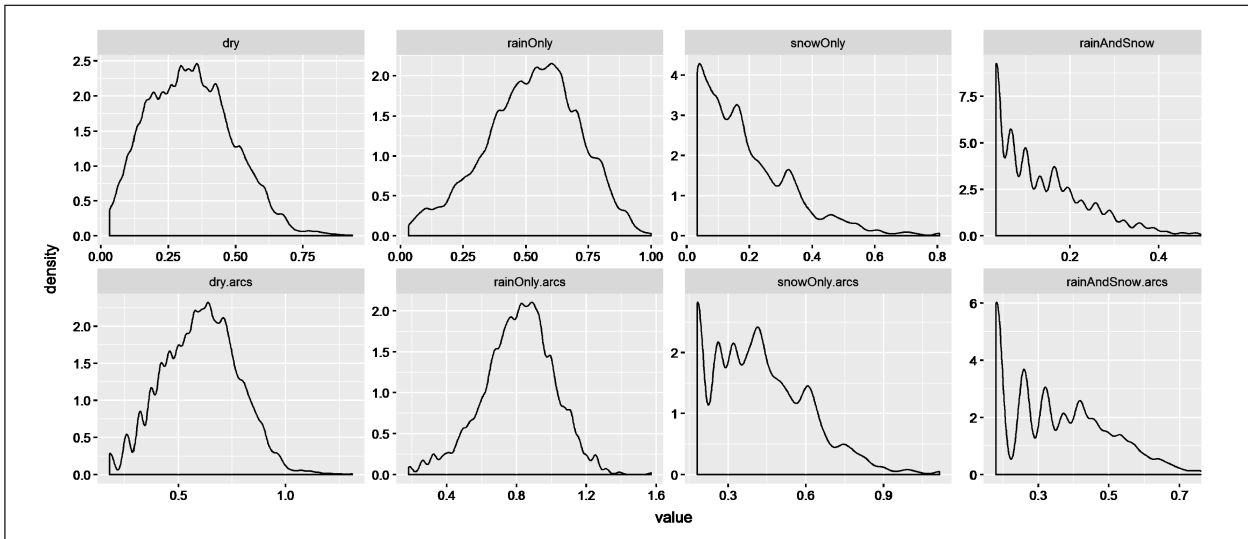
Die Indikatorvariablen (Dry, Rain, Snow, Rainand-snow) sind 1 bei Auftreten des beschriebenen Phänomens und sonst 0. Der monatliche Mittelwert ist also die Proportion der Tage, an denen das

Variable	Bedeutung	Einheit	Transformation
LUFTTEMPERATUR	Tagesmitteltemperatur in 2 m Höhe	°Celsius	
LUFTTEMPERATUR_MAXIMUM	Tagesmaximumtemperatur	°Celsius	
LUFTTEMPERATUR_MINIMUM	Tagesminimumtemperatur	°Celsius	
LUFTTEMP_AM_ERDB_MINIMUM	Minimumtemperatur am Erdboden	°Celsius	
DAMPFDRUCK	Dampfdruck	hPa	Sqrt
BEDECKUNGSGRAD	Wolkenbedeckung	1/8	ArcSin
LUFTDRUCK STATIONSHÖHE	Luftdruck	hPa	Sqrt
REL FEUCHTE	Relative Luftfeuchte	%	ArcSin
WINDGESCHWINDIGKEIT	Tagesmittel Windgeschwindigkeit	m/s	Sqrt
Tageswindspitze	Maximale Windspitze	m/s	Sqrt
NIEDERSCHLAGSHÖHE	Tägliche Niederschlagshöhe	mm	Sqrt
DRY	Tag ohne Niederschlag	Indikator	ArcSin
RAIN	Tag mit Regen (ohne Schnee)	Indikator	ArcSin
RAINandSNOW	Tag mit Regen und Schnee	Indikator	ArcSin
SONNENSCHNEIDAUER	Tageswert Sonnenscheindauer	h	Sqrt
SCHNEEHÖHE	Tageswert Schneehöhe	cm	Sqrt

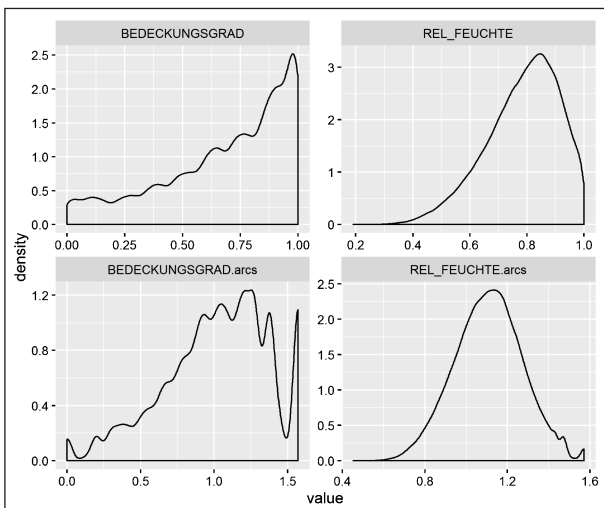
Tab. 3: Meteorologische Variablen aufgenommen in die Modellierung



**Bild 8:** Verteilungsform vor und nach der Transformation von rechtsschiefen Variablen



**Bild 9:** Verteilungsform vor und nach der Transformation von Indikatorvariablen



**Bild 10:** Verteilungsform vor und nach der Transformation von linksschiefen Variablen

jeweilige Phänomen aufgetreten ist. Um die Rechts-Schiefe dieser Verteilungen zu korrigieren, werden diese durch die Arkussinus-Wurzel transformiert ( $\sin^{-1}\sqrt{x}$ ) (Bild 9).

### 2.3.2 Z-Transformation

Die Z-Transformation der Werte gibt an, wie ungewöhnlich ein gemessener Wert für den betreffenden Monat und die betreffende Station ist. Von den (bereits schiefe-korrigierten) Werten wird dabei der durchschnittliche Monatswert der letzten 25 Jahre (seit 1991) abgezogen und die Differenz wird durch die Standardabweichung dividiert (siehe exakte Beschreibung im technischen Anhang).

- Beispiel: Wenn es im Januar in Hamburg 10 Grad und die mittlere Januartemperatur 3 Grad ist, dann liegt der beobachtete Wert 7 Grad über seinem Monats/Stationsmittelwert. Diese +7 Grad werden dann noch durch die Standardabweichung über alle Hamburger Januar-messungen seit 1991 geteilt. Derselbe Messwert im Juni würde eine ganz andere Z-Transformation ergeben, da die mittlere Junitemperatur viel höher liegt.

Für jede der oben beschriebenen Variablen gehen also zwei Werte in die Modellierung ein: (1) der (schiefe-korrigierte) Messwert und (2) der Z-transformierte Wert.

### 2.3.3 Aggregation

Die (schiefe-korrigierten) Messwerte und die Z-Werte aller Variablen werden in zwei Schritten gemittelt, als erstes über die Tage eines jeden Monats und als zweites über alle 9 genutzten Wetterstationen. Bei der Aggregation über die Stationen werden die Werte einer jeden Station mit der Summe der Bevölkerung der ihr zugeordneten NUTS2-Regionen gewichtet.

### 2.3.4 Hauptkomponenten-Analyse

Zwischen den Stationsvariablen besteht häufig ein enger Zusammenhang. So sind z. B. die mittlere Tagestemperatur, die maximale Tagestemperatur, die minimale und auch die Indikatorvariable für Schneefall stark miteinander korreliert. In einer Regression führen korrelierte Prädiktoren zu instabilen Schätzungen der Regressionsgewichte und der ihnen zugeschriebenen Fehlerterme. Kleinste Änderungen in den Daten können dann zu völlig anderen Resultaten führen.

Ein übliches Verfahren zur Selektion der Prädiktorvariablen ist der schrittweise Ausschluss von nicht-signifikanten Prädiktoren (Stepwise). Bei hochkorrelierten Prädiktoren ist dieses Verfahren jedoch nicht sehr valide, weil die Signifikanzschätzung fehlerbehaftet ist. Aus diesem Grunde werden die aggregierten Stationsmesswerte wie auch die aggregierten Z-Werte erst einer Hauptkomponenten-Analyse unterzogen, wobei die den beobachteten Werten zugrunde liegenden Dimensionen extrahiert werden. Alle extrahierten Komponenten sind unkorreliert und werden bestimmt durch ihre Fähigkeit, die Varianz der Inputvariablen zu erklären. Alle

Komponenten, die mindestens 3 % der Varianz in den Messwerten oder ihren Z-Transformationen erklären, werden in die folgenden Modellerschritte aufgenommen. Siehe die Details dieser Analyse im technischen Anhang.

### 2.3.5 Vervollständigung der Wetterdaten (Imputation)

Die Hauptkomponenten der Wetterdaten werden für alle Monate berechnet, von 1991 bis zum ersten Tag des laufenden Monats. Auf diese Weise wird vermieden, dass Mittelwerte für einen Monat berechnet werden, für den es nur für die ersten Tage Messwerte gibt. Für jede Hauptkomponente wird eine eigene Zeitreihenanalyse berechnet, um die Werte für Monate ohne (vollständige) Wetterdaten vorherzusagen. Auch hier wird der allgemeine Modellrahmen (Kapitel 2.1.1) verwendet mit Level, Steigung und saisonaler Komponente. Die Details dieser Vorhersage der fehlenden Wetterkomponenten sind im technischen Anhang beschrieben.

## 2.4 Selektion der Prädiktoren

Für jede der 27 Reihen (siehe Kapitel 1.1) werden die folgenden Schritte durchlaufen:

1. Die Hauptkomponenten aus der vorhergehenden Analyse wie auch die Interaktionen erster Ordnung zwischen den Komponenten werden zuerst alle als Prädiktoren in die Modellierung aufgenommen. Die Aufnahme der Interaktion ermöglicht es auch, Effekte des Zusammenwirkens von Komponenten zu berücksichtigen. So kann zum Beispiel Regen bei kaltem Wetter einen anderen Effekt auf das Radfahrverhalten haben als Regen bei warmem Wetter.
2. Alle nicht signifikanten Interaktionen ( $\alpha = .01$ ) werden aus dem Modell entfernt.
3. Alle nicht signifikanten Hauptkomponenten ( $\alpha = .10$ ) werden aus dem Modell entfernt. Für die Hauptkomponenten wird also das gängige Signifikanzniveau erhöht, um sicher zu sein, jede relevante Information im Modell zu behalten.

Das zur Vorhersage benutzte Modell für jede Zeitreihe enthielt also alle Hauptkomponenten, die auf dem 10 % Niveau signifikant sind und alle Interaktionen, die auf dem 1 % Niveau signifikant sind.

## 2.5 Wettereffekte

Zusätzlich zu den Schätzungen die sich entweder auf beobachtete oder imputierte Wetterdaten stützen wird für jede Reihe monatlich auch ein wetterkorrigierter Wert geschätzt. Dabei wird ein alternatives Modell geformt in dem alle vorhergehenden Prozesse identisch sind, die Wetterwerte des betreffenden Monats jedoch durch ihre Langzeitmittelwerte ersetzt werden. Das Verhältnis zwischen beiden Schätzwerten gibt an, inwieweit die Unfall- oder Verunglücktenzahlen aufgrund der spezifischen Wetterumstände des betreffenden Monats erhöht oder gesenkt sind.

$$\text{Wettereffect} = \frac{n \text{ bei tatsächlichem Wetter}}{n \text{ bei Langzeitmittelwert}}$$

## 3 Resultate

### 3.1 Vorhersagequalität

Der Kalman Filter ist ein Verfahren, mit dem die Vorhersage des folgenden (noch unbekannt) Messpunktes optimiert wird. Die Modellkomponenten und die Selektion der Prädiktoren ermöglichen also im Prinzip die bestmögliche Vorhersage des Folgemonats. Der hauptsächliche Verwendungszweck des Modells ist jedoch die Vorhersage der Jahresendwerte auf der Basis der vorläufigen Daten bis August (bzw. September) einschließlich.

Um die Qualität der Modelle bezüglich dieses Ziels zu evaluieren, wird für die letzten 15 Jahre berechnet, wie gut verschiedene Prognoseverfahren auf Basis der Daten von Januar bis August die endgültige Jahresanzahl vorhergesagt haben/hätten. Für diese Vorhersagen wird der Root Mean Squared Error (RMSE) für die Jahre 2010 bis 2015 berechnet: die mittlere Abweichung der Vorhersagen vom tatsächlichen Jahresendergebnis.

$$rmse = \sqrt{\frac{\sum(\text{predicted} - \text{final})^2}{\text{years}}}$$

Als Vergleichsgrundlage werden die bisherigen Vorhersagen der BAST benutzt. Die Qualität der Vorhersagen des BAST Expertenteams kann dabei als ehrgeiziges Ziel für die Modellierung angesehen werden.

Um zu evaluieren, welche Rolle die meteorologischen Variablen bei der Vorhersage spielen, wird der RMSE für Modelle mit und ohne Wettervariablen verglichen.

Aufgrund dieser Überlegungen wird also für jede Zeitreihe der Root Mean Squared Error (RMSE) 2010-2015 für die folgenden Vorhersagemethoden berechnet.

- Prgnss: Vorhersagen des Wettermodells,
- BAST: Vorhersagen des BAST Teams,
- NoWeather: Vorhersagen des Modells ohne Wettervariablen.

Die RMSE-Werte für jede Reihe sind in Tabelle 4 gegeben. Der Anstieg der Präzision wird dabei berechnet als

$$\text{Anstieg Präzision} = \frac{rmse(alt) - rmse(neu)}{rmse(alt)}$$

Im Vergleich zu den Vorhersagen der BAST in den letzten 5 Jahren sehen wir bei 23 der 27 Reihen eine Verbesserung der Vorhersagequalität für die Prognosen des Wettermodells, wobei die Präzision der Vorhersagen bis zu 55 % gestiegen ist. Bei zwei Reihen gibt es keinen Unterschied (Getötete auf Autobahnen und getötete Insassen von Güterkraftfahrzeugen) und lediglich bei zwei Reihen ist die Vorhersagegenauigkeit geringfügig schlechter als die des BAST Teams (Getötete Insgesamt: -8 %, Getötete 25-64 Jahre: -6 %).

Die Einbeziehung von Wettervariablen führt in 20 der 27 Reihen zu einer verbesserten Vorhersagegenauigkeit. Die Zunahme ist am höchsten für Unfälle mit Personenschaden innerorts (42 %). Für Getötete unter 15 Jahren, getötete Insassen von Güterkraftfahrzeugen und getötete Mopednutzer hat die Aufnahme von meteorologischen Variablen keinen Effekt, was sich auch darin äußert, dass für diese Reihen kaum signifikante Wettervariablen in der Modellgleichung enthalten sind.

Für einige Reihen ist die Vorhersage ohne Wettervariablen etwas besser als die mit (Unfälle mit Personenschaden auf Autobahnen, Unfälle mit schwerem Sachschaden, Getötete zwischen 15 und 17 Jahren und die Anzahl aller polizeilich registrierten Unfälle). Dies ist bemerkenswert, da in den Modellen für diese Reihen doch verschiedene Wettervariablen signifikant geworden sind, was heißt, dass diese die Vorhersage des Folgemonats verbesserten.

Reihe	Unfälle / Verunglückte	RMSE			% Reduction	
		Wettermodell	BASt	Ohne Wetter	BASt	Ohne Wetter
G00	Getötete insgesamt	81	75	88	-8	8
GA1	Getötete Kinder unter 15 Jahren	8	9	8	11	0
GA2	Getötete im Alter von 15 bis 17 Jahren	12	14	11	14	-9
GA3	Getötete im Alter von 18 bis 24 Jahren	23	44	26	48	12
GA4	Getötete im Alter von 25 bis 64 Jahren	50	47	59	-6	15
GA5	Getötete im Alter von 65 Jahren und älter	32	48	38	33	16
GO1	Getötete innerorts	24	32	31	25	23
GO2	Getötete außerorts ohne BAB	57	72	76	21	25
GO3	Getötete auf BAB	31	31	32	0	3
GO4	Getötete auf Bundesstraßen außerorts	39	40	40	3	3
GV1	Getötete Fahrer und Mitfahrer von Kfz. (ohne mot. Zweiräder)	57	62	61	8	7
GV2	Getötete Fahrer und Mitfahrer von Pkw	55	62	59	11	7
GV3	Getötete Fahrer und Mitfahrer von Güterkraftfahrzeugen	10	10	10	0	0
GV4	Getötete Fahrer und Mitfahrer von Kraftfahrzeugen/Kraftrollern	24	26	31	8	23
GV5	Getötete Fahrer und Mitfahrer von Moped/Mofas	6	9	6	33	0
GV6	Getötete Fahrer und Mitfahrer von Fahrrädern	22	24	25	8	12
GV7	Getötete Fußgänger	28	34	30	18	7
U00	Polizeilich registrierte Unfälle insgesamt	16.713	31.981	15.068	48	-11
UP0	Unfälle mit Personenschaden insgesamt	2.600	3.598	3.389	28	23
UP1	Unfälle mit Personenschaden innerorts	1.616	3.182	2.792	49	42
UP2	Unfälle mit Personenschaden außerorts ohne BAB	1.009	1.313	1.133	23	11
UP3	Unfälle mit Personenschaden auf BAB	329	614	321	46	-2
UP4	Unfälle mit Personenschaden auf Bundesstraßen außerorts	343	481	385	29	11
UPA	Alkoholunfälle mit Personenschaden	319	475	389	33	18
USS	Schwerwiegende Unfälle mit Sachschaden (i.e.S.)	1.366	1.383	1.309	1	-4
V00	Verunglückte insgesamt	3.754	4.843	4.529	22	17
VSV	Schwerverletzte insgesamt	858	1.888	1.345	55	36

Tab. 4: Vorhersagegenauigkeit (RMSE) per Reihe für das Wettermodell, die Vorhersagen des BASt Teams (BASt) und einem Modell ohne Wettervariablen

## 3.2 Outputs per Zeitreihe

Für jede Zeitreihe werden bei der Prognose zwei Outputdateien generiert, eine tabellarische und eine Datei mit grafischem Output. Der grafische Output für jede Reihe ist im technischen Anhang wiedergegeben. Am Beispiel der Zeitreihe G00 (Getötete insgesamt) wird hier der Output erklärt.

### 3.2.1 Tabellarisch

In Tabelle 5 wurden die endgültigen Werte der vergangenen Jahre ergänzt durch die geschätzten Anzahlen des laufenden Jahres.

Im unteren Teil der Tabelle werden die beobachteten endgültigen Anzahlen der vergangenen Jahre pro

Monat wiedergegeben. Dabei beinhaltet die obere Reihe die Monatszahlen und die untere Reihe (mit orangem Hintergrund) die kumulierte Häufigkeit. Im oberen Teil der Tabelle werden die vorläufigen Anzahlen der ersten Monate in blau angegeben und die durch das Wettermodell geschätzten endgültigen Anzahlen in grün. Die Wettereffekte für jeden Monat sind in fuchsia angegeben. Sie geben das Verhältnis zwischen der Standardschätzung und der unter Annahme von durchschnittlichem Wetter. Alle Schätzwerte sind einerseits für den jeweiligen Monat angegeben (mit weissem Hintergrund) und andererseits kumuliert über alle vorhergehenden Monate des laufenden Jahres (mit blauem Hintergrund).



G00: Getötete insgesamt												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2016	242	191	190	246	320	259	331	327	322			
kum	242	433	623	869	1189	1448	1779	2106	2428			
	242	189	188	245	320	259	333	331	321	261	247	265
kum	242	431	619	864	1184	1443	1776	2107	2428	2689	2936	3201
	264	194	208	268	335	292	357	349	306	302	265	265
kum	264	458	666	934	1269	1561	1918	2267	2573	2875	3140	3405
2015	223	202	230	284	312	343	348	335	316	286	294	286
kum	223	425	655	939	1251	1594	1942	2277	2593	2879	3173	3459
2014	219	186	244	297	286	339	341	283	327	292	262	301
kum	219	405	649	946	1232	1571	1912	2195	2522	2814	3076	3377
2013	208	181	226	240	279	305	374	323	338	310	257	298
kum	208	389	615	855	1134	1439	1813	2136	2474	2784	3041	3339

Tab. 5: Tabellarischer Output am Beispiel der Reihe G00 (Getötete Insgesamt)

### 3.2.2 Grafischer Output

#### Vorhersagen und Wettereffekte

Für jede Reihe werden die durch das Wettermodell vorhergesagten Anzahlen für jeden Monat des laufenden Jahres abgetragen (grüne Linie). Der Wettereffekt wird als Pfeil nach oben oder unten angegeben. Die Endpunkte der Pfeile geben den wetterkorrigierten Wert an, nämlich die geschätzte Anzahl wenn man im Wettermodell die beobachteten Wetterdaten ersetzt durch ihren Langzeitmittelwert. In Bild 11 liegen beispielsweise für die Reihe G00 (alle Getöteten) in 2016 die wetterbereinigten Werte beinahe alle über denen, die aufgrund der tatsächlich beobachteten Wetterbedingungen geschätzt sind.

#### Validierung

Zur Modellvalidierung ist der Vorhersagefehler verschiedener Vorhersagen für die letzten 5 Jahre geplottet. Neben den Vorhersagen, die die BAST pub-

liziert hat, werden immer zwei Vorhersagen evaluiert, die das Wettermodell unter verschiedenen Annahmen für die vergangenen Jahre gemacht hätte. Das BAST-Team erstellt seine Vorhersagen meist Anfang Dezember oder Ende November. Davon abhängig beinhaltet der Modellinput entweder die Wetterdaten bis einschließlich November oder nur bis Oktober. Für die Monate ohne beobachtete Wetterdaten werden, wie in Kapitel 2.2.6 beschrieben, die Wetterwerte mit einem Zeitreihenmodell vorhergesagt. Je mehr Monate vorhergesagt werden müssen, desto größer ist die Qualitätseinbuße. Die Vorhersagegenauigkeit des Wettermodells in den letzten Jahren wird darum einmal unter der Annahme berechnet, dass die Wetterdaten bis einschließlich November erhältlich sind und einmal unter der Annahme, dass für November noch keine Wetterdaten erhältlich sind. In beiden Fällen wird der rmse (Root means squared error) berechnet wie in Kapitel 3.1 beschrieben.

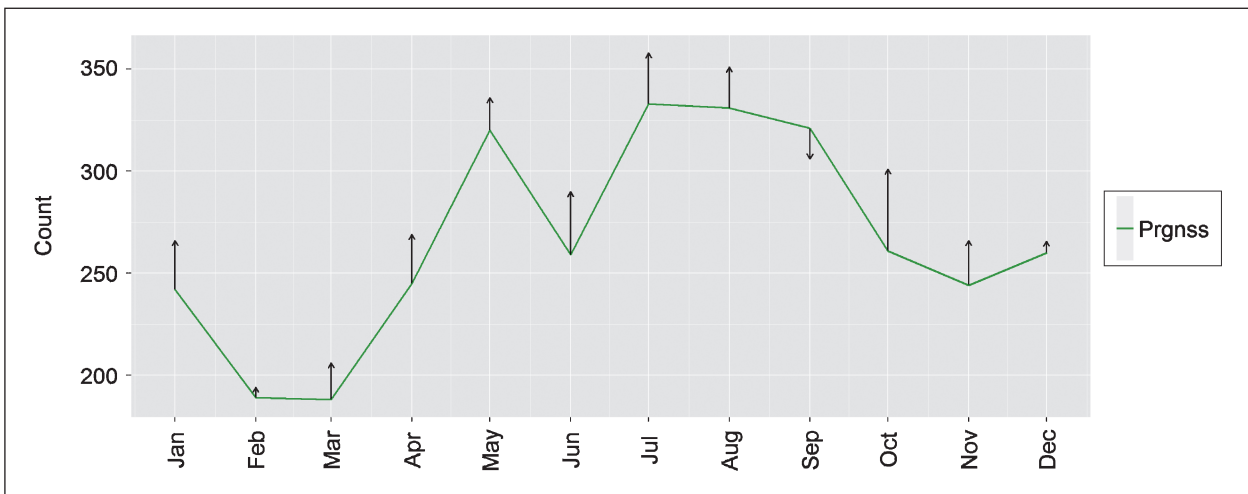


Bild 11: Grafischer Output zur Evaluation der Wettereffekte am Beispiel der Reihe G00 (Getötete Insgesamt)

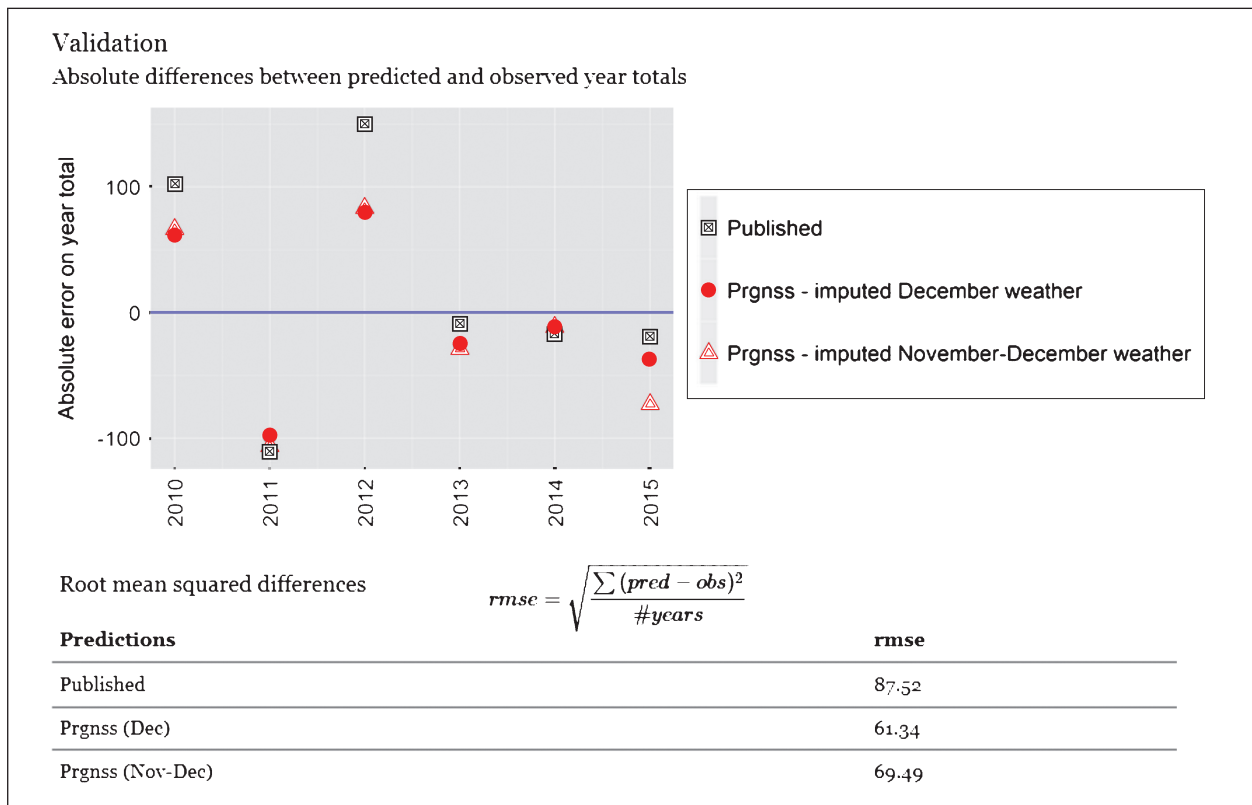


Bild 12: Grafischer Output zur Evaluation der Vorhersagegenauigkeit am Beispiel der Reihe G00 (Getötete Insgesamt)

## 4 Diskussion

So wie in vielen Ländern ist auch in Deutschland das Zeitintervall zwischen Unfallgeschehen und der Veröffentlichung der offiziellen Unfallzahlen relativ lang. Das Ziel dieser Studie ist die Entwicklung eines Verfahrens, um auf der Basis der monatlichen Unfall- und Verunglücktenzahlen von Januar bis August/September die Jahresendwerte vorherzusagen.

Kernstück des hier verwendeten Ansatzes ist ein strukturelles Zeitreihenmodell, das anhand der monatlichen Daten seit 1991 kalibriert wurde. Diese Daten enthalten nicht nur die Unfall- und Verunglücktenzahlen, sondern auch meteorologische Prädiktoren. Die Evaluation der Zeitreihenmodelle ergibt, dass – verglichen mit dem bisher benutzten heuristischen Ansatz – die hier entwickelten Modelle die Vorhersagegenauigkeit bis zu 55 % erhöhen. Lediglich zwei der 27 Reihen zeigten eine geringfügige Verschlechterung. Wir können also folgern, dass der Modell-basierte Ansatz eine gute Alternative darstellt, wenn es darum geht, Unfalldaten für das gerade beendete Jahr zur Verfügung zu stellen.

### 4.1 Konstruktionsprinzipien

Der hier beschriebene Modellierungsansatz wurde spezifisch für die Jahresendvorhersage entwickelt. Alle Modellier- und Vorhersageschritte sind als unabhängig laufendes Modul implementiert und sind in praktischen Erwägungen begründet. Die Behandlung von Witterungseinflüssen ist automatisiert – und zwar nicht nur in der Schätzung der Parameter, sondern auch in der Selektion der Prädiktoren. Das hat den Vorteil, dass in zukünftigen Jahren das Modell automatisch stets wieder kalibriert wird und damit auch sich ändernden Witterungs- oder Unfallbedingungen gerecht werden kann.

Um eine automatische Prozedur zu erreichen, muss diese über die 27 Zeitreihen homogen sein. Obwohl möglicherweise die Performanz innerhalb jeder einzelnen Zeitreihe durch zeitreihen-spezifische Prinzipien verbessert werden könnte, wird der Vorteil der sich selbst kalibrierenden Modelle, die alle demselben Typ entsprechen, als größer eingeschätzt als eine individuelle Behandlung jeder Zeitreihe.

Um optimale Vorhersagen zu erreichen, werden die ursprünglichen Messwerte in mehreren Schritten vorverarbeitet. Die täglichen Wetterdaten der indivi-

duellen Stationen werden zuerst transformiert, dann skaliert, und schließlich zu monatlichen nationalen Werten aggregiert. Unter der Annahme, dass in dicht besiedelten Gebieten auch der meiste Verkehr ist, werden in der Aggregation die am dichtesten besiedelten Gebieten am stärksten gewichtet. Schließlich werden diese Werte in einer Hauptkomponenten-Analyse noch einmal transformiert, weil die Wettervariablen untereinander stark korreliert sind. Die Aufnahme der resultierenden orthogonalen Komponenten stabilisiert die Schätzung der Koeffizienten und macht sie robuster gegenüber kleinen Veränderungen des Dateninputs. Außerdem verhindert die Identifikation von Komponenten, die keine signifikante Beziehung zum Unfallgeschehen aufweisen, dass die Modelle overfitted werden. Overfitting (also die Aufnahme von Variablen, die lediglich zufällig einen Zusammenhang mit dem zu modellierenden Kriterium zeigen) verschlechtert die Generalisierbarkeit der Resultate.

## 4.2 Evaluation der Modelle

Die Resultate bestätigen den zuvor schon gefundenen Zusammenhang zwischen Wetter und Unfallgeschehen (z. B. BERGEL-HAYAT et al., 2013; BIJLEVELD & CHURCHILL, 2009; HERMANS et al., 2006; MARTENSEN & FOCANT, 2014). Sie demonstrieren auch, dass die Nutzung von Wettervariablen die Vorhersageleistung von Regressionsmodellen verbessert, vor allem da die Witterungsumstände wesentlich früher erhältlich sind als die Unfalldaten. Im Vergleich zwischen Modellen mit und ohne Wettervariablen zeigt sich, dass mit nur wenigen Ausnahmen durch die Aufnahme von meteorologischen Prädiktorvariablen eine erhöhte Vorhersagegenauigkeit erreicht wird. Lediglich in vier Reihen sehen wir eine geringfügige Verschlechterung der Vorhersage. Die Tatsache, dass auch in diesen vier Reihen eine signifikante Beziehung zwischen Wetter und Unfallgeschehen identifiziert wird, deutet darauf hin, dass der möglicherweise komplexe Zusammenhang zwischen Wetter- und Unfallvariablen in den hier verwendeten Modellen noch nicht vollständig erfasst ist. Dennoch bietet der hier vorgestellte Ansatz eine deutliche Verbesserung der Vorhersagegenauigkeit gegenüber dem bisher verfolgten heuristischen Ansatz.

## 4.3 Einschränkungen und Ausblick

Durch die Reihe von Vorverarbeitungsschritten erlaubt die hier beschriebene Methodologie keine direkte Interpretation der Regressionskoeffizienten auf das Unfallgeschehen (wie etwa „eine Zunahme der Temperatur um 1 Grad führt zu einer Zunahme der Anzahl der Getöteten um xxx %“). Hier wurde also die Verbesserung der Vorhersagen über die Interpretierbarkeit der Resultate gestellt.

Die angewendeten Transformationen stehen einer Interpretation nicht prinzipiell entgegen. Zum Beispiel hilft die Hauptkomponentenanalyse bei der Identifikation von latenten Dimensionen im Wettergeschehen. Die damit verbundene Abstraktion von physikalischen Messwerten zu Wetter-Szenarien eignet sich auf lange Sicht wahrscheinlich besser dazu, den Zusammenhang zwischen Wetter und Unfallgeschehen zu beschreiben als die Betrachtung einzelner Variablen. Weitere Forschung bezüglich der inhaltlichen Interpretation der resultierenden Hauptkomponenten in diesem Sinne würde eine gute Basis für ein weiteres Fine-tuning der Modelle darstellen.

Wie oben schon im Kapitel 4.2 angegeben, implizieren die Resultate an einigen Stellen, dass die derzeitigen Modelle die Beziehung zwischen Wetter und Unfalldaten noch nicht vollständig abbilden. Vor allem der lineare Charakter (nach Transformation) und die beschränkte Aufnahme von Interaktionen werden der wirklichen Beziehung wahrscheinlich nicht gerecht. In Bezug auf Komplexität, gehen die jetzigen Modelle einerseits bis zur Aufnahme von Interaktionen zwischen jeweils zwei Variablen und andererseits zur zusätzlichen Aufnahme eines jeden Wertes in der Form von Abweichungen vom Monats- und Stationsmittelwerts. Dadurch ist es zum Beispiel möglich, Unterschiede im Effekt von starkem Regenfall bei hohen und bei niedrigen Temperaturen zu modellieren. Obwohl die Z-Transformation, die den Ausnahmegehalt eines Messwertes wiedergibt, eine spezifische Form von Non-Linearität ist, könnte die Präzision der Vorhersagen durchaus auch durch andere Formen von nicht linearen Wettereffekten zu steigern sein. Zum Beispiel birgt Regen, wenn er auf eine lange Trockenheit folgt, einen besonders hohes Risiko, da sich der Staub auf der Straße mit dem Wasser zu einer schmierigen Masse vermischt, die besonders unfallträchtig ist (EISENBERG, 2004). Phänomene dieser Art zu modellieren bleibt eine Herausforderung und müsste

für jede Zeitreihe einzeln geplant und evaluiert werden.

Für die Spezifikation von Gegenmaßnahmen wäre es auch wichtig, die tatsächliche Wirkungsweise der Wettereffekte zu identifizieren. Wie in der Einleitung schon erwähnt, beruht der Zusammenhang auf zwei verschiedenen Mechanismen. Einerseits beeinflussen die Witterungsumstände das Reiseverhalten und damit das Verkehrsaufkommen der verschiedenen Arten der Verkehrsteilnahme (SABIR, 2011). Andererseits kann das Wetter auch das Unfall- (oder Verletzungs-)risiko als solches beeinflussen. Dies kann zum Beispiel durch verminderte Sicht oder Fahrbahnhaftung entstehen. Im hiesigen Modellierungsansatz wird zwischen diesen beiden Typen nicht unterschieden. Dies könnte beispielsweise im „Latent Risk“ Ansatz modelliert werden, in dem die Entwicklung des Risikos getrennt von der Entwicklung des Verkehrsvolumens abgebildet wird (BIJLEVELD et al., 2008). Der Hauptgrund, dies nicht zu tun, liegt in der mangelnden Verfügbarkeit von (monatlichen) Verkehrs-/ Mobilitätsdaten. Insbesondere für Fußgänger und Zweiradfahrer haben die Daten keinesfalls die nötige Zeitauflösung und Aktualität, obwohl gerade bei dieser Gruppe das Mobilitätsverhalten stark durch das Wetter mitbestimmt wird. Für das Ziel der vorliegenden Studie – die Qualität der Vorhersagen – ist die getrennte Schätzung von Auswirkungen des Wetters auf das Risiko und Auswirkungen auf das Verkehrsaufkommen jedoch nicht entscheidend.

Inhaltlich wurden die Prädiktoren in dieser Studie beschränkt auf Wetterindikatoren. Natürlich gibt es auch andere Variablen, die Einfluss auf das Verkehrs- und Unfallgeschehen haben können. Hauptsächlich sind dabei Konjunkturindikatoren, wie zum Beispiel die Arbeitslosenquote oder das Bruttoinlandsprodukt, zu nennen. In einem Report der OECD wurde festgestellt, dass wirtschaftlicher Rückgang – etwa wie er seit der Krise im Jahr 2007 in vielen europäischen Ländern mehr oder weniger lang anhaltend zu beobachten war – häufig verbunden ist mit einem starken Rückgang der Verunglücktenzahlen. Obwohl der Zusammenhang noch nicht vollständig erklärt werden kann, setzt er sich wahrscheinlich zusammen aus einer allgemeinen Verringerung des Verkehrsaufkommens, einem verminderten Zugang von Risikogruppen (wie zum Beispiel jungen Männern) zu Fahrzeugen und Fahrerlaubnis und eventuell einer langfristigen Einstellungsänderung, die zu einem vorsichtigeren Fahrverhalten führt (OECD/ITF, 2015). Es handelt

sich bei diesen Entwicklungen allerdings zumeist um langfristige Phänomene. Der Effekt von Konjunkturdaten ist darum nicht unbedingt relevant für die kurzfristige Vorhersage der Jahresenddaten.

## 5 Schlussfolgerung

Die Anwendung von strukturellen Zeitreihenmodellen führt zu einer deutlichen Verbesserung der Vorhersagegenauigkeit der Jahresendvorhersagen der Unfall- und Verunglücktenzahlen in Deutschland. Dieser Ansatz stellt also eine gute Alternative zum heuristischen Ansatz dar, wie er bislang gewählt wurde.

Die Studie zeigt außerdem den Einfluss der Witterungsbedingungen auf das Unfallgeschehen. Zeitreihenmodelle mit meteorologischen Prädiktoren stellen ein leistungsfähiges Werkzeug dar, mit dem dieser Zusammenhang erfasst werden kann.

## 6 Literatur

- BBR (2011): Aktualisierte und erweiterte Testreferenzjahre von Deutschland für mittlere, extreme und zukünftige Witterungsverhältnisse. Offenbach: Bundesamtes für Bauwesen und Raumordnung
- BERGEL-HAYAT, R.; DEBBARH, M.; ANTONIOU, C.; YANNIS, G. (2013): Explaining the road accident risk: weather effects. *Accident Analysis and Prevention*, 60, pp. 456-465
- BERGEL-HAYAT, R.; ZUKOWSKA, J. (2015): Road safety trends at national level in Europe: a review of time-series analysis performed during the period 2000-2012. *Transport Reviews: A Transnational Transdisciplinary Journal*, DOI: 10.1080/01441647.2015.1030005.
- BIJLEVELD, F.; CHURCHILL, T. (2009): The influence of weather conditions on road safety. An assesment of the effect of precipitation and temperature. Leidschendam (The Netherlands): Swov Institute for Road Safety Research
- BIJLEVELD, F.; COMMANDEUR, J.; GOULD, P.; KOOPMAN, S. (2008): Model-based measurement of latent risk in time series with applications. *Journal of the Royal Statistical Society, Series A*

- BOX, G.; JENKINS, G. (1970): *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day
- CHRISTOFFER, J.; DEUTSCHLÄNDER, T.; WEBS, M. (2004): TRY – Testreferenzjahre von Deutschland für mittlere und extreme Witterungsverhältnisse. Deutscher Wetterdienst
- DUPONT, E.; MARTENSEN, H. (2012): *Forecasting road traffic fatalities in European countries*. Brussels: Deliverable 4.4 of the EC PF7 project DaCoTA
- DUPONT, E.; COMMANDEUR, J.; LASSARRE, S.; BIJLEVELD, F.; MARTENSEN, H.; ANTONIOU, C.; GIUSTINIANI, G. (2014): Latent risk and trend models for the evolution of annual fatality numbers in 30 European countries. pp. 327-336: *Accident Analysis and Prevention* 71
- DURBIN, J. (2001): *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press
- DURBIN, J.; KOOPMAN, S. J. (2012): *Time Series Analysis by State Space Methods*. Second Edition. Oxford: Oxford University Press
- EISENBERG, D. (2004): The mixed effects of precipitation on traffic crashes. *Accident Analysis and Prevention*, 36: pp. 637-647
- FOCANT, N.; MARTENSEN, H. (2014): Y a-t-il plus d'accidents de la route quand il pleut? Analyse exploratoire de l'influence des conditions météorologiques sur le nombre d'accidents de la route en Belgique. Bruxelles: Centre de Connaissance Sécurité Routière, BIVV-IBSR
- HARVEY, A. (1989): *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press
- HERMANS, E.; WETS, G.; van den BOSSCHE, P. (2006): Frequency and severity of Belgian road traffic accidents studied by state-space methods. *Journal of Transportation and Statistics* 9, pp. 63-75
- KALMAN, R. (1960): A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 82, doi:10.1115/1.3662552
- KOOPMAN, S. J.; HARVEY, A. C.; DOORNIK, J. A.; SHEPHARD, N. (1999a): *Structural Time Series Analysis, Modelling, and Prediction using STAMP*. London: Timberlake Consultants Press
- LERNER, M.; SCHEPERS, A.; PÖPPEL-DECKER, M.; LEIBNITZ, J.; FITSCHEN, A. (2013): *Voraussichtliche Entwicklung von Unfallzahlen und Jahresfahrleistung in Deutschland – Ergebnisse 2013*. Bergisch Gladbach: Bundesanstalt für Straßenwesen
- LERNER, M.; SCHEPERS, A.; PÖPPEL-DECKER, M.; LEIPNITZ, C.; FITSCHEN, A. (2014): *Voraussichtliche Entwicklung von Unfallzahlen und Jahresfahrleistung in Deutschland – Ergebnisse 2014*. Bergisch Gladbach: Teil der Vergabeunterlagen
- MARTENSEN, H.; BERGEL-HAYAT, R. (in Vorbereitung): *Factsheet forecasting – Joint methodology for IRTAD countries*. Draft version. IRTAD
- MARTENSEN, H.; DUPONT, E. (2010): *Forecasting road traffic fatalities in European countries: model and first results*. Brussels: Deliverable 4.2 of the EC FP7 project DaCoTA
- SABIR, M. (2011): *Weather and travel behaviour*. Vrije Universiteit Amsterdam, Amsterdam
- OECD/ITF (2015): *Why does road safety improve when economic times are hard?* <http://www.itf-oecd.org/why-does-road-safety-improve-when-economic-times-are-hard> (last accessed on May 5<sup>th</sup>, 2017)

## APPENDIX

### A1 CRASH DATA

Input crash data are read from the Excel file `prgnss\data\master.xlsx`. The data import is handled by the script `prgnss\src\read_master.r`. The script first looks for the sheet with the name *Reihen* that lists the different data series:

series	series.de	series.en
U00	Polizeilich registrierte Unfälle insgesamt	total number of police recorded accidents
UP0	Unfälle mit Personenschaden insgesamt	total number of injury accidents
USS	Schwerwiegende Unfälle mit Sachschaden (i.e.S.)	severe accidents with damage only (towed away)
G00	Getötete insgesamt	fatalities total
V00	Verunglückte insgesamt	total number of casualties
GO1	Getötete innerorts	fatalities inside urban areas
GO2	Getötete außerhalb ohne BAB	fatalities outside urban areas without motorways
GO3	Getötete auf BAB	fatalities on motorways
GO4	Getötete auf Bundesstraßen außerhalb	fatalities on national roads outside urban areas
GA1	Getötete Kinder unter 15 Jahren	fatalities under 15 years
GA2	Getötete im Alter von 15 bis 17 Jahren	fatalities from 15 to 17 years
GA3	Getötete Personen im Alter von 18 bis 24 Jahren	fatalities from 18 to 24 years
GA4	Getötete im Alter von 25 bis 64 Jahren	fatalities from 25 to 64 years
GA5	Getötete Personen im Alter von 65 Jahren und älter	fatalities 65 years and older
GV1	Getötete Fahrer und Mitfahrer von Kfz. (ohne mot. Zweiräder)	fatally injured user of motor vehicles (without motorised two wheelers)
GV2	Getötete Fahrer und Mitfahrer von Pkw	fatally injured user of cars
GV3	Getötete Fahrer und Mitfahrer von Güterkraftfahrzeugen	fatally injured user of goods vehicles
GV4	Getötete Fahrer und Mitfahrer von Krafträdern/Kraftrollern	fatally injured user of motor bikes
GV5	Getötete Fahrer und Mitfahrer von Moped/Mofas	fatally injured user of mopeds
GV6	Getötete Fahrer und Mitfahrer von Fahrrädern	fatally injured user of bicycles
GV7	Getötete Fußgänger	fatally injured pedestrians
UP1	Unfälle mit Personenschaden innerorts	injury accidents inside urban areas
UP2	Unfälle mit Personenschaden außerhalb ohne BAB	injury accidents outside urban areas without motorways
UP3	Unfälle mit Personenschaden auf BAB	injury accidents on motorways
UP4	Unfälle mit Personenschaden auf Bundesstraßen außerhalb	injury accidents outside urban areas without motorways
UPA	Alkoholunfälle mit Personenschaden	injury accidents under the influence of alcohol
VSV	Schwerverletzte insgesamt	total number of seriously injured (hospitalized)

Next, it goes through all the individual data sheets whose name occurs in the first column of *Reihen*. This implies that users have the possibility to add data series to the analysis or to remove them. In the case of adding, the user should (1) add a row in *Reihen* (with a new abbreviation, German and English name) and (2) add a new data sheet whose name matches the chosen abbreviation.

The data sheets have the following format:

year	data	Jan	Feb	Mrz	Apr	Mai	Jun	Jul	Aug	Sep	Okt	Nov	Dez
1991	provisional	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
1991	final	864	702	801	840	976	1053	1069	1070	998	1041	945	941
1992	provisional	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
1992	final	800	762	740	801	926	915	966	1009	928	967	963	854
1993	provisional	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
1993	final	840	616	728	705	911	874	938	903	858	868	778	930
1994	provisional	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
1994	final	800	596	676	733	807	909	938	886	818	909	867	875
1995	provisional	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
1995	final	678	644	714	721	833	874	916	876	826	881	766	725
1996	provisional	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
2013	provisional	209	181	228	240	277	302	370	314	345	#N/A	#N/A	#N/A
2013	final	208	181	226	240	279	305	374	323	338	310	257	298
2014	provisional	218	186	245	296	284	339	344	278	337	#N/A	#N/A	#N/A
2014	final	219	186	244	297	286	339	341	283	327	292	262	301
2015	provisional	224	202	234	287	315	338	344	326	321	#N/A	#N/A	#N/A
2015	final	223	202	230	284	312	343	348	335	316	286	294	286
2016	provisional	242	191	190	246	320	259	331	327	322	#N/A	#N/A	#N/A
2016	final	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

Per year, there are two rows. One with the provisional crash counts and one with the consolidated (final) crash counts. Wherever data is missing, this is indicated with #N/A. Provisional data is only available since the year 2000. Final data are available 9 years earlier (1991). Unless there is a clear a priori reason, the user is advised *not* to shift the onset of the final data to a more recent date. They are the main input for the *prgnss* predictions and any data restrictions can have a negative effect on the quality of predictions. If needed, however, any data can be excluded by manually replacing values with #N/A.

Before running *prgnss* the user should make sure that the master input file is fully updated and coherent. This also implies updating the final sheet *Publiziert*, which lists the counts that were published by BASt for each crash series in previous years. These numbers are used for comparison in the validation output.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
U00	2350000	2300000	2260000	2249567	2240000	2230000	2290000	2270000	2290000	2340000	2340000	2375000	2377000	2370000	2470000
UP0	371000	365000	353500	334500	335000	323000	339000	321000	311000	289000	299900	302000	287000	298000	301000
USS	108500	106600	103000	100615	98000	96400	96500	90000	88400	90000	83000	83000	82550	72500	70000
G00	7050	6910	6575	5750	5400	5050	5070	4500	4150	3750	3898	3750	3330	3360	3440
V00	498000	489000	471000	440000	438000	419000	440000	414000	403000	380000	390000	392500	373000	387000	391000
GO1	1730	1680	1654	1412	1470	1420	1320	1260	1210	1075	1100	1100	970	980	1020
GO2	4520	4440	4070	3680	3300	2970	3132	2730	2421	2232	2357	2250	1920	1990	2030
GO3	760	790	847	631	660	655	640	510	455	435	420	400	440	390	380
GO4	1750	1740	1615	1450	1250	1160	1170	984	900	835	900	870	730	720	790
GA1	250	220	196	150	173	126	120	105	92	100	84	75	68	55	80
GA2	310	300	290	280	250	179	160	172	144	105	125	110	92	95	70
GA3	1650	1590	1440	1230	1149	955	1040	845	800	720	720	680	510	490	476

## A2 WEATHER DATA

### A2.1 RAW WEATHER DATA

Raw weather data is downloaded from the Deutscher Wetterdienst (DWD) at the following location:  
[ftp://ftp-cdc.dwd.de/pub/CDC/observations\\_germany/climate/daily/kl/](ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/daily/kl/)

The R-code is provided in the script *prgnss\src\download\_dwd.r*. The data represent daily values since January 1<sup>st</sup> 1991 for the following variables:

*LUFTTEMPERATUR, DAMPFDRUCK, BEDECKUNGSGRAD, LUFTDRUCK\_STATIONSHOEHE, REL\_FEUCHTE, WINDGESCHWINDIGKEIT, LUFTTEMPERATUR\_MAXIMUM, LUFTTEMPERATUR\_MINIMUM, LUFTTEMP\_AM\_ERDB\_MINIMUM, WINDSPITZE\_MAXIMUM, NIEDERSCHLAGSHOEHE, NIEDERSCHLAGSHOEHE\_IND, SONNENSCHEINDAUER, SCHNEEHOEHE*

The nature of these variables is detailed in [ftp://ftp-cdc.dwd.de/pub/CDC/observations\\_germany/climate/daily/kl/recent/BESCHREIBUNG\\_obsgermany\\_climate\\_daily\\_kl\\_recent\\_de.pdf](ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/daily/kl/recent/BESCHREIBUNG_obsgermany_climate_daily_kl_recent_de.pdf)

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 the table below. The selected weather stations appear in italics.

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

More information about these weather stations is available in:

[ftp://ftp-cdc.dwd.de/pub/CDC/help/stations\\_map\\_soil.png](ftp://ftp-cdc.dwd.de/pub/CDC/help/stations_map_soil.png)

and

[ftp://ftp-cdc.dwd.de/pub/CDC/observations\\_germany/climate/daily/kl/recent/KL\\_Tageswerte\\_Beschreibung\\_Stationen.txt](ftp://ftp-cdc.dwd.de/pub/CDC/observations_germany/climate/daily/kl/recent/KL_Tageswerte_Beschreibung_Stationen.txt)

After downloading, the data is stored in *prgnss\data\weather\_raw.csv*. This file also lists all missing values up until the last day of the current month (at the time of downloading). By default, the raw weather data is updated incrementally, i.e., data from an earlier download is not overwritten unless there was an update of historic data on the DWD server. To override this behaviour, i.e., to initiate an update of all data since 01/01/1991, the user should delete the following Rdata file prior to downloading: *prgnss\data\\_rda\weather\_raw.rda*.

When running *prgnss* through the main command file *prgnss.cmd*, the raw weather data are updated automatically. By default, the program will only take into account weather data from 01/01/1991 until the first day of the current month (to avoid averaging over a limited amount of data for the current month). The month values of the current month are imputed through time series modelling, as detailed in Section A2.4.

The user can, however, change this default behaviour by executing the alternative *prgnss\_custom.cmd*. There are two differences with the process in *prgnss.cmd*: (a) there is no automatic update of the raw weather variables and (b) the program will take into account all weather data, as given in *prgnss\data\weather\_raw.csv*, excluding missing values. This also enables the user to manipulate raw weather data (i.e., the values in *prgnss\data\weather\_raw.csv*) – for instance, impute missing daily values for the current month. Importantly, when manually imputing missing values, the user should do this separately for each weather station. In order to

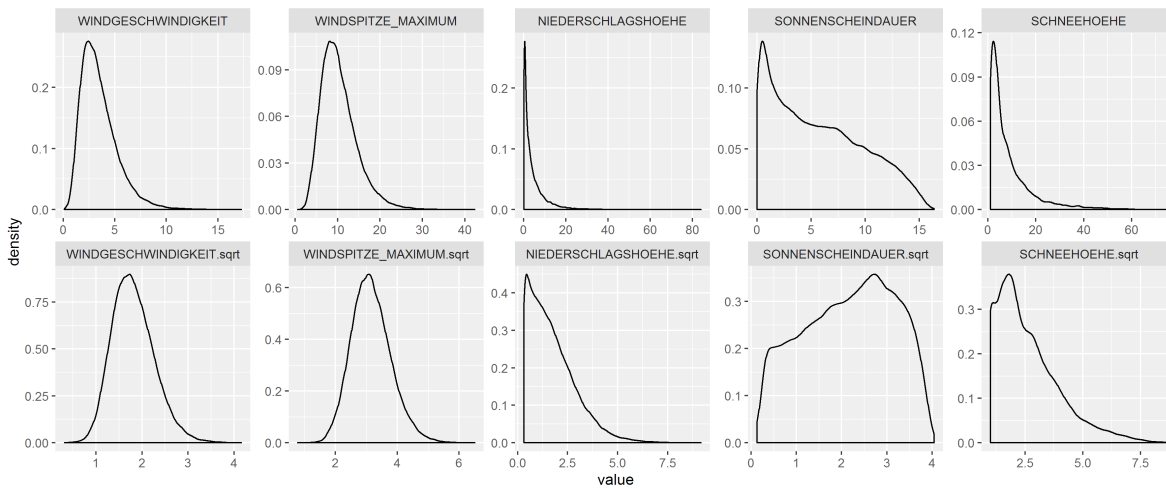


prepare `prgnss\data\weather_raw.csv` manually, the user should update the file with the most recent available data first. This is possible by executing the command script `dwd.cmd`.

## A2.2 AGGREGATED WEATHER DATA

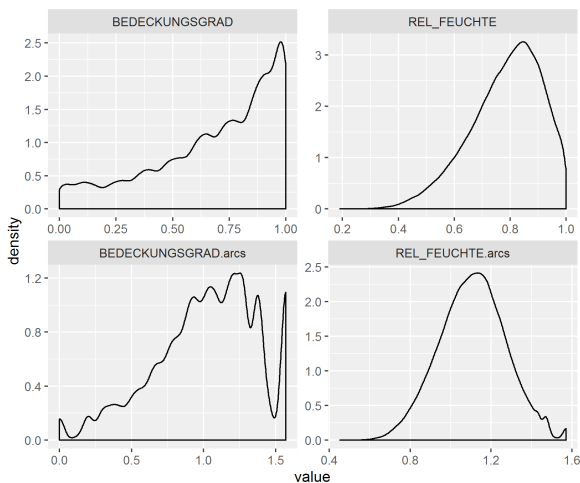
The daily station-level values are aggregated into monthly values. This process is implemented in the R-script `prgnss\src\aggregate_dwd.r`. Before aggregation, the daily station-level values undergo a number of transformations.

The variables `WINDGESCHWINDIGKEIT`, `WINDSPITZE_MAXIMUM`, `NIEDERSCHLAGSHOEHE`, `SONNENSCHNEINDAUER` and `SCHNEEHOEHE` are transformed into their square root to reduce the right skew in their distribution (see Figure 1).



**Fig. 1:** Distribution Square root transformation of daily station-level values for `WINDGESCHWINDIGKEIT`, `WINDSPITZE_MAXIMUM`, `NIEDERSCHLAGSHOEHE`, `SONNENSCHNEINDAUER` and `SCHNEEHOEHE` (zero-values not shown)

The variables `BEDECKUNGSGRAD` and `REL_FEUCHTE` are also transformed to reduce the skew in their distribution but through the arcsine square root transformation:  $\sin^{-1} \sqrt{x/\max(x)}$  (see Figure 2).



**Fig. 2:** Distribution Arcsine square root transformation of daily station-level values for `BEDECKUNGSGRAD` and `REL_FEUCHTE`

The indicator variable *NIEDERSCHLAGSHOEHE\_IND* is translated onto four new variables:

**dry**

$$\sin^{-1} \sqrt{\text{monthly proportion of } NIEDERSCHLAGSHOEHE\_IND = 0}$$

**rainOnly**

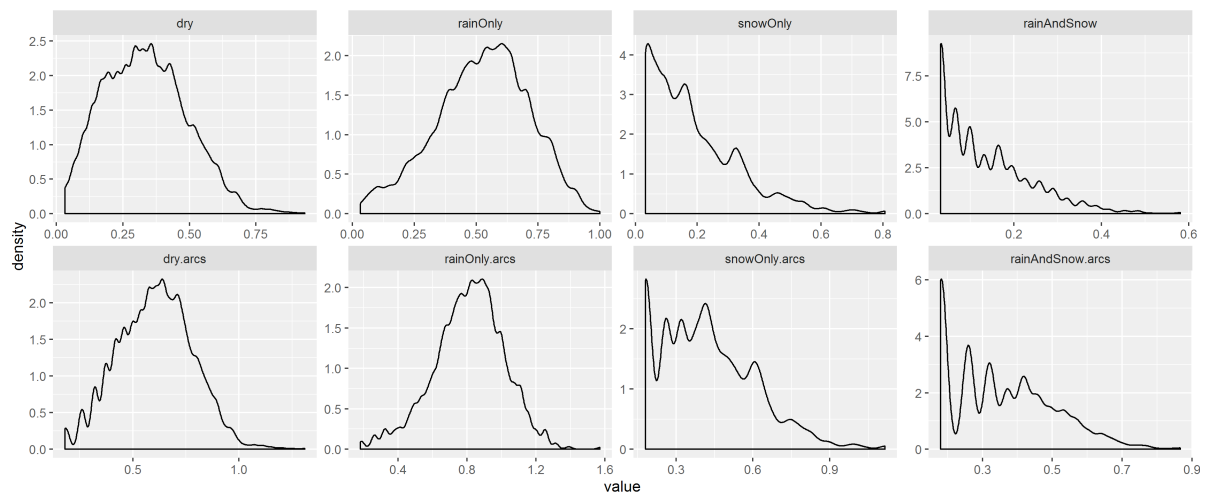
$$\sin^{-1} \sqrt{\text{monthly proportion of } NIEDERSCHLAGSHOEHE\_IND = 6}$$

**snowOnly**

$$\sin^{-1} \sqrt{\text{monthly proportion of } NIEDERSCHLAGSHOEHE\_IND = 7}$$

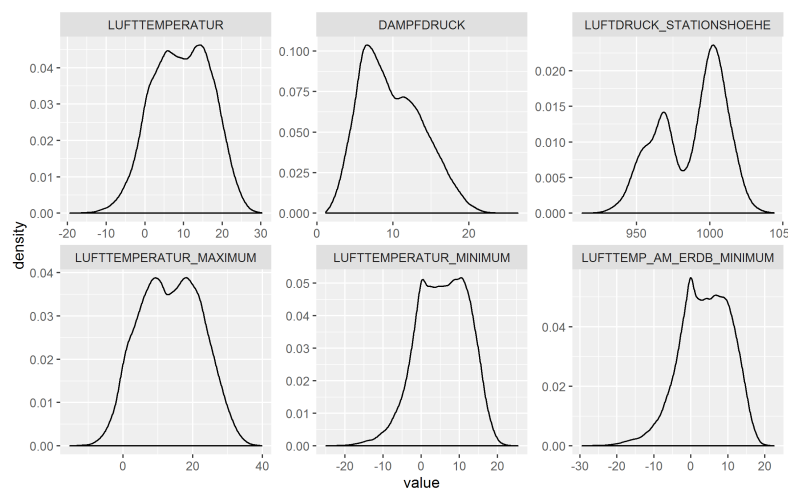
**rainAndSnow**

$$\sin^{-1} \sqrt{\text{monthly proportion of } NIEDERSCHLAGSHOEHE\_IND = 8}$$



**Fig. 3:** Distribution of the four variables that are created from *NIEDERSCHLAGSHOEHE\_IND*

The remaining variables, *LUFTTEMPERATUR*, *DAMPFDRUCK*, *LUFTDRUCK\_STATIONSHOEHE*, *LUFTTEMPERATUR\_MAXIMUM*, *LUFTTEMPERATUR\_MINIMUM* and *LUFTTEMP\_AM\_ERDB\_MINIMUM* are kept in their original scale.



**Fig. 4:** Distribution of the daily station-level values for the weather variables that remain untransformed prior to aggregation

After the initial variable-specific transformations, an additional set of variables is created wherein all original variables are expressed as station- and month-specific z-scores. For a given variable  $x$ , each individual score expresses the deviation on day  $i$  from the mean for month  $j$  at station  $k$  in standard deviation units:

$$z_{x_{ijk}} = \frac{x_{ijk} - \bar{x}_{jk}}{\sqrt{\text{Var}(x_{jk})}}$$

where  $\bar{x}_{jk}$  is the sample mean of variable  $x$  in month  $j$  at station  $k$  since 1991 and  $\text{Var}(x_{jk})$  the sample variance. The z-score variables serve as an additional set of input data, quantifying how usual or unusual the daily values are for a given variable at a given weather station, in the light of all other observations in the same month at that station since 1991.

The aggregation itself is performed in two steps. During the first step, the arithmetic mean of each variable is calculated per station, year and month – omitting missing daily values. If, for a given variable, station, year and month, all daily values are missing, the mean of that variable for that station and month is chosen.

During the second step, the weather data from the different weather stations are combined into a single value per variable, year and month. This is achieved by taking a weighted mean across stations, where the weight is proportional to the yearly population size associated with each station since 1991. The population data are obtained from Eurostat: [http://ec.europa.eu/eurostat/en/web/products-datasets/-/DEMO\\_R\\_PJANGROUP](http://ec.europa.eu/eurostat/en/web/products-datasets/-/DEMO_R_PJANGROUP)

The download procedure is implemented in the script `prgnss\src\download_eurostat.r`. The download can be initiated at any time by executing the command file `eurostat.cmd`. After downloading, the data can be viewed in the Excel file `prgnss\data\eurostat.xlsx`. The linkage between weather stations and NUTS2code is provided in `prgnss\data\_doc\climazonesXstations.txt`. At the time of execution, population data is often missing for the most recent (two) year(s). By default, the aggregation protocol will impute the missing values with the most recent available data (assuming no growth or decline in population size). The user can override this behaviour by manually imputing missing values in `prgnss\data\eurostat.xlsx`.

The weighted averages per variable, year and month are stored in `prgnss\data\_rda\weather_weighted.rda`.

### A2.3 ORTHOGONAL WEATHER DATA

The  $N \times M$  matrix of aggregated weather data, with  $N$  individual time points and  $M$  individual variables, is characterised by a naturally high degree of multicollinearity. In the current context, this is not a problem per se, since it is *not* within the scope of *prgnss* to link specific patterns in crash counts to individual weather variables. However, the aim of *prgnss* is to predict future crash counts, using only those weather variables that have a proven statistically significant relationship with crash counts. Hence, there is a need for variable selection during the calibration of the prediction models. Multicollinearity poses an important problem in this regard. Correlated weather variables map onto the same portion of the variance in the crash data, which leads to unstable results among them. A small change in the data can drastically change the pattern of estimated coefficients and the associated error distributions. Deciding which predictor variables can be dropped and which should be retained thus becomes a difficult task.

The solution that is used in *prgnss* is to transform the  $N \times M$  matrix of aggregated weather data into its principal components (PCs). These  $M$  PCs, by definition, provide orthogonal (uncorrelated) dimensions across all weather variables. They are ordered such that each PC accounts for the highest variability among the  $N$  time points, given the preceding PCs. Put differently, the first PC provides the dimension that allows the highest degree of discrimination between the time points (in terms of the source weather variables). The second PC provides the second highest degree, etc. PCs that explain less than 3% of the variance in the aggregated weather data, are discarded at this stage. Further selection is realised on the basis of the estimated relationships with crash counts. Since the PCs are uncorrelated, the calibration process is most straightforward (see Section A3.3).

## A2.4 IMPUTED WEATHER DATA

In the final stage, before models for crash data are built, *prgnss* determines whether or not weather data is available for the full extent of the crash count time series. As described in Section A2.1, by default, the program will only take into account PC values from 01/01/1991 until the first day of the current month. The missing values are imputed through the prediction of a classical structural time series model, fitted independently to each PC. Specifically, each individual PC is modelled as the sum of a stochastic level, slope, (monthly) seasonal and irregular process. The model is defined by the following equations:

$$y_t = \mu_t + \sum_{j=1}^{[s/2]} \gamma_{jt} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2) \sum_{j=1}^{[s/2]} \gamma_{jt}$$

$$\mu_{t+1} = \mu_t + \nu_t + \eta_t, \quad \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2)$$

$$\nu_{t+1} = \nu_t + \xi_t, \quad \xi_t \sim \mathcal{NID}(0, \sigma_\xi^2)$$

$$\gamma_{j,t+1} = \gamma_{jt} \cos \lambda_j + \gamma_{jt}^* \sin \lambda_j + \omega_{jt},$$

$$\gamma_{j,t+1}^* = -\gamma_{jt} \sin \lambda_j + \gamma_{jt}^* \cos \lambda_j + \omega_{jt}^*, \quad \omega_{jt}, \omega_{jt}^* \sim \mathcal{NID}(0, \sigma_\omega^2)$$

$$\lambda_j = 2\pi j/s$$

with

$y_t$  the observed value for a given PC at time  $t$ , with random disturbance  $\varepsilon_t$

$\mu_t$  the trend component, with random disturbance  $\eta_{t-1}$ , slope  $\nu_{t-1}$  and slope disturbance  $\xi_{t-1}$

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

All disturbances are modelled as independent stationary Gaussian processes. The parameters are estimated using the same generic procedure as for the crash data described in Section A3.1. The only difference is that the response variable at time  $t$  is assumed to follow normal distribution (instead of a log-normal distribution).

After estimation, the model is used to obtain the smoothed predictions

$$\hat{\alpha}_t = E(\alpha_t | y_1, \dots, y_n), \quad = (\hat{\mu}_t \hat{\nu}_t \hat{\gamma}_{1,t} \hat{\gamma}_{1,t}^* \dots \hat{\gamma}_{6,t})'$$

for  $t = n + 1, \dots, T$

with

$$\hat{\varepsilon}_t = \hat{\eta}_t = \hat{\xi}_t = \hat{\omega}_{jt} = \hat{\omega}_{jt}^* = 0$$

The time point  $n + 1$  corresponds to the first month for which all subsequent values of a given PC are missing. Time point  $T$  denotes December of the running year.

## A3 WEATHER MODEL

### A3.1 DEFINITION

Predictions in *prgnss* are based on independent models of the *final* crash data. These models extend the classical decomposition in structural time series modelling by including fixed weather effects (i.e., fixed regression components). By default, these models are recalibrated every time *prgnss* is run. The generic model is defined by the following equations:

$$y_t = \exp \left( \mu_t + \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt} + \sum_{k=1}^r \beta_k x_{kt} + \varepsilon_t \right), \quad \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2) \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt}$$

$$\mu_{t+1} = \mu_t + \nu_t + \eta_t, \quad \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2)$$

$$\nu_{t+1} = \nu_t + \xi_t, \quad \xi_t \sim \mathcal{NID}(0, \sigma_\xi^2)$$

$$\gamma_{j,t+1} = \gamma_{jt} \cos \lambda_j + \gamma_{jt}^* \sin \lambda_j + \omega_{jt},$$

$$\gamma_{j,t+1}^* = -\gamma_{jt} \sin \lambda_j + \gamma_{jt}^* \cos \lambda_j + \omega_{jt}^*, \quad \omega_{jt}, \omega_{jt}^* \sim \mathcal{NID}(0, \sigma_\omega^2)$$

$$\lambda_j = 2\pi j/s$$

with

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

$\mu_t$  the trend component, with random disturbance  $\eta_{t-1}$ , slope  $\nu_{t-1}$  and slope disturbance  $\xi_{t-1}$ ,

$\gamma_{jt}$  the trigonometric seasonal components with periodicity  $s = 12$  and disturbances  $\omega_{jt}$  and  $\omega_{jt}^*$ ,

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

$\beta_k$  the time-invariant coefficient of the  $k$ -th weather regressor and

$\varepsilon_t$  the additive observation level noise term

All disturbances are modelled as independent stationary Gaussian processes. The model is non-linear since the observed counts are modelled as the exponent of a linear decomposition. More specifically, the counts at time  $t$  are assumed to have a log-normal density.

### A3.2 ESTIMATION

Models are fit to the data in two steps. These steps rely on Kalman filtering, mode estimation and importance sampling, as implemented in the R-package KFAS (HELSKE, 2016).

#### A3.2.1 LOG-COUNTS

In the first step, the observation equation is rewritten as

$$\log(y_t) = \mu_t + \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt} + \sum_{k=1}^r \beta_k x_{kt} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{NJD}(0, \sigma_\varepsilon^2) \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt}$$

and the model is fit using the regular Kalman filter as described below. Importantly, in this step, any zero-counts are replaced with missing values.

The model is fit through the Kalman filter with exact diffuse initialisation of state vector  $\alpha_1$  and its associated variance-covariance matrix  $P_1$  (KOOPMAN and DURBIN, 2003).

$$\alpha_1 = (\mu_1 \nu_1 \gamma_{1,1} \gamma_{1,1}^* \gamma_{2,1} \gamma_{2,1}^* \dots \gamma_{6,1} \beta_1 \dots \beta_r)', \quad \beta_{k,1} = \beta_{k,2} = \dots = \beta_{k,T}$$

$$P_1 = \begin{pmatrix} \sigma_{\mu_1}^2 & \dots & \sigma_{\mu_1 \beta_r} \\ \vdots & \ddots & \vdots \\ \sigma_{\mu_1 \beta_r} & \dots & \sigma_{\beta_r}^2 \end{pmatrix}$$

The Kalman filter provides the equations to calculate

$$\begin{aligned} a_{t|t} &= E(\alpha_t | y_1, \dots, y_t), \\ a_{t+1} &= E(\alpha_{t+1} | y_1, \dots, y_t), \\ P_{t|t} &= \text{Var}(\alpha_t | y_1, \dots, y_t), \\ P_{t+1} &= \text{Var}(\alpha_{t+1} | y_1, \dots, y_t) \end{aligned}$$

recursively from

$$\begin{aligned} a_t &= E(\alpha_t | y_1, \dots, y_{t-1}), \\ P_t &= \text{Var}(\alpha_t | y_1, \dots, y_{t-1}) \end{aligned}$$

given the values of the disturbance parameters  $\sigma_\varepsilon^2$ ,  $\sigma_\eta^2$ ,  $\sigma_\xi^2$  and  $\sigma_\omega^2$ . The results of the Kalman filter are used in a backward recursion (the so-called state smoothing recursion) to obtain the expected value of  $\alpha_t$ , given the entire series, i.e.,  $\hat{\alpha}_t = E(\alpha_t | y_1, \dots, y_n)$  and the corresponding variance  $V_t = \text{Var}(\alpha_t | y_1, \dots, y_n)$ . (see KOOPMAN and DURBIN, 2012, for details).

The model is optimized with respect to the unknown disturbance parameters  $\sigma_\varepsilon^2$ ,  $\sigma_\eta^2$ ,  $\sigma_\xi^2$  and  $\sigma_\omega^2$ . This happens on the basis of the (diffuse) log-likelihood that is maximized through the Broyden-Fletcher-Goldfarb-Shanno optimization algorithm (BFGS e.g., FLETCHER, 1987).

To impose the constraints  $\sigma_\varepsilon^2 \geq 0$ ,  $\sigma_\eta^2 \geq 0$ ,  $\sigma_\xi^2 \geq 0$ ,  $\sigma_\omega^2 \geq 0$  and  $\sigma_\varepsilon^2 + \sigma_\eta^2 + \sigma_\xi^2 + \sigma_\omega^2 > 0$ , the transformation  $\boldsymbol{\psi} = \frac{1}{2} \log \boldsymbol{\sigma}^2$  is applied during optimization, where  $-\infty < \boldsymbol{\psi} < \infty$ .

For each model, optimization is done ten times with a different set of starting values for the disturbance parameters. The fitted model with the highest likelihood is chosen. The starting values in each run are obtained by drawing an independent random sample of vector  $\boldsymbol{\psi}$  from the uniform distribution  $[-2, -3]$ . After setting  $\boldsymbol{\sigma}^2 = \exp(2\boldsymbol{\psi})$ , the Kalman filter is applied to the model and parameter vector  $\boldsymbol{\sigma}^2$  is updated by multiplying it with the

sample variance of the innovations. Before running the BFGS algorithm the estimates are updated through 10 runs of the Expectation Maximisation (EM) algorithm (see KOOPMAN 1993).

### A3.2.2 COUNTS

To incorporate zero-counts, there is a second stage of model fitting where untransformed crash counts are modelled explicitly under the log-normal assumption in a generalized state space modelling approach. Models are fit through the method of mode estimation, which allows to find an approximating linear Gaussian state space model, and importance sampling, which allows to correct for approximation errors. The reader is referred to KOOPMAN and DURBIN (2012) for the details of these methods. Starting values for the disturbance parameters are taken directly from the best fitting model of the first stage. The number of simulation runs in the importance sampling procedure is set to 1000.

### A3.3 CALIBRATION

As noted in Section A2.3, the aim of *prgnss* is to predict future crash counts, using only those weather variables that have a proven statistically significant relationship with crash counts. To this end, models are calibrated with respect to the vector of regressors  $x$ . This happens in three steps. Initially, the model is fit including all weather PCs (i.e., those that explain at least 3% of the variance in the aggregated weather data) and their first order interactions. During the second step, the model is refit excluding all interactions with a  $p$ -value  $\geq .01$  in the full model. In the final step, the model is refit excluding all simple effects (of PCs without significant first order interactions) with a  $p$ -value  $\geq .10$ . The final model thus contains PCs that yield (a) significant ( $\alpha = .01$ ) first order interaction(s) or a significant ( $\alpha = .10$ ) simple effect. The  $p$ -values that are used in this calibration are taken from the log-counts model described in Section A3.2.1.

### A3.4 PREDICTION

For each crash data series, the weather model is fit to all available *final* data that are provided in the Excel file *prgnss\data\master.xlsx*. However, the main objective of *prgnss* is to provide predictions in the following situation.

2013	provisional	209	181	228	240	277	302	370	314	345	#N/A	#N/A	#N/A
2013	final	208	181	226	240	279	305	374	323	338	310	257	298
2014	provisional	218	186	245	296	284	339	344	278	337	#N/A	#N/A	#N/A
2014	final	219	186	244	297	286	339	341	283	327	292	262	301
2015	provisional	224	202	234	287	315	338	344	326	321	#N/A	#N/A	#N/A
2015	final	223	202	230	284	312	343	348	335	316	286	294	286
2016	provisional	242	191	190	246	320	259	331	327	322	#N/A	#N/A	#N/A
2016	final	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A

In this example, the value 286 for December 2015 is the last available count in the final data that is used for building the weather model. The objective is to predict all subsequent missing/unobserved values, i.e., Jan-Dec 2016. This is achieved by two routines. The first routine predicts missing final data on the basis of available provisional data (Jan-Sep 2016). The second routine predicts missing/unobserved final data in the period where provisional data are also unavailable (Oct-Dec 2016). This is done on the basis of the weather model (1) the observed/imputed final data (...-Sep 2016) and (2) the observed/imputed weather data (...-Dec 2016) as input.

#### A3.4.1 FINAL-PROVISIONAL MODEL PREDICTIONS

This routine provides the predictions for those months where final data are missing, but provisional data are available. In the example above, this concerns the period Jan-Sep 2016. Predictions are based on a time series model of the *difference* between available final and provisional counts. The model is fitted independently to each data series and has the following generic form:

$$d_t = \mu_t + \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{NID}(0, \sigma_\varepsilon^2) \sum_{j=1}^{\lfloor s/2 \rfloor} \gamma_{jt}$$

$$\mu_{t+1} = \mu_t + \eta_t, \quad \eta_t \sim \mathcal{NID}(0, \sigma_\eta^2)$$

$$\gamma_{j,t+1} = \gamma_{jt} \cos \lambda_j + \gamma_{jt}^* \sin \lambda_j + \omega_{jt},$$

$$\gamma_{j,t+1}^* = -\gamma_{jt} \sin \lambda_j + \gamma_{jt}^* \cos \lambda_j + \omega_{jt}^*, \quad \omega_{jt}, \omega_{jt}^* \sim \mathcal{NID}(0, \sigma_\omega^2)$$

$$\lambda_j = 2\pi j/s$$

with

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

$\mu_t$  the level component, with random disturbance  $\eta_{t-1}$  and

$\gamma_{jt}$  the  $j$ -th trigonometric seasonal component with periodicity  $s$  and disturbances  $\omega_{jt}$  and  $\omega_{jt}^*$

The months for which provisional data are available are determined automatically and separately for each data series. Subsequently, the series is compressed by eliminating those months for which provisional data are never available, resulting in a new periodicity value  $s < 12$ . In the example above, provisional data is available from January to September. Hence,  $s$  is set to 9. The parameters are estimated using the same generic procedure as for the crash data described above. The only differences are that the response variable at time  $t$  is assumed to follow normal distribution and that there is no slope component to be estimated.

The fitted model is used to obtain smoothed predictions

$$\hat{\alpha}_t = E(\alpha_t | d_1, \dots, d_n), = (\hat{\mu}_t \hat{\gamma}_{1,t} \hat{\gamma}_{1,t}^* \dots \hat{\gamma}_{s/2,t})', \text{ for } t = n+1, \dots, n+s$$

with

$$\hat{\varepsilon}_t = \hat{\eta}_t = \hat{\omega}_{jt} = \hat{\omega}_{jt}^* = 0$$

In the example above,  $t = n+1$  corresponds to January 2016 and  $t = n+s$  corresponds to September 2016. After calculating  $\hat{d}_t$ , the estimated difference is added to the observed provisional data for  $t = n+1, \dots, n+s$ , which yields the prediction for the final data in this period.

#### A3.4.2 WEATHER MODEL PREDICTIONS

The second prediction routine applies to the final months of the year where both final and provisional data are unavailable. This range is determined automatically, as it varies across the different data series. The predictions for each data series are made on the basis of the corresponding fitted weather model. The input data are (a) the final data, including the final-provisional model predictions, and (b) the weather data. In the example above, this implies (a) the final data up to September 2016 and (b) the weather data up to December 2016. If the weather series is incomplete, this will include imputed weather data. As explained in Section A2.1, the user can decide to use default imputation of weather PC values by time series modelling or manual imputation of station-level raw weather values.



Formally, the fitted weather model for a given crash data series is used to obtain the smoothed predictions

$$\hat{\alpha}_t = E(\alpha_t | y_1, \dots, y_n, x_{1,t}, \dots, x_{r,t}), = (\hat{\mu}_t \hat{\nu}_t \hat{\gamma}_{1,t} \hat{\gamma}_{1,t}^* \dots \hat{\gamma}_{6,t} \hat{\beta}_1 \dots \hat{\beta}_r)', \text{ for } t = n + 1, \dots, T$$

with

$$\hat{\varepsilon}_t = \hat{\eta}_t = \hat{\xi}_t = \hat{\omega}_{jt} = \hat{\omega}_{jt}^* = 0$$

The time point  $n$  corresponds to the month for which all subsequent provisional data is missing (Sep 2016 in the above example). Time point  $T$  is the last month in the data series (Dec 2016). The actual prediction  $\hat{y}_t$  is obtained by applying the model equations of the weather model, including (imputed) weather PC values

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \dots & x_{r,1} \\ \vdots & \ddots & \vdots \\ x_{1,T} & \dots & x_{r,T} \end{pmatrix}$$

## A4 VALIDATION

For each year since 2010 and each accident data series, predictions are 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 are kept constant across all validation years and are taken from the newly fitted models. The predictions for each month of a given year are summed and compared with the sum of the eventual final data for that year. The same is done for the predictions that have been published, as provided in the penultimate sheet of the master input file `prgnss\data\master.xlsx`. In both cases, accuracy is evaluated by calculating the root mean squared error (RMSE) relative to the final year totals.

## A5 WEATHER INFLUENCE

To evaluate the influence of specific weather conditions at a given time point, the fitted weather models are used to calculate

$$\hat{y}_t^* = \exp \left( \hat{\mu}_t + \sum_{j=1}^6 \hat{\gamma}_{jt} + \hat{\beta}' \mathbf{x}_t^* + \hat{\varepsilon}_t \right)$$

Specifically,  $\hat{y}_t^*$  is obtained by applying the observation equation of the weather model, using the smoothed state vector  $\hat{\alpha}_t$  and the average weather vector  $\mathbf{x}_t^*$ . This vector provides the average weather PC values in  $\mathbf{X}$  for the corresponding month, excluding  $\mathbf{x}_t$  (i.e., the actual values at time  $t$ ). We refer to  $\hat{y}_t^*$  as the average weather prediction. The weather influence at time  $t$  is then defined as  $y_t - \hat{y}_t^*$ . Importantly, the prediction  $\hat{y}_t^*$  is defined at the month level and does not take into account the observations in previous or following months. Hence, these predictions should not be considered as a time series on their own.

## A6 REPORTING

Each run results in 3 standard outputs: (1) a graphical report (*report.html*), (2) a numerical report with the historical final data and the current predictions (*report.xlsx*) and (3) an additional numerical report including estimated weather influences for the whole series (*weather.xlsx*). These three files are stored under *prgnss\runs\* in a folder whose name corresponds to the date and time of the run (e.g., *prgnss\runs\20170203 9h11\*).

While *prgnss* is running, the progress can be monitored in a console window. This output has the following output.

```
G00 - Getötete insgesamt :
(Gaussian ..... done)(Poisson . simulating done)
(Imputing ..... done)(Predicting done)(Validating ..... done)
(Reporting done)
```

If for some reason an error occurs when running, this is indicated as follows.

```
G00 - Getötete insgesamt :
(Gaussian ..... done)(Poisson . simulating done)
(Imputing ..... failed!)
```

The program will continue to the next data series. The data series that are affected by errors will not appear in the standard outputs that are described below. However, a text file *error\_log.txt* will be generated in the same folder. This file contains the exact error message(s). If no such file is generated, this means that *prgnss* did not encounter any errors.

### A6.1 REPORT.HTML

The graphical report provides an overview for each series of the predicted counts for the running year and the observed (final) counts in previous years. Three graphs are provided with a different historical perspective; the current year, a 10 year perspective and a full perspective since 1991.

There is also a graphical representation of the validation results. The graph shows the absolute errors since 2010 with respect to the year totals (predicted – final data). In the legend, *Published* refers to the predictions that were actually published since 2010. These predictions are read from the penultimate sheet in the input file *prgnss\data\master.xlsx*. The two remaining categories refer to the predictions of *prgnss* with (1) imputed weather values for December only and (2) imputed weather values for November and December.

Finally, for each series, a table is provided with the total root mean squared error (*rmse*) of the validation output:

$$rmse = \sqrt{\frac{\sum(predicted - final)^2}{\#years}}$$

This is again calculated for the published predictions, *prgnss* predictions with imputed weather values for December (Dec) and the *prgnss* predictions with imputed weather values for November and December (Nov-Dec).

## A6.2 REPORT.XLS

In this Excel file, there is sheet for each data series. Each sheet contains a historical overview of the final data. In this overview, there are two rows per year, one with single month values and one with cumulative values per month (*kum*):

...

<b>2011</b>	<b>246</b>	<b>196</b>	<b>305</b>	<b>346</b>	<b>374</b>	<b>352</b>	<b>355</b>	<b>381</b>	<b>392</b>	<b>369</b>	<b>321</b>	<b>372</b>
kum	246	442	747	1093	1467	1819	2174	2555	2947	3316	3637	4009
<b>2010</b>	<b>193</b>	<b>184</b>	<b>260</b>	<b>363</b>	<b>312</b>	<b>356</b>	<b>431</b>	<b>300</b>	<b>375</b>	<b>339</b>	<b>354</b>	<b>181</b>
kum	193	377	637	1000	1312	1668	2099	2399	2774	3113	3467	3648
<b>2009</b>	<b>279</b>	<b>249</b>	<b>284</b>	<b>380</b>	<b>399</b>	<b>375</b>	<b>360</b>	<b>386</b>	<b>392</b>	<b>383</b>	<b>362</b>	<b>303</b>
kum	279	528	812	1192	1591	1966	2326	2712	3104	3487	3849	4152

...

The top rows of each sheet, give an overview of the predictions for the current year. The two first rows, in blue, show the provisional data (single month and cumulative values). The next two rows, in green, show the *prgnss* predictions, i.e., (1) the predictions of the final-provisional model for those months where provisional data are available and (2) the predictions of the weather model, including observed and imputed weather values. Whether this includes imputed weather data will depend on the input that is provided in *prgnss\data\weather\_raw.csv*. Finally, pink values give the average weather prediction for each month (i.e.,  $\hat{y}_t^*$ ; see above). It is important to keep in mind that these predictions are defined at the month level and do not take into account the observations in previous or following months. Hence, they should not be considered as a time series and therefore cumulative values are not provided.

G00: Getötete insgesamt												
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
<b>2016</b>	242	191	190	246	320	259	331	327	322			
kum	242	433	623	869	1189	1448	1779	2106	2428			
	242	189	188	245	320	259	333	331	321	261	247	265
kum	242	431	619	864	1184	1443	1776	2107	2428	2689	2936	3201
	264	194	208	268	335	292	357	349	306	302	265	265
												Average weather Prgnss

...

## A6.3 WEATHER.XLS

This Excel file also contains one sheet per data series. It simply provides all monthly final data in one column, the *prgnss* predictions in the next, followed by the average weather predictions for all time points.

## Schriftenreihe

### Berichte der Bundesanstalt für Straßenwesen

#### Unterreihe „Mensch und Sicherheit“

## 2012

M 220: Mobilitätsstudie Fahranfänger – Entwicklung der Fahrleistung und Autobenutzung am Anfang der Fahrkarriere  
Funk, Schneider, Zimmermann, Grüninger € 30,00

M 221: Maßnahmen zur Erhöhung der Verkehrssicherheit von Kleintransportern  
Roth € 15,00

M 222: Neue Aufgabenformate in der Theoretischen Fahrerlaubnisprüfung  
Malone, Biermann, Brünken, Buch € 15,00

M 223: Evaluation der bundesweiten Verkehrssicherheitskampagne „Runter vom Gas!“  
Klimmt, Maurer € 15,00

M 224: Entwicklung der Verkehrssicherheit und ihrer Rahmenbedingungen bis 2015/2020  
Maier, Ahrens, Aurich, Bartz, Schiller, Winkler, Wittwer € 17,00

M 225: Ablenkung durch fahrfremde Tätigkeiten – Machbarkeitsstudie  
Huemer, Vollrath € 17,50

M 226: Rehabilitationsverlauf verkehrsauffälliger Kraftfahrer  
Glitsch, Bornewasser, Dünkel € 14,00

M 227: Entwicklung eines methodischen Rahmenkonzeptes für Verhaltensbeobachtung im fließenden Verkehr  
Hautzinger, Pfeiffer, Schmidt € 16,00

M 228: Profile von Senioren mit Autounfällen (PROSA)  
Pottgießer, Kleinemas, Dohmes, Spiegel, Schädlich, Rudinger € 17,50

M 229: Einflussfaktoren auf das Fahrverhalten und das Unfallrisiko junger Fahrerinnen und Fahrer  
Holte € 25,50

M 230: Entwicklung, Verbreitung und Anwendung von Schulwegplänen  
Gerlach, Leven, Leven, Neumann, Jansen € 21,00

M 231: Verkehrssicherheitsrelevante Leistungspotenziale, Defizite und Kompensationsmöglichkeiten älterer Kraftfahrer  
Poschadel, Falkenstein, Rinkenauer, Mendzheritskiy, Fimm, Worringer, Engin, Kleinemas, Rudinger € 19,00

M 232: Kinderunfallatlas – Regionale Verteilung von Kinderverkehrsunfällen in Deutschland  
Neumann-Opitz, Bartz, Leipzig € 18,00

## 2013

M 233: 8. ADAC/BAST-Symposium 2012 – Sicher fahren in Europa  
CD-ROM / kostenpflichtiger Download € 18,00

M 234: Fahranfängervorbereitung im internationalen Vergleich  
Genschow, Sturzbecher, Willmes-Lenz € 23,00

M 235: Ein Verfahren zur Messung der Fahrsicherheit im Realverkehr entwickelt am Begleiteten Fahren  
Glaser, Waschulewski, Glaser, Schmid € 15,00

M 236: Unfallbeteiligung von Wohnmobilen 2000 bis 2010  
Pöppel-Decker, Langner  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.

M 237: Schwer erreichbare Zielgruppen – Handlungsansätze für eine neue Verkehrssicherheitsarbeit in Deutschland  
Funk, Faßmann € 18,00

M 238: Verkehrserziehung in Kindergärten und Grundschulen  
Funk, Hecht, Nebel, Stumpf € 24,50

M 239: Das Fahrerlaubnisprüfungssystem und seine Entwicklungspotenziale – Innovationsbericht 2009/2010 € 16,00

M 240: Alternative Antriebstechnologien – Marktdurchdringung und Konsequenzen – Berichtsjahr 2011 – Abschlussbericht  
Küter, Holdik, Pöppel-Decker, Ulitzsch  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.

M 241: Intervention für punkteauffällige Fahrer – Konzeptgrundlagen des Fahrreignungsseminars  
Glitsch, Bornewasser, Sturzbecher, Bredow, Kaltenbaek, Büttner € 25,50

M 242: Zahlungsbereitschaft für Verkehrssicherheit – Vorstudie  
Bahamonde-Birke, Link, Kunert € 14,00

## 2014

M 243: Optimierung der Praktischen Fahrerlaubnisprüfung  
Sturzbecher, Mörl, Kaltenbaek € 25,50

M 244: Innovative Konzepte zur Begleitung von Fahranfängern durch E-Kommunikation  
Funk, Lang, Held, Hallmeier € 18,50

M 245: Psychische Folgen von Verkehrsunfällen  
Auerbach € 20,00

M 246: Prozessevaluation der Kampagnenfortsetzung 2011-2012 „Runter vom Gas!“  
Klimmt, Maurer, Baumann € 14,50

#### AKTUALISIERTE NEUAUFLAGE VON:

M 115: Begutachtungsleitlinien zur Kraftfahreignung – gültig ab 1. Mai 2014  
Gräcmann, Albrecht € 17,50

M 247: Psychologische Aspekte des Unfallrisikos für Motorradfahrerinnen und -fahrer  
von Below, Holte € 19,50

M 248: Erkenntnisstand zu Verkehrssicherheitsmaßnahmen für ältere Verkehrsteilnehmer  
Falkenstein, Joiko, Poschadel € 15,00

M 249: Wirkungsvolle Risikokommunikation für junge Fahrerinnen und Fahrer  
Holte, Klimmt, Baumann, Geber € 20,00

M 250: Ausdehnung der Kostentragungspflicht des § 25a StVG auf den fließenden Verkehr  
Müller € 15,50

M 251: Alkohol-Interlocks für alkoholauffällige Kraftfahrer  
Hauser, Merz, Pauls, Schnabel, Aydeniz, Blume, Bogus, Nitzsche, Stengl-Herrmann, Klipp, Buchstaller, DeVol, Laub, Müller, Veltgens, Ziegler € 15,50

M 252 Psychologische Aspekte des Einsatzes von Lang-Lkw  
Glaser, Glaser, Schmid, Waschulewski  
Dieser Bericht liegt nur in digitaler Form vor, ist interaktiv und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.

## 2015

- M 253: Simulatorstudien zur Ablenkungswirkung fahrfremder Tätigkeiten  
Schömig, Schoch, Neukum, Schumacher, Wandtner € 18,50
- M 254: Kompensationsstrategien von älteren Verkehrsteilnehmern nach einer VZR-Auffälligkeit  
Karthaus, Willemssen, Joiko, Falkenstein € 17,00
- M 255: Demenz und Verkehrssicherheit  
Fimm, Blankenheim, Poschadel € 17,00
- M 256: Verkehrsbezogene Eckdaten und verkehrssicherheitsrelevante Gesundheitsdaten älterer Verkehrsteilnehmer  
Rudinger, Haverkamp, Mehlig, Falkenstein, Hahn, Willemssen € 20,00
- M 257: Projektgruppe MPU-Reform  
Albrecht, Evers, Klipp, Schulze € 14,00
- M 258: Marktdurchdringung von Fahrzeugsicherheitssystemen  
Follmer, Geis, Gruschwitz, Hölscher, Raudszus, Zlocki € 14,00
- M 259: Alkoholkonsum und Verkehrsunfallgefahren bei Jugendlichen  
Hoppe, Tekaat € 16,50
- M 260: Leistungen des Rettungsdienstes 2012/13  
Schmiedel, Behrendt € 16,50
- M 261: Stand der Radfahrausbildung an Schulen und motorische Voraussetzungen bei Kindern  
Günther, Kraft € 18,50
- M 262: Qualität in Fahreignungsberatung und fahreignungsfördernden Maßnahmen  
Klipp, Bischof, Born, DeVol, Dreyer, Ehlert, Hofstätter, Kalwitzki, Schattschneider, Veltgens € 13,50
- M 263: Nachweis alkoholbedingter Leistungsveränderungen mit einer Fahrverhaltensprobe im Fahrsimulator der BAST  
Schumacher  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.

## 2016

- M 264: Verkehrssicherheit von Radfahrern – Analyse sicherheitsrelevanter Motive, Einstellungen und Verhaltensweisen von Below € 17,50
- M 265: Legalbewährung verkehrsauffälliger Kraftfahrer nach Neuerteilung der Fahrerlaubnis  
Kühne, Hundertmark € 15,00
- M 266: Die Wirkung von Verkehrssicherheitsbotschaften im Fahrsimulator – eine Machbarkeitsstudie  
Wandtner  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.
- M 267: Wahrnehmungspsychologische Analyse der Radfahraufgabe  
Platho, Paulenz, Kolrep € 16,50
- M 268: Revision zur optimierten Praktischen Fahrerlaubnisprüfung  
Sturzbecher, Luniak, Mörl € 20,50
- M 269: Ansätze zur Optimierung der Fahrschulausbildung in Deutschland  
Sturzbecher, Luniak, Mörl € 21,50
- M 270: Alternative Antriebstechnologien – Marktdurchdringung und Konsequenzen  
Schleh, Bierbach, Piasecki, Pöppel-Decker, Ulitzsch  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.

## 2017

- M 271: Evaluation der Kampagnenfortsetzung 2013/2014 „Runter vom Gas!“  
Klimmt, Geber, Maurer, Oschatz, Süflow € 14,50
- M 272: Marktdurchdringung von Fahrzeugsicherheitssystemen 2015  
Gruschwitz, Hölscher, Raudszus, Zlocki € 15,00
- M 273: Verkehrswahrnehmung und Gefahrenvermeidung – Grundlagen und Umsetzungsmöglichkeiten in der Fahranfängervorbereitung  
TÜV | DEKRA arge tp 21 € 22,00
- M 273b: Traffic perception and hazard avoidance – Foundations and possibilities for implementation in novice driver preparation  
Bredow, Brünken, Dressler, Friedel, Genschow, Kaufmann, Malone, Mörl, Rüdell, Schubert, Sturzbecher, Teichert, Wagner, Weiße  
Dieser Bericht ist die englische Fassung von M 273 und liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.
- M 274: Fahrschulüberwachung in Deutschland – Gutachten im Auftrag der Bundesanstalt für Straßenwesen  
Sturzbecher, Bredow  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.
- M 275: Reform der Fahrlehrerausbildung  
Teil 1: Weiterentwicklung der Fahrlehrerausbildung in Deutschland  
Teil 2: Kompetenzorientierte Neugestaltung der Qualifizierung von Inhabern/verantwortlichen Leitern von Ausbildungsfahrschulen und Ausbildungsfahrlehrern  
Brünken, Leutner, Sturzbecher  
Dieser Bericht liegt nur in digitaler Form vor und kann unter <http://bast.opus.hbz-nrw.de/> heruntergeladen werden.
- M 276: Zeitreihenmodelle mit meteorologischen Variablen zur Prognose von Unfallzahlen  
Martensen, Diependaele € 14,50

---

Fordern Sie auch unser kostenloses Gesamtverzeichnis aller lieferbaren Titel an! Dieses sowie alle Titel der Schriftenreihe können Sie unter der folgenden Adresse bestellen:

Fachverlag NW in der Carl Schünemann Verlag GmbH  
Zweite Schlachtpforte 7 · 28195 Bremen  
Tel. +(0)421/3 69 03-53 · Fax +(0)421/3 69 03-63

Alternativ können Sie alle lieferbaren Titel auch auf unserer Website finden und bestellen.

[www.schuenemann-verlag.de](http://www.schuenemann-verlag.de)