Accident Analysis and Prevention 130 (2019) 136–150

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Methodological considerations with data uncertainty in road safety analysis

Matthias Schlögl*, Rainer Stütz

AIT Austrian Institute of Technology, Center for Mobility Systems, Transportation Infrastructure Technologies, Austria

ARTICLE INFO

Article history:
Received 29 July 2016
Received in revised form 18 January 2017
Accepted 2 February 2017
Available online 16 February 2017

MSC:
62-07

Keywords:
Uncertainty
Road safety
Accident analysis
Linear referencing
GIS

ABSTRACT

The analysis of potential influencing factors that affect the likelihood of road accident occurrence has been of major interest for safety researchers throughout the recent decades. Even though steady methodological progresses were made over the years, several impediments pertaining to the statistical analysis of crash data remain. While issues related to methodological approaches have been subject to constructive discussion, uncertainties inherent to the most fundamental part of any analysis have been widely neglected: data. This paper scrutinizes data from various sources that are commonly used in road safety studies with respect to their actual suitability for applications in this area. Issues related to spatial and temporal aspects of data uncertainty are pointed out and their implications for road safety analysis are discussed in detail. These general methodological considerations are exemplary illustrated with data from Austria, providing suggestions and methods how to overcome these obstacles. Considering these aspects is of major importance for expediting further advances in road safety data analysis and thus for increasing road safety.

1. Introduction

Road traffic accidents are the root of considerable harmful effects on societies all over the world. With as much as 1.25 million fatalities and approximately 50 million people who get injured every year, road crashes are globally recognized as a leading cause of death as well as a major cause of severe losses to society (OECD/ITF, 2015; WHO, 2015). In addition to untold pain and suffering, this entails immeasurable harmful economic and social effects.

Against this background, great efforts were made to reduce the number of crashes by establishing targeted policies and implementing appropriate safety-related countermeasures (WHO, 2004). Scientists and researchers all over the world made notable contributions to improving the situation by pushing methodological advances that facilitate gaining a better understanding of the underlying factors which affect the likelihood of crashes (Tarko et al., 2013). Great endeavors were made in recent years to improve the capability of predicting the probability and severity of road accidents by developing increasingly sophisticated accident analysis models (Lord and Mannering, 2010; Savolainen et al., 2011; Mannering and Bhat, 2014).

The development of advanced modeling approaches went hand in hand with an increase of methodological complexity. This has led to the evolvement of a dichotomy between the highly sophisticated methods applied by pioneering researchers and approaches used in applied science and by practitioners (Mannering and Bhat, 2014). Even though it is incontestable that scientific curiosity regarding methodological advances is one of the very foundations of science that helps to gain important new inferences, it is not advisable to lose sight of the big picture – in particular as far as general applicability of the proposed methods is concerned. While previous work in this area has mainly focused on developing new modeling approaches that offer superior fit, few researchers have addressed the issues inherent to the underlying data. Against the background of new emerging data sources (Antoniou et al., 2011; Mannering and Bhat, 2014; Wu et al., 2014a), it is thus wise to acknowledge issues related data management as well as to various aspects of data uncertainty (e.g. precision, accuracy, reliability). Usually, a lot of effort has to be put into preparing unruly raw data for further use in statistical modeling. It is estimated that data scientists spend up to 80% of their time on data wrangling (Dasu and Johnson, 2003; Wickham, 2014). In many cases, neither possible decisions taken in this process nor their implications on the further analysis are reflected and discussed comprehensively. However, the straightforward use of raw data without being distinctly aware of their

* Corresponding author.
E-mail addresses: matthias.schloegl@ait.ac.at (M. Schlögl), rainer.stuetz@ait.ac.at (R. Stütz).

https://doi.org/10.1016/j.aap.2017.02.001
0001-4575/© 2017 Elsevier Ltd. All rights reserved.

characteristics and limitations will lead to a biased estimation of statistical models and consequently to wrong inferences.

Even though the overall data availability has increased in recent years, this does not necessarily imply that data quality is keeping pace with neither the methodological progresses nor the genuine understanding of the data genesis from various sources.

In many cases, researchers are hence initially confronted with problems related to data preparation and data preprocessing. Given various restrictions inherent to road safety related data in terms of both availability and quality, a great deal of effort has to be put into preparing data for further use. While the focus was laid on the development of models that offer the best possible fit to some data, pressing concerns related to data themselves somehow faded from the spotlight. As a matter of fact, several issues related to methodological problems in accident modeling were taken up and pointed out by various publications (Lord and Mannering, 2010; Mannering and Bhat, 2014). However, previous research has tended to focus on the modeling part of the whole accident analysis procedure, while collected data are implicitly considered to be an adequate reflection of reality. Hence, this paper takes up on methodological considerations one stage earlier on the data level and seeks to point out several serious issues related to data preparation as well as data preprocessing against the background of uncertainty. The questions raised in this context are practically highlighted with examples from Austria.

This paper is structured as follows: In Section 2 available data sources that are potentially relevant for analyzing road accidents are described. Section 3 assesses the uncertainties inherent to these data sources. Issues related to data quality are presented, including examples from various Austrian data sets for demonstration purposes. Following a review of suitable spatial and temporal resolution of data for accident analysis in Section 4, issues related to data processing are illustrated and discussed in Section 5. Finally, Section 6 summarizes and concludes this paper.

2. Data sources

2.1. Accident data

Traditional road safety data used in accident analysis are usually gathered by the police. These police crash reports are forwarded to ministries of transport or national statistical offices, which collect all data, perform descriptive data analysis and publish official national road accident statistics. In addition, they provide accident data sets for scientific purposes. In Austria these accident records are collected by the national bureau of statistics (Statistics Austria).

As a matter of fact, most existing road accident studies have used data from police crash reports for modeling road accident risk. These data sets usually include a vast amount of different parameters, which may potentially serve as explanatory variables within various models for assessing both accident frequency and accident severity (OECD/ITF, 2015).

The information included in these data sets can basically be classified into three main aspects:

- information related to the accident location level: date and time of the accident, coordinates of the accident, name of municipalities in which the accident happened, road type, information on whether the accident occurred on a rural or urban road, speed limit, accident type, presumptive main accident cause, road-surface conditions, weather conditions, roadway lighting, basic roadway geometrics, kilometer of road, etc.
- information related to the involvement level: type of vehicles involved, involvement of trailers, engine power, hit-and-run, transport of hazardous materials, license plate/country code, date of first registration, possible contributing circumstances to the crash, etc.
- information related to the person level: type of involvement (driver, passenger, pedestrian), date of birth, gender, injury-severity, nationality, driver sobriety, safety belt usage, air-bag deployment, etc.

It has to be noted that the emergence of electronic data processing led to changes in how police-reported data are collected. Therefore, accident data ascertainment underwent a major change in Austria in January 2012, when a system change-over for accident data acquisition was implemented. While accidents were reported with a pencil-and-paper method via so-called *accident count sheets* until 2011, an electronic data processing system for accident data management (ADM) was introduced in 2012 (BMVIT, 2011, 2013). The resulting break in the time series entails that data are not directly comparable.

2.2. Road geometry and road conditions

Data on road geometry and road conditions are available from different sources.

Firstly, there are official transport graphs, i.e. digital maps of the transport network. These maps are intended to provide infrastructure- and traffic-related information to administrative units, road management authorities and transport companies. In Austria, the official transport graph is provided by the Graph Integration Platform (GIP), which represents the common intermodal traffic reference system for public authorities. GIP data include geographic information of traffic network (nodes and links), reference points and many parameters related to routing (functional road classes, number of lanes, trafficability, speed limit, average speed, lane directions, turn allowances, etc.). Since January 2016 an export of the GIP data is provided as open government data. The data sets are updated in a bi-monthly interval (GIP, 2016).

Secondly, information on road characteristics can be obtained from mobile measurement systems. The EU Directive 2008/96/EC on road infrastructure safety management provides for periodic road safety inspections on the trans-European road network (Council of European Union, 2008). In addition, road condition data are required for pavement management, as they serve as a basis for decision-making with respect to road maintenance and reconstruction projects. Therefore, measurement campaigns are not only carried out on the high-level road network, but also on rural roads. Road safety inspections as well as data ascertainment for pavement management are carried out by some sort of mobile measurement device like rebuilt trucks, trailers or other measuring vehicles.

In Austria, precise measurements of the road surface are obtained via periodic measurement campaigns with various mobile measurement systems that are mounted on a rebuilt truck (RoadSTAR system). These measurement systems deliver high-resolution data on road condition, road geometry and the road environment (FSV, 2009b). Measured parameters include skid resistance, transverse and longitudinal evenness, curve radii, texture and water-film thickness (FSV, 2002, 2004a,b,c). Measurements are performed in a single run under normal traffic conditions (40–120 km/h), with a standard measurement speed of 60 km/h. The road geometry can be derived from an inertial measurement unit (IMU) and differential GPS with a spatial resolution of 1 m, while the evenness is measured by different laser scanning system with resolution up to 5–20 mm. Several camera systems provide information about surface defects and the road environment (FSV, 2009a). The high-level road network is periodically monitored in five year cycle in Austria, where the latest measurement campaign started in the

2014. Rural roads are (successively) measured at irregular intervals and different levels of coverage every couple of years.

2.3. Weather

Various studies have already looked into weather effects on road accident occurrence (Fridstrøm et al., 1995; Shankar et al., 1995; Eisenberg, 2004; Brijs et al., 2008; Koetse and Rietveld, 2009; Jung et al., 2010; Bergel-Hayat et al., 2013; Theofilatos and Yannis, 2014). In particular, the influence of rainfall events on road accidents are been investigated since several decades (Sherretz and Farhar, 1978; Brodsky and Hakkert, 1988; Andrey and Yagar, 1993; Edwards, 1996). Therefore, including weather data into accident prediction models in order to eventually assess potential safety risks is desirable.

Weather data are available in different formats and from different sources. Firstly, time series of various types of weather information can be obtained from meteorological measuring stations (in-situ data). Automated weather stations operated by national meteorological services under a standardized framework provided by the World Meteorological Organization (WMO) as a part of WMO's Global Observing System (GOS) are a viable source of information. Measurements obtained from these stations are comparable in a way that they follow standardized procedures regarding general requirements, types, location and composition, frequency and timing of observations (WMO, 2008, 2010, 2011).

Currently, the meteorological measuring network in Austria comprises more than 250 semi-automatic weather stations (some of which are additionally supervised or include manned observations), which are operated by the Central Institution for Meteorology and Geodynamics (ZAMG). In addition, the measuring network also comprises stations from the Austrian hydrographical services and the Austrian aviation weather service Austro Control. Weather data from these stations are currently available at a temporal resolution of 10 min, with the exception of precipitation amount, which is measured every minute.

While meteorological data is measured with relatively high confidence at these stations, data are only available at certain points in space. In order to obtain local-scale weather information across whole areas of interest, two general procedures can be applied.

On the one hand, various spatialization methods for interpolating climate variables obtained from in situ station measurements have been developed (Daly et al., 1994, 2008; Steinacker et al., 2011b; Bica et al., 2007; Szentimrey et al., 2011; Teegavarapu et al., 2012; Frei, 2014; Gyasi-Agyei and Pegram, 2014; Hiebl and Frei, 2016; Boudevillain et al., 2016). Basically, these methods are based on statistical techniques for interpolating scattered station reports, integrating physical constraints, various additional measurements (e.g. weather radar) and additional environmental variables like for instance land cover/vegetation, topography/digital elevation model, sun exposure, water bodies, etc.

On the other hand, meteorological and climate models can be used to derive spatially comprehensive climate information. Most notably, this encompasses meteorological reanalyses and climate models. Meteorological reanalysis make use of an unchanging, consistent data assimilation scheme that combines (heterogeneous) data obtained from a huge observational network (including numerous different, ever changing data sources like e.g. weather stations, satellites, radiosondes, buoys, ship and aircraft weather reports) with dynamic meteorological models (Kalnay et al., 1996; Uppala et al., 2005; DeGaetano and Belcher, 2007; Dee et al., 2011; Luhamaa et al., 2011; Rienecker et al., 2011; Kobayashi et al., 2015) in order to obtain estimates for the meteorological variables on a regular grid. As far as climate models are concerned, general circulation models (GCMs), which incorporate mathematical methods to model the general circulation of the earth's atmosphere (and ocean) have to be mentioned in this context. Such coarse grids can be further refined by either empirical-statistical downscaling (ESD) or by dynamical downscaling methods. While ESD uses statistical relationships between large-scale climate variables (predictors) and the observed local climate (predictands) for deriving regional scale patterns from either GCMs (Matulla et al., 2002; Benestad, 2004; Matulla, 2005) or observational data (Schmidli et al., 2001), dynamical downscaling is based on nested numerical modeling, i.e. on model chains composed of regional climate models (RCM), which are driven by GCMs (Giorgi and Gutowski, 2015). In both cases, the resulting data sets consist of time series of meteorological variables, which are available as area-covering regular grids (raster stacks). Depending on the method used for obtaining these meteorological grids, the spatial and temporal resolution at which these grids are available may vary considerably. In particular, spatial and temporal resolution depend on both the region (i.e. on the scale: national – continental – global) and on the length of the time series under consideration. In order to guarantee homogeneity within the time series, some data sets comprise only selected stations which are assumed to provide consistent measurements that can be used in long-term time series without compromising consistency (Auer et al., 2007; Hiebl and Frei, 2016). Consequently, this entails a lower spatial and temporal resolution compared to e.g. nowcasting systems, which include all information currently available from stations at the expense of long-term temporal homogeneity (Haiden et al., 2011; Kann et al., 2012).

Apart from the E-OBS European gridded data sets on temperature and precipitation, which are available on a daily basis on both a 0.25 degree regular lat-lon grid as well as on a 0.22 degree rotated pole grid (Haylock et al., 2008), specific gridded data sets for the Alpine region and particularly for Austria are available. Daily temperature and precipitation grids covering Austria at 1 km-resolution on a daily basis have been derived e.g. by Hiebl and Frei (2016). These data sets mainly aim at offering long-term homogeneous meteorological data, which is usually not required for accident analysis. Thus, high-resolution data sets like INCA (Haiden et al., 2011) or VERAflex reanalyses (based on VERA (Steinacker et al., 2011b)) are available at a temporal resolution of down to 10 min and at a spatial grid of as small as 100 m.

2.4. Traffic volume

Counting vehicular traffic along roads can be carried out either automatically (via temporal or permanent devices) or manually (FHWA, 2013). Continuous automated measurements can be undertaken by using e.g. inductive loops, piezoelectric detectors, pneumatic road tubes, radar counters or videos (Ni, 2016; Xia et al., 2016). Generally speaking, permanent measurement devices are obviously better suited for providing detailed data about the traffic volume, as the road is continuously monitored. However, in most cases only the primary road network is monitored by means of continuous count stations. In other cases, short duration measurements carried out by either temporal devices or via visual counting and recording by observers can also be used to estimate the traffic volume. Even though both methods yield comparable results and despite counting errors of manual counting are small (Mao et al., 2012), manual traffic counts are considered to be cumbersome and inefficient (Findley et al., 2011). Anyway, these short-term traffic counts can be used to derive indicators for annual average daily traffic (AADT) or similar benchmarks.

Traffic volume on the Austrian highway network is constantly monitored by the traffic management and information system VMIS (Verkehrsmanagement- und Informationssystem), which comprises – among other sensors – more than 500 permanent traffic count stations. Data from approximately 280 of these stations are officially validated and published as monthly AADT estimates,

Table 1
Summary statistics of absolute deviations of reported accident locations (in meters). The last character in the road identifier denotes the direction ('R' direction with increasing kilometrage (right), 'L' direction with decreasing kilometrage (left)).

Motorway	n_{Acc} [−]	Min. [m]	Q_1 [m]	Mean [m]	Median [m]	Q_3 [m]	Max. [m]
A01L	195	0.3	37.4	646.7	179.0	485.4	50,090.3
A01R	188	1.1	26.9	443.0	113.1	401.9	10,215.2
A02L	179	1.8	49.0	393.5	204.6	554.2	4521.8
A02R	181	0.7	22.4	1419.4	111.5	398.1	139,737.2
A08L	6	26.0	61.9	250.3	124.5	393.9	702.0
A08R	25	2.8	17.3	381.4	140.7	301.4	5539.3
A10L	48	2.1	33.1	1394.3	139.8	372.6	42,875.4
A10R	40	2.7	53.0	1392.9	199.1	393.2	29,057.6
A12L	62	3.6	30.2	374.6	123.4	279.9	4896.7
A12R	52	2.5	34.1	162.7	87.5	218.3	708.3
A13L	15	10.6	27.0	143.3	41.3	170.5	652.0
A13R	7	9.0	27.2	70.3	38.2	89.5	211.6

including heavy good vehicles. However, raw data are gathered at these stations at a temporal resolution of one minute. In addition, national, regional and local roads are also monitored by numerous permanent traffic count stations, of which some 300 are officially validated by the regional government offices (BMVIT, 2016).

3. Analysis of uncertainties

3.1. Spatial uncertainties

3.1.1. Accident data

While police-reported data have been used in a wide variety of studies, which doubtlessly made important contributions to providing deeper insights into various aspects of road safety research, this data source has several drawbacks, too. Some of these downsides are documented quite well, others somehow seem to go by the board or are simply overlooked.

Previous work has mainly focused on ongoing methodological issues in the context of statistical modeling of crash frequencies and injury severities and, consequently, on correction techniques that allow obtaining unbiased inference (Lord and Mannering, 2010; Mannering and Bhat, 2014). Albeit uncertainties inherent to road safety data also pose a serious threat to the general validity of statistical models, this aspect has remained a neglected area thus far.

Only one of the commonly recognized and well documented problems mentioned in Lord and Mannering (2010) directly emerges from the quality of the data themselves, namely the issue of under-reporting in official accident statistics. That is, accidents with less severe injuries are likely to fall victim to not appearing in crash databases, as these accidents are simply not recorded by the police. For the sake of completeness, it should be mentioned at this point that usually only accidents involving personal injuries are considered in many accident analysis models, while accidents involving only damage to property are neglected. Obviously, this problem of under-reporting leads to a slightly biased sample of road accident occurrences. For details on the problem of under-reporting see e.g. Kumara and Chin (2005), Yamamoto et al. (2008), Ma (2009), Lord and Mannering (2010), Ye and Lord (2011), Patil et al. (2012), Yasmin and Eluru (2013) and Mannering and Bhat (2014).

However, there has been little discussion on the quality of police-reported data apart from the issue of under-reporting. In fact, there are several sources of uncertainty that have not yet been discussed in greater detail. Basically, it can be stated that the crux of the matter is related to various sources of inaccuracy, which entail different types of uncertainty. As a matter of fact, sizable inconsistencies regarding data ascertainment can be observed in numerous parameters. Most notable, time and location of the accident are subject to considerable uncertainties.

As far as spatial accuracy of crash site localization is concerned, two main sources of uncertainty arise. On the one hand, accidents are not events that can be located exactly due to the process of an accident occurring over a certain distance along a road segment, ranging from some "accident-inducing" point to the final end position of the vehicles. On the other hand, additional uncertainty is introduced by the way the crash scene is finally localized by the police. In Austria, accident location is usually reported in two ways since the introduction of ADM in 2012. On the one hand, police officers manually set WGS84-coordinates for the final end point of the vehicles within a web mapping service. On the other hand, they also provide a specification of the respective road kilometer based on driver location signs. This double reporting leads to spatial inconsistencies which can be nicely used for illustrating this aspect of spatial uncertainty.

Therefore, the accuracy of the reported accident location was evaluated on a sample of six motorways in Austria. The two location attributes can be linked using a linear referencing system (LRS) in order to compute the distance between the resulting locations. In turn, to build a LRS, a set of reference points are required. In order to check consistency, two variants for creating the LRS were considered: On Austrian motorways, driver location signs are placed every 250 or 500 m either along the hard shoulder at each side or on the median strip. The first set of reference points consists of manually set markers for each of these distance marker posts during measurement runs of the RoadSTAR system, which are georeferenced through a differential GPS system. Secondly, a stereo camera system allows a highly accurate 3D-calibration of the position of the mileage marker in stereo video frames. This referenced mileage markers provide the basis for the reference points in the GIP, which were used as second evaluation method.

The considered road network, which has a length of 1124 km, covers ca. 65.4% of the Austrian motorway network (cf. Fig. 2). Both motorway directions were evaluated separately, and only those accidents were taken into account, where the direction attribute has been provided. In total, 998 accidents from the year 2014 were used to compute the deviations of the two location specifications. Both evaluation methods yield similar and consistent results, which showed an overall median absolute deviation of approximately 136 m. Table 1 contains a five-number summary (sample minimum, first quartile, median, third quartile and sample maximum), plus the sample mean and the number of observations n_{Acc}. Results show that several extreme outliers strongly affect the mean absolute differences. This does not only become visible in the maximum values reported in the last column, but also with respect to large deviations between the mean and the median.

In addition, the outliers become visible in the graphical representation of these results, which additionally shows the differences in deviations between both different highways and different directions of the same highway. Note that the y-axis limit is capped

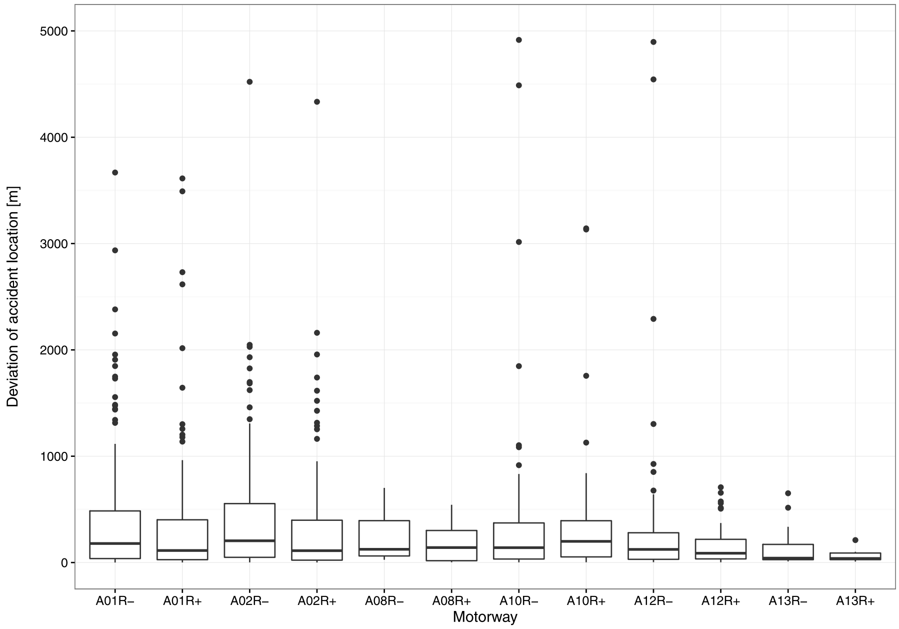

Fig. 1. Absolute deviations between accident locations based on driver location signs and accident locations based on WGS84-coordinates for selected Austrian highways.

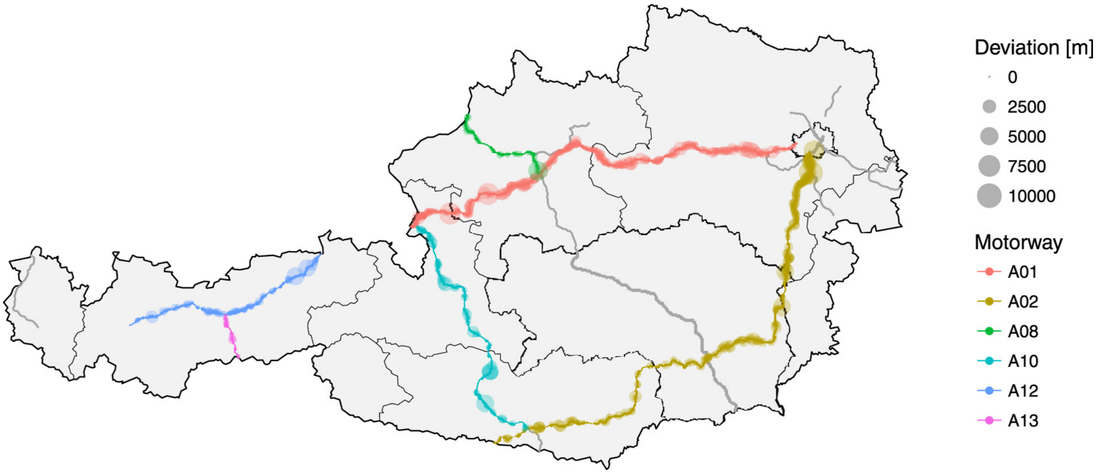

Fig. 2. Spatial representation of absolute deviations of the reported road kilometers from WGS84-coordinates. Magnitudes of deviations are illustrated as scaled circles.

to 5000 m, but there are several points above this limit (up to ca. 140 km, cf. last column in Table 1). Such high deviations are typically errors in the accident database. For instance, some accidents apparently occurred on a motorway ramp, but were erroneously encoded as motorway accidents (Fig. 1).

Despite these differences between highways and directions, no systematic spatial patterns could be derived (Fig. 2).

While results illustrate the uncertainties in the specification of the accident locations, it cannot be derived from this evaluation which of the two location specifications – WGS84-coordinates or road kilometer – is more accurate.

3.1.2. *Road geometry and road conditions*

Data about road geometry and road conditions are usually available at high spatial resolution. On top of that, measurement and mapping uncertainties with respect to spatial aspects are very small, as measurements and digital mapping are carried out with high precision. Road geometries and course of the road may of course change over time, which is considered as temporal uncertainty, though.

3.1.3. *Weather data*

Most of the aforementioned studies use observed data from weather stations in close proximity to the area of investigation. However, various constraints have to be considered when incorporating weather data into accident models, and trade-offs have to be made taking into account the properties of data derived from different data sources.

While meteorological data are measured with relatively high confidence at the standardized WMO stations, data are only available at certain points in space (WMO, 2008). Assigning road segments to weather stations in close proximity is problematic due to several reasons. Primarily, different topographic and microclimatic conditions entail a varied exposure to weather. Especially in heterogeneous, small-scale landscapes and complex terrains, in-situ observations are hence only conclusive to a limited extent. Potentially important variations in weather conditions that cannot be captured by a scattered measuring network In addition, extreme events are taking place at small scales. Therefore, some extreme events might not be captured by any station, which most likely leads to a bias towards underestimation of extreme events.

While accuracy and reliability of data derived via interpolation techniques and via downscaling of meteorological reanalysis output are subject to higher uncertainty than observed data – as they rely upon interpolation and physical models in addition to mere measurement uncertainty – these data sets open up new possibilities of analyses due to their capability to present relatively precise individual values for each point in space. However, the occasionally considerable uncertainty related to the estimates of these weather data grids has to be taken into account.

Generally speaking, uncertainties associated to these gridded data sets are assessed and reported thoroughly in meteorological and climatological literature, most commonly by means of established model verification techniques like cross-validation and statistical model performance indicators (Haiden et al., 2011; Brands et al., 2012; Borsche et al., 2015; Rose and Apt, 2016; Hiebl and Frei, 2016). In addition, many validation efforts as well as method and model comparisons have been undertaken (Hofstra et al., 2008; Mooney et al., 2011; Jakobson et al., 2012; Abatzoglou, 2013; Fu et al., 2016; Hu et al., 2016; Song et al., 2016).

Regarding spatial resolution, well documented methods are available for obtaining grids at a 1 km resolution (Steinacker et al., 2011a; Haiden et al., 2011; Hiebl and Frei, 2016). While data with lower resolutions might be potentially useful for countries featuring vast homogeneous landscapes (e.g. Great Plains), high spatial resolution is required for adequately accounting for small-scale variations in complex landscapes.

Uncertainties in these data depend on the parameter under consideration. While temperature can be accurately modeled with an average accuracy of 1–2 °C, the mean relative analysis error for precipitation amounts is 50–100% (Haiden et al., 2011; Steinacker et al., 2011a). Downscaling methods are only capable of describing a part of the daily precipitation variability (Wilby and Wigley, 1997), and precipitation is particularly difficult to model along mountain ranges like the central Alps (Gao et al., 2014). As far as wind speed is concerned, the relative mean absolute error has been estimated to be roughly of the order of 50% (Haiden et al., 2011). In addition, seasonal patterns of wind estimation uncertainty have been found (Haiden et al., 2011; Rose and Apt, 2016).

Despite a relatively high spatial resolution of 1 km, subgrid-scale variations within these high-resolution grids may entail representativeness problems (Haiden et al., 2011). Uncertainties for high-resolution VERAflex reanalyses – which are capable of addressing these subgrid-scale variations by providing a spatial resolution of up to 100 m – are not known to the authors.

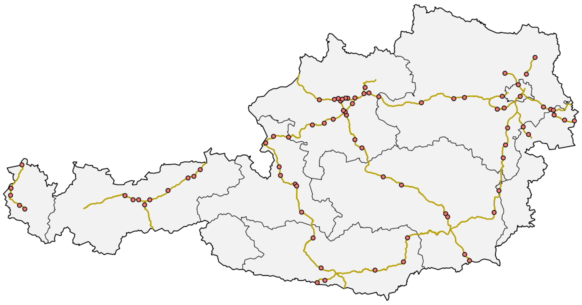

Fig. 3. Location of permanent traffic counters on the Austrian highway network. Only those counters where measurements are available for at least 11 months are included.

3.1.4. Traffic volume

Automated traffic counters are located along road networks in irregular intervals. Even though coverage may be good, measurements are usually not available for every section of each highway, let alone roads of lower hierarchy. Since measurements are only carried out at certain points in space, efforts were undertaken to impute missing values or predict them at unmeasured locations (Zhao and Park, 2004; Selby and Kockelman, 2013; Lowry, 2014). However, uncertainties related to these statistical methods (like e.g. Kriging or geographically weighted regression) are further aggravated by uncertainties inherent to traffic count measurements. Widely scattered stations hamper accurate, network-covering estimates of traffic volume, especially off the highway. The spatial distribution of officially validated traffic counters along the Austrian highway network illustrates that traffic volume actually has to be estimated at large portions of the highway network without drawing on permanent traffic counters (Fig. 3).

3.2. Temporal uncertainties

3.2.1. Accident data

The time of the accidents is estimated by police forces on site. This results in rounding effects, which produce a distinct pattern of accident occurrence over the course of 1 h (Fig. 4). The resulting plot clearly deviates from a discrete uniform distribution, showing that rounding effects lead to an accumulation at five minute bins. The two biggest peaks emerge at full and half hours and 43% of observed accidents are recorded at quarter hours.

These uncertainties are additionally underlined by discrepancies that become visible when comparing police-reported accident data with alerting times of the air emergency service.

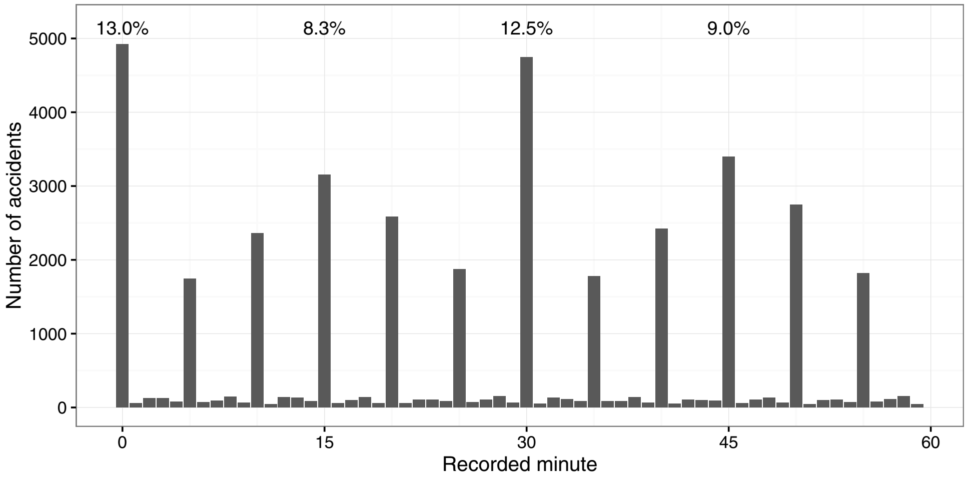

Fig. 4. Distribution of road accident occurrence by minutes of the hour, based on 37,957 accidents with personal injuries in Austria in 2014.

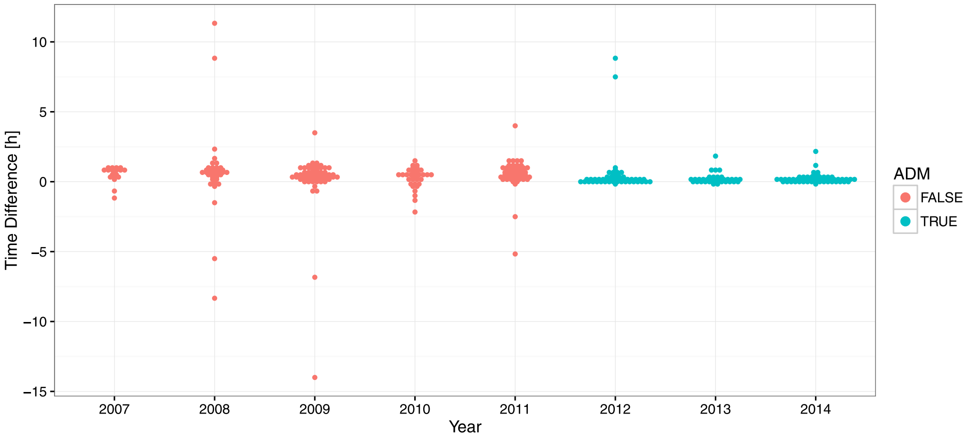

Fig. 5. Time difference between alerting time of the rescue helicopter and recorded time in the accident database binned to 10 min intervals, based 346 run-off road accidents from 2007 to 2014.

The main helicopter emergency medical service in Austria is operated by the Christophorus Flight Rescue through the Austrian Automobile, Motorcycle and Touring Club (ÖAMTC). The fleet consists 30 helicopters, which operate from 16 bases across Austria. At each base at least one helicopter is stationed and typically 3–7 missions are run per helicopter per day.

In this study, a sample of 346 run-off-road accidents taken care of by the air emergency in the period 2007–2014 were considered. A linkage with the official accident database allows a comparison between the alert times of the rescue helicopters and the timestamps in the accident records.

Albeit some deviations should be expected due to the aforementioned rounding effects, certain patterns emerge (Fig. 5). Set aside these rounding effects, it is reasonable to assume that the alert time of the air emergency service is slightly later than the actual accident time. Positive outliers can be easily explained by helicopters demanded by rescue sources in the field already present at the accident locations. It can be observed that major negative deviations do not occur anymore following the introduction of ADM.

3.2.2. Road geometry and road conditions

While data about road geometry and road condition can be gathered at high precision and with high accuracy as far as spatial aspects are concerned, these data are just mere snapshots in time. Presuming these data are constant over the entire time period of consideration (i.e. up to five years) by using a fixed version of the road graph and data from a single measurement campaign will introduce a bias in the resulting model. To some extent, certain parameters like e.g. road gradients can reasonably be assumed to remain constant. However, many other parameters, most notably skid resistance, are subject to considerable variations over time. Furthermore, changes in the course of the road as well as reconstruction works may entail distinct breaks in the properties of road features, which are difficult to be captured. Since taking information about construction works into account is usually cumbersome, short term changes caused by maintenance activities may also not be captured. The extent of uncertainty caused by temporal changes and variations in road properties is difficult to estimate, as this strongly depends on the parameters used and the actual situational conditions.

3.2.3. Weather

Temporal uncertainties of weather data are mainly related to the temporal resolution of the data sets. The smaller the temporal scale, the higher the associated uncertainties. While data sets comprising daily aggregates of weather parameters (e.g. daily precipitation totals; temperature minimum, maximum and mean; mean wind speed and maximum wind gusts; etc.) are more robust than data with a higher temporal resolution, important within-period variations might be averaged out in data sets on a daily basis. Hence, a trade-off has to be found between robustness and precision. As the resulting uncertainties are usually not evaluated separately with respect to temporal and spatial resolution, aforementioned uncertainties of weather data, which referred to a 1 km grid, are also valid for a temporal resolution of 15 min (Haiden et al., 2011).

Another aspect related to temporal uncertainty is also somehow connected to the temporal resolution of the data. In order to guarantee long-term consistency of meteorological and climatological time series, only measurements on a daily basis can be used. In Austria, the majority of automated weather stations has been installed around the mid-1990s. Measurements on an hourly basis are only available at certain stations situated at selected locations.

3.2.4. Traffic volume

While automated traffic counts can be assessed at high temporal resolutions, usually estimates of AADT values are used in accident prediction models. Since accidents occur at certain points in time, average daily values are merely capable of reflecting the actual traffic volume at this time spot to some extent. Naturally, it can be reasonably argued that including interactions between AADT, date and time is entirely sufficient for appropriately describing variations of the traffic volume. However, a detailed analysis of raw measurement data from all automated traffic count stations along the Austrian highway network gives rise to doubts in this regard.

Apart from the permanent counting stations which are used as a basis for calculating official values for AADT on the Austrian highway network, traffic counts can be obtained from the traffic management and information system operated by the Austrian motorway operator ASFiNAG. Among other duties (like for instance traffic control), this system carries out traffic counts with a temporal resolution of one minute at more than 500 locations on the highway network. Due to this high resolution, these sensitive measuring devices are prone to sensor malfunctions. In order to guarantee a necessary degree of robustness and reliability, measurement data were aggregated to hourly counts. All resulting aggregates which did not feature at least 75% data coverage were discarded. Consequently, resulting valid values contain at least 45 samples per hour.

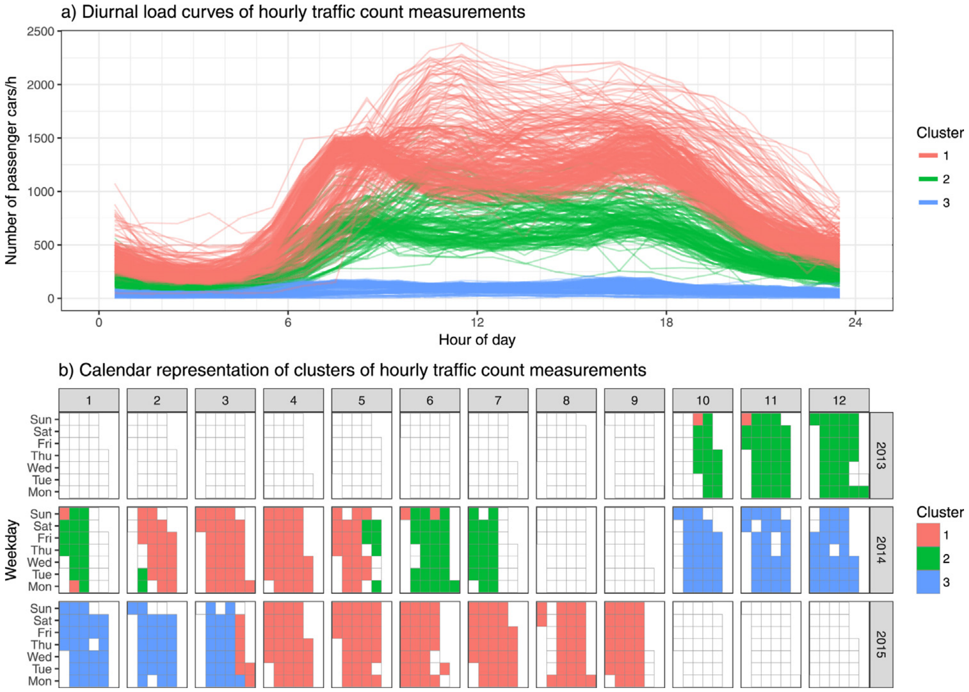

Fig. 6. Clustered load curves of hourly traffic count data. Data include all available measurements fulfilling data quality requirements.

These hourly data were used to derive diurnal traffic load curves. Therefore, the 1h-aggregated data were subsequently clustered in a two step procedure. First, the estimated number of clusters was determined with the aid of partitioning around medoids (PAM) using the optimum average silhouette width criterion. The final cluster result was obtained by a (longitudinal) *k*-medians clustering algorithm. This procedure was replicated for different week durations (Mon–Thu, Mon–Fri and Mon–Sun). It has to be noted that the classification of these days was carried out according to the Austrian guideline for traffic counts (FSV, 2015). In order to establish a comparable characterization of weekdays, (movable) feasts are taken into account. For instance, if a feast occurs on a Monday, this Monday is characterized as a Sunday and the following Tuesday is classified as a Monday.

While no consistent result representing the whole road network could be derived, some issues were observed. Traffic load curves show inconsistencies that are ascribed to sensor malfunctions. Despite the elimination of apparently invalid measurement data, clusters consisting of implausible measurements become visible (Fig. 6a). Since the blue colored lines form a cluster consisting of too low values which exhibit no diurnal variations, it has to be doubted that the sensor device was in a fully operational state. The two remaining clusters exhibit plausible load curves and are separated according to their traffic volumes. A calender representation of the clustering result (Fig. 6b) reveals strong seasonal patterns. The presented data set ranges from October 2013 to September 2015. Due to a database fault, no data are available for the period between August and September 2014. Single white cells within the period under consideration represent values that were discarded due to data quality reasons. While the blue cluster that occurs between October 2014 and March 2015 is likely due to sensor malfunction, explanations for the occurrence of the patterns related to the other two clusters are not evident.

Even though several authors have investigated the effects of hourly traffic flow on road accidents (Zhou and Sisiopiku, 1997; Martin, 2002), studies comprehensively investigating a possible benefit of using hourly traffic count data as opposed to AADT values have not been conducted yet.

3.3. *Structural uncertainties*

3.3.1. *Uncertainties caused by changes in data structure*

Furthermore, longer time series of accident data have to be analyzed against the background of changes in the database format or in the method of data collection. These changes may relate to alterations of variables, changes in how data is collected, modifications of data elements recorded or accidents reported.

For instance, the comprehensive change in the data collection method of Austrian accident data leads to a distinct break in the time series, which entails that data collected before and after the introduction of the ADM-system are not directly comparable. Effects caused by the definition of mandatory fields, the increase in the number of variables or an increase in the number of reported crashes due to electronic transmission led to inconsistencies with previously collected accident data. These inconsistencies were already briefly pointed out before when validating the accident times against the alert times of the air emergency service (Fig. 5). Another illustrative example can be provided with respect to sun glare accidents. Few studies have so far investigated the effects of sun glare on road accident occurrence. Some studies conclude that sun glare is a potentially relevant variable for accident prediction models – in particular at intersections – which is often omitted (Mitra and Washington, 2012; Mitra, 2014).

In the aforementioned examples, spatial and temporal uncertainties have been derived using external data sets. Another approach is to derive certain features, which allow a validation of the consistency of specific attributes in the accident database.

In the following example, sun glare accidents of the period 2004–2014 were considered. Based on the timestamp of each accident, the solar elevation angles[1] were computed using an implementation of algorithms provided by the National Oceanic & Atmospheric Administration (NOAA) (Bivand and Lewin-Koh, 2016). For visualization purposes, solar elevation is not only

[1] The solar elevation angle is the angular height of the geometric center of the sun relative to the horizon.

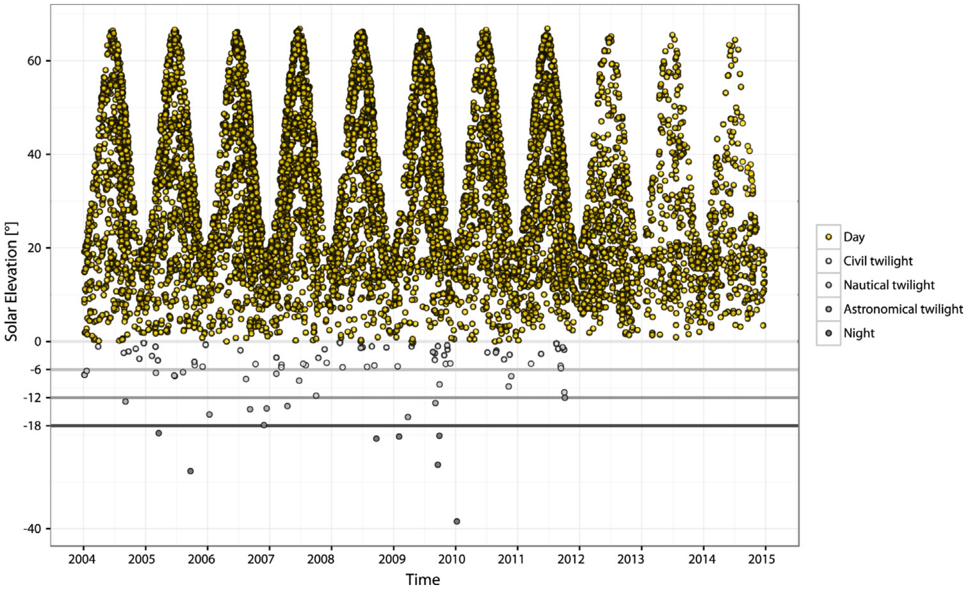

Fig. 7. Calculated solar elevation angles for 7873 sun glare accidents in the period 2004–2014; negative values indicate inconsistencies in the database.

classified into day and night, but also into various stages of twilight, namely civil, nautical and astronomical twilight. Basically, twilight is the time between day and night, when the atmosphere is still illuminated, even though the sun is not directly visible. Civil twilight occurs when the geometric center of the sun is no more than 6 degrees below the horizon at either sunrise or sunset. Nautical twilight is the time period when the elevation angle is between 6 degrees and 12 degrees below the horizon. And finally, astronomical twilight occurs when the elevation angle is between 12 and 18 degrees below the horizon. During this period, the illumination due to scattered light from the sun is already less than that from starlight and other natural light sources in the sky.

As expected, results show a strong seasonal pattern. Surprisingly though, implausible negative values were observed for several records (cf. Fig. 7). This can be either due to erroneous entries in the accident database, e.g. glare caused by artificial light sources, or – as shown above – imprecisely recorded timestamps. Note also that – due to the changes in the data collection process – these issues are no more present in the accident database since 2012. In addition, the overall occurrence of accidents being influenced by sun glare clearly plummeted in 2013.

3.4. *Uncertainties of injury severity estimation*

As far as injury severity is concerned, the degree of severity is most commonly represented on an ordinal scale. Referring to the KABCO-scale, these discrete categories usually comprise the following categories (National Safety Council, 2007; Savolainen et al., 2011; Burch et al., 2014): fatal injury or killed (K), incapacitating injury (A), non-incapacitating evident injury (B), possible injury (C) and non-injury or property damage only (O). Basically, the degree of injury severity is estimated and recorded by police officers on site (Compton, 2005; Burch et al., 2014). Apart from obvious cases, this classification is prone to misjudgements, as a serious assessment of injury severity solely based on preclinical examination is difficult in many cases. More sophisticated methods – like for instance the Abbreviated Injury Scale (AIS), which includes detailed information on traumatic injuries and a classification according to the threat-to-life (Gennarelli and Wodzin, 2007) – include indicators for injury severity which are assigned in hospitals. While it was argued that a simple classification of injury severity according to the KABCO-scale is an appropriate tool for basically discriminating between more serious and minor crashes (Compton, 2005). Farmer (2003) comprehensibly concludes that the lack of professional medical evaluations leads to misclassifications of injury-severity in police-reported data. This entails that these data are to inexact for more specific research applications. While efforts have been made to link hospital and police traffic crash records systemic restrictions due to privacy issues and methodological limitations of e.g. probabilistic linking have to be considered (Wilson et al., 2012).

Further methodological issues, like for instance under-reporting of crashes, omitted variable bias and endogeneity problems as well as issues related to sample size or difficulties caused by correlations are comprehensively presented in Savolainen et al. (2011).

4. Data resolution

Selecting an appropriate spatial and temporal scale when conducting a statistical analysis of accident occurrence is one of the most challenging and crucial parts in accident prediction modeling. While using a coarse resolution might conceal potentially important phenomena by averaging out certain effects (e.g. rainfall extremes reported as precipitation amount accumulated on a daily level), using a too high resolution implies an extremely high precision which in fact cannot be achieved due to aleatoric and epistemic uncertainties inherent to the data (e.g. one-minute precipitation data on a one-meter grid). Thus, a trade-off between robustness and precision has to be made in order to obtain unbiased and accurate results. To some extent, both spatial and temporal resolution are predetermined by the data, their properties, and their associated uncertainties. Choice of spatial and temporal resolution has a determining influence on the choice of the statistical model for accident frequency and accident occurrence analysis. At the same time, model output might yield different results depending on the spatial and temporal scales chosen. Thus, it is important to bring to mind and discuss potential issues in the context of temporal and spatial aspects of data.

4.1. *Spatial resolution*

Spatial resolution plays an important role for accident prediction models with respect to two different aspects:

Firstly, the spatial resolution of the underlying raster and vector data layers (like for instance digital elevation models (DEM), weather data or road condition data) is of importance. Some of these data are potentially available at very high resolutions based on highly precise measurements (e.g. LiDAR based DEM, road condition data), while there are certain physical or practical restrictions with respect to spatial resolution in other data sets (e.g. weather data).

Secondly, taking the aforementioned sources of uncertainty into account, the spatial resolution in relation to the segment length has to be considered carefully. The fact that an accident is a process rather than a specific point in space entails that environmental conditions and other spatial information actually should be considered for the point where the accident was triggered, not for the point where the vehicle finally came to a halt. In combination with the uncertainties regarding the accuracy of accident localization by the police, lack of information with respect to the accident-inducing point makes the assignment of accidents to specific segments of a road somewhat cumbersome. Hence, the length of the road sectors under consideration has to be chosen carefully. On the one hand, segments should be long enough to allow accidents to be mapped to these segments with high confidence. On the other hand, segments should be short enough to reflect distinct properties of the road and the environment. In addition, the choice of segment length also relates to the issue of handling the preponderance of zeros in count data models. The shorter the segment length, the higher the share of segments which do not contain any observed accidents. Albeit several methods have been proposed to deal with this issue (Lord et al., 2005; Malyshkina et al., 2009), a well thought-out choice of segment length may contribute to alleviate this problem. Segment definition can be undertaken in two different ways. Firstly, segments can be defined as sections of equal length along the roads under consideration. Secondly, segments can be defined as homogeneous roadway sections. These natural breaks between the segments lead to varying segment lengths based on e.g. road geometrics (Abdel-Aty and Radwan, 2000; Barua et al., 2015) or uniform risk (Deublein et al., 2013). It has been pointed out by Shankar et al. (1995) that using fixed-length sequences is superior to using homogeneous segments, in particular if detailed data about geometry and weather are available. The authors further underline that disadvantages of using fixed-length sections can hence be overcome by appropriately adjusting methodological specifications, while disadvantages of using homogeneous sections are far more severe. Using bundles of segments which feature large variations in their segment lengths (e.g. between 0.1 miles and 12.22 miles in Anastasopoulos (2016)) may result in detriments with respect to model estimation efficiency, as potential heteroskedasticity problems caused by unequal sample sizes might be exacerbated Shankar et al. (1995).

These two levels of spatial resolution – i.e. resolution of underlying layers as well as segment length – depend on each other and clearly expose where the meaningfulness of high-resolution data is currently stretched to its limits. Thinking about the spatial resolution of the underlying input layers, different levels of uncertainty come to light. While the uncertainties in connection with digital elevation models derived from LiDAR measurements (Goulden et al., 2016) and road condition data are small, limitations related to the accuracy of e.g. weather and land cover data become visible. Both data sets are basically available at spatial resolutions of down to several hundred meters (Broxton et al., 2014), but the estimation is associated with considerable uncertainties (Kann et al., 2012; Congalton et al., 2014; Quaife and Cripps, 2016).

High resolution data is generally valuable as spatially distinct properties can be adequately represented. However, if uncertainties regarding the accident location are larger than the spatial resolution of data sets of potentially relevant explanatory variables, unnecessary uncertainties and efforts can be avoided by using data on a slightly coarser scale. In order to provide some point of reference, recommendations are provided for various data sets against the background of the aforementioned data quality and availability in Austria. Given a median uncertainty of almost 140 m regarding the accident localization on highways, plus considering the uncertainty with respect to the accident inducing point, weather grids of 250 m-resolution are considered to be appropriate. Gridded data sets up to 1 km may be still applicable in less complex terrains, but the use of high-resolution raster data is encouraged in small-scale, mountainous landscapes like the Alpine region, which are topographically and climatically heterogeneous. The same applies to all other raster data sets. While many data sets are gathered at high precision with a higher spatial resolution (e.g. LiDAR DEM), its usefulness is limited by the uncertainties inherent to the accident location itself.

Naturally, data availability and data accessibility has to be taken into account. Depending on the area under investigation, an appropriate spatial scale has to be reasoned against the background of data availability and associated uncertainties.

Against the backdrop of uncertainties inherent to underlying layers and considering uncertainty regarding the exact localization of accidents, segment length has to be chosen accordingly (Lee and Mannering, 2002). Even though general-purpose recommendations are difficult to derive, the following aspects can be stated:

- segment length depends on the hierarchy of roads under consideration, and particularly on the speed limit;
- segment lengths below 100 m are neither serious nor feasible;
- segment lengths above 1000 m are likely to water down site-specific effects;
- operational definitions of hazardous road locations in various European countries stipulate the use of sliding windows with lengths ranging from 100 m to 1 km for identifying black spots (Elvik, 2008);
- site definitions (segments, tunnels, bridges, speed-change lanes, ramps, intersections, etc.)

As there is no panacea for this crucial issue, the importance of clearly elaborating the choice of segment length in any study is emphasized at this point.

4.2. Temporal resolution

As with spatial resolution, temporal resolution is also an important aspect of accident occurrence analysis that features some pitfalls worthy of discussion.

The major impediments with respect to temporal resolution are caused by rounding errors when estimating the exact accident time. Uncertainties resulting from these rounding effects restrict the reasonableness of data with a high temporal resolution. Any form of accident analysis which uses sub-hourly data should thus be taken with a grain of salt if awareness and treatment of uncertainties are not tackled properly. In addition, this biased distribution is a challenge for setting the breakpoints of temporal stratification. Usually, temporal classes start on the hour, which is in fact the point in time where the highest number of accidents is affected. Hence, the uncertainty related to the assignment of the accident towards the earlier or the later interval affects a good deal of accidents.

Another issue in the context of temporal uncertainties has already been recognized in recent reviews by Lord and Mannering (2010) and Washington et al. (2010). They mention that ignoring variations in explanatory variables over the time period under consideration may lead to information loss if the time intervals are too large. However, while this fact is mainly attributed to the lack of detailed data within the time period in Lord and Mannering (2010), we have showed that more detailed data – in terms of a higher temporal resolution – does not necessarily prove to be of much help if the underlying processes generating these variables are complex and thus implicate considerable variations that have to be taken into account.

Thus, it is important to choose the strata accordingly. On the one hand, already data at daily resolution might conceal potentially relevant information. This is particularly true if weather effects are investigated. On the other hand, given the uncertainties with respect to the actual accident time in terms of rounding effects, using temporal resolutions smaller than one hour are hardly tenable. While gridded meteorological data sets featuring a sub-hourly temporal resolution are difficult to obtain, the precision of these data is elusive due to the lack of certainty with respect to accurately assigning accidents to the appropriate time slots. While robustness can be increased by using hourly values, distinct variations in the variables are still clearly observable. It is hypothesized that using a different break point instead of the full hour might further alleviate this issue, as uncertainties are highest around the full hour.

As far as traffic load is concerned, hourly values are also considered to be the most useful resolution. While counts aggregated at an hourly level exhibit a beneficial amount of robustness, they are specific enough to capture distinct variations.

Furthermore, the length of the time period under consideration is another aspect worth of discussion. Again, this issue has to be considered between the conflicting priorities of accuracy and bias. While longer time series naturally contribute to increasing confidence of analysis results, temporal changes and variations of relevant variables over time may severely impede a sound analysis. In particular, bias is introduced by activities related to reconstruction, like road works, changes of the road superstructure (including gradients, skid resistance, etc.) or changes in the course of the road (e.g. straightening). These parameters are usually considered to be constant within the time period under consideration, even though there are most certainly – at least minor – changes over time. If these possible variations are not considered, the resulting analysis is likely to yield severely distorted results.

In order to capture seasonal effects, using at least one year of data is strongly recommended. Besides, shorter series will negatively affect the sample size. Using several years of data is beneficial in terms of model validity. However, care has to be taken when analyzing time periods spanning more than five years. On the one hand, structural changes of roads might lead to inconsistencies. On the other hand, possible temporal trends have to be considered (Lord and Persaud, 2000; Wu et al., 2014b).

5. Data (pre)processing

Having gained clarity about quality and properties of available data, as well as having thought about appropriate spatial and temporal resolution of the input data sets, the next important step before being able to eventually come up with a statistical model is data preprocessing. Data preparation is usually performed by using appropriate software environments (like for instance R (R Core Team, 2016) or Python (Python Software Foundation, 2016)), often in connection with some (preferably spatial) database (e.g. PostGIS (Refractions Research Inc, 2016)) for managing large amounts of data. Concerning data for road safety management, guidelines on how to set up data systems that should produce comparable data are available (WHO, 2010; OECD/ITF, 2015).

Despite the availability of both powerful software packages and computational power, the easy access to capable software should not hide the fact that there are still several serious methodological barriers that have to be broken down when conducting a statistical analysis of accident data. In fact, the intellectual work and various preliminary considerations related to data processing – taking the limitations of various data sources into account – are probably more demanding than the actual programming work itself.

One challenging task is to derive meaningful and representative values for the various parameters from the underlying raster and vector data sets. This is rather straightforward when modeling accident frequency, as every accident is just a single event that has a flag in terms of an exact timestamp and location. Values for this specific coordinates and time can simply be extracted from any other layer. However, as far as accident frequency analysis is concerned, a rule has to be defined how these parameters can be matched to the segments at which the accidents are counted. In this case, the target variable is not recorded on a single spot, but rather on a linestring at a certain point in time. Naturally, these linestrings may traverse multiple raster cells or across/along various vector features (points, lines, polygons). In order to obtain a single value for each segment that can be used as a covariate for further modeling, two different methods seem plausible. On the one hand, values can be derived as weighted average according to the share of segments located within raster cells or vector feature. On the other hand, simply the minimum/maximum of all values can be assigned to the respective segment.

This issue gets even more crucial when it comes to deriving representative indicators for road condition data and road geometrics, which can be assigned to a whole road segment. Determining meaningful and informative feature definitions is particularly important if fixed-length segments are used, where roadway properties may change within the segment. Aggregating data, which are measured (with some noise) at intervals of 1 m, into one value representing a whole segment, is a complex task. Feature engineering (and consequently feature selection) is a difficult yet potentially rewarding process that may provide new insights into characterizing road segments. Unfortunately, not much effort has been undertaken in this direction so far. In many cases, simply maxima, averages or indices (e.g. pavement condition index) are used. It seems worthwhile to foster further research on feature engineering in the area of road segment characterization in order to explore the possible benefits of using new indicators (e.g. certain quantiles) off the beaten track for characterizing road segments.

Given a variety of potentially relevant covariates – comprising data related to persons involved (age, gender, sobriety, etc.), vehicles involved (engine power, type of vehicle, etc.), accident location (road type, rural vs urban area, speed limit, etc.), road geometry (grades, skid resistance, etc.), traffic load (AADT, share of heavy good vehicles, etc.), road environment (land cover, etc.) or weather (precipitation, wind, temperature, humidity, sun glare, etc.) and so forth – this begs the question of how to identify the actually relevant variables. A carefully considered and suitable dimensionality reduction is one of the keys to success for performing a sound analysis. Various machine learning algorithms for dimension reduction, such as principal component analysis, random forests or artificial neural networks might prove useful in this case. Thus, collinearities can be identified and taken into account.

Lord and Mannering (2010) state that the Poisson-gamma/negative binomial model is the probably most frequently used model in crash-frequency modeling. If variable selection is performed within such a parametric regression model, regularization techniques like LASSO (Tibshirani, 1996, 2011; Friedman et al., 2010) or Boosting (Hothorn et al., 2010) should be given preference to often used stepwise model selection methods, which are prone to overfitting (Hastie et al., 2009; Harrell, 2015). This is essential if a model is used for prediction, and not just for statistical inference.

Of course, care has to be taken to consider the underlying causality between the covariates and the target variable by specifying appropriate hypothesis. Otherwise, the whole analysis degenerates to a process of mere data dredging.

While there are issues related to the plethora of possible covariates, one of the common problems is caused by an opposite effect, namely the problem of unobserved heterogeneity. The issue that certain factors that most likely affect the occurrence of accidents are either difficult to collect or cannot be observed at all may lead to erroneous conclusions (Mitra and Washington, 2012; Garnowski and Manner, 2011).

As far as further considerations about methodological issues closely related to the aspects of model fitting and interpretation are concerned, it is referred to Lord and Mannering (2010), Savolainen et al. (2011) and Mannering and Bhat (2014).

6. Summary and conclusions

The preceding discussion of uncertainties inherent to various data types and sources illustrates that care has to be exercised when linking different kinds of information in order to perform accident analyses. While a lot of effort has been put into advancing the predictive quality of accident models, detailed discussions of data properties related to information used within these models have been largely eclipsed. Research has tended to focus on modeling approaches, upstaging important considerations related to data themselves.

Lord and Mannering (2010), Savolainen et al. (2011) and Mannering and Bhat (2014) have done seminal work on highlighting methodological considerations with respect to statistical properties of data used in accident modeling. In this review we have gone beyond the scope of their work, expanding these methodological considerations to the process of data wrangling. The underlying uncertainties associated with the data that emerge during this process have been illustrated, allowing us to tackle several assumptions that were implicitly and succinctly made in various other studies.

We have outlined the issues that may occur when trying to expanding police reported data further by linking it to various other data, like roadway geometric characteristics, traffic volumes or weather-related data. From a data scientist's perspective, this assumption, which implies a straightforward procedure, grossly neglects the efforts related to data preparation for road accident analysis. Challenges range from issues of data engineering to various conceptual concerns with respect to spatial and temporal resolution. Uncertainties caused by human factors (e.g. sloppy data entries or rounding errors), by technical (e.g. measurement accuracy) or methodological (e.g. downscaling of reanalysis data) limitations or due to circumstances beyond control (e.g. practical impossibility to determine exact accident location and time) have to be kept in mind when drawing inferences. Even though high resolution data as described in this article can be considered to be the best estimate for reflecting reality, many of these data sets are still just – albeit comparatively good – approximations of the true values and processes, which are used due to default of better alternatives. Of course, actual data availability and quality for specific regions of interest might differ substantially from the theoretically best options, imposing further restrictions.

As it is virtually impossible to utterly eliminate uncertainties from any type of statistical analysis, accident modeling always has to be carried out in due consideration of the limitations imposed by available data themselves. Therefore, both description of the raw data used and the efforts which are put in data preparation, including important data processing steps, are an integral part of any accident prediction model.

This work has demonstrated some important sources of uncertainty that can commonly be observed in accident modeling. In the knowledge that various implications of both data and model uncertainty have to be conceded, accident analysis will continue to prove to be a highly valuable tool for improving safety policies and preventing injuries and loss of life.

Acknowledgments

The authors gratefully acknowledge ASFiNAG for providing traffic volume data and ÖAMTC for providing alert times of the air rescue fleet.

References

Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol. 33, 121–131, http://dx.doi.org/10.1002/joc.3413.

Abdel-Aty, M.A., Radwan, A., 2000. Modeling traffic accident occurrence and involvement. Accid. Anal. Prev. 32, 633–642, http://dx.doi.org/10.1016/S0001-4575(99)00094-9.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Anal. Methods Accid. Res. 11, 17–32, http://dx.doi.org/10.1016/j.amar.2016.06.001.

Andrey, J., Yagar, S., 1993. A temporal analysis of rain-related crash risk. Accid. Anal. Prev. 25, 465–472, http://dx.doi.org/10.1016/0001-4575(93)90076-9.

Antoniou, C., Balakrishna, R., Koutsopoulos, H.N., 2011. A synthesis of emerging data collection technologies and their impact on traffic management applications. Eur. Transp. Res. Rev. 3, 139–148, http://dx.doi.org/10.1007/s12544-011-0058-1.

Auer, I., Böhm, R., Jurkovic, A., Lipa, W., Orlik, A., Potzmann, R., Schöner, W., Ungersböck, M., Matulla, C., Briffa, K., Jones, P., Efthymiadis, D., Brunetti, M., Nanni, T., Maugeri, M., Mercalli, L., Mestre, O., Moisselin, J.-M., Begert, M., Müller-Westermeier, G., Kveton, V., Bochnicek, O., Stastny, P., Lapin, M., Zalai, S., Szentimrey, T., Cegnar, T., Dolinar, M., Gajic-Capka, M., Zaninovic, K., Majstorovic, Z., Nieplova, E., 2007. HISTALP – historical instrumental climatological surface time series of the Greater Alpine Region. Int. J. Climatol. 27, 17–46, http://dx.doi.org/10.1002/joc.1377.

Barua, S., El-Basyouny, K., Islam, M.T., 2015. Effects of spatial correlation in random parameters collision count-data models. Anal. Methods Accid. Res. 5-6, 28–42, http://dx.doi.org/10.1016/j.amar.2015.02.001.

Benestad, R.E., 2004. Empirical-statistical downscaling in climate modeling. Eos Trans. Am. Geophys. Union 85, 417–422, http://dx.doi.org/10.1029/2004EO420002.

Bergel-Hayat, R., Debbarh, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: weather effects. Accid. Anal. Prev. 60, 456–465, http://dx.doi.org/10.1016/j.aap.2013.03.006.

Bica, B., Steinacker, R., Lotteraner, C.J., Suklitsch, M., 2007. A new concept for high resolution temperature analysis over complex terrain. Theor. Appl. Climatol. 90, 173–183, http://dx.doi.org/10.1007/s00704-006-0280-2.

Bivand, R., Lewin-Koh, N., 2016. maptools: Tools for Reading and Handling Spatial Objects. https://CRAN.R-project.org/package=maptools, r package version 08-39.

BMVIT, 2011. Verkehrssicherheitsprogramm 2011–2020 (Traffic Safety Program 2011–2020). Austrian Federal Ministry for Transport, Innovation and Technology.

BMVIT, 2013. Verkehrssicherheit in Österreich – Jahresbericht 2012 (Road Safety in Austria Annual Report 2012). Austrian Federal Ministry for Transport, Innovation and Technology.

BMVIT, 2016. Automatische Straßenverkehrszählung 2014 – Bundesweite Auswertung (Automated Road Traffic Census – Nationwide Evaluation). Austrian Federal Ministry for Transport, Innovation and Technology.

Borsche, M., Kaiser-Weiss, A.K., Undén, P., Kaspar, F., 2015. Methodologies to characterize uncertainties in regional reanalyses. Adv. Sci. Res. 12, 207–218, http://dx.doi.org/10.5194/asr-12-207-2015.

Boudevillain, B., Delrieu, G., Wijbrans, A., Confoland, A., 2016. A high-resolution rainfall re-analysis based on radar-raingauge merging in the Cévennes-Vivarais region, France. J. Hydrol., http://dx.doi.org/10.1016/j.jhydrol.2016.03.058 (in press).

Brands, S., Gutiérrez, J.M., Herrera, S., Cofi no, A.S., 2012. On the use of reanalysis data for downscaling. J. Clim. 25, 2517–2526, http://dx.doi.org/10.1175/JCLI-D-11-00251.1.

Brijs, T., Karlis, D., Wets, G., 2008. Studying the effect of weather conditions on daily crash counts using a discrete time-series model. Accid. Anal. Prev. 40, 1180–1190, http://dx.doi.org/10.1016/j.aap.2008.01.001.

Brodsky, H., Hakkert, A.S., 1988. Risk of a road accident in rainy weather. Accid. Anal. Prev. 20, 161–176, http://dx.doi.org/10.1016/0001-4575(88)90001-2.

Broxton, P.D., Zeng, X., Sulla-Menashe, D., Troch, P.A., 2014. A global land cover climatology using modis data. J. Appl. Meteorol. Climatol. 53, 1593–1605, http://dx.doi.org/10.1175/JAMC-D-13-0270.1.

Burch, C., Cook, L., Dischinger, P., 2014. A comparison of KABCO and AIS injury severity metrics using CODES linked data. Traffic Inj. Prev. 15, 627–630, http://dx.doi.org/10.1080/15389588.2013.854348, PMID: 24261347.

Compton, C.P., 2005. Injury severity codes: a comparison of police injury codes and medical outcomes as determined by {NASS} {CDS} investigators. Proceedings of the Traffic Records Forum, Buffalo, NY, USA, August 2, 2005. J. Saf. Res. 36, 483–484, http://dx.doi.org/10.1016/j.jsr.2005.10.008.

Congalton, R.G., Gu, J., Yadav, K., Thenkabail, P., Ozdogan, M., 2014. Global land cover mapping: a review and uncertainty analysis. Remote Sens. 6, 12070–12093, http://dx.doi.org/10.3390/rs61212070.

Council of European Union, 2008. Directive 2008/96/EC of the European Parliament and of the Council on Road Infrastructure Safety Management. http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008L0096.

Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Doggett, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol. 28, 2031–2064, http://dx.doi.org/10.1002/joc.1688.

Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J. Appl. Meteorol. Climatol. 33, 140–158, http://dx.doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Dasu, T., Johnson, T., 2003. Exploratory Data Mining and Data Cleaning. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J.P., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597, http://dx.doi.org/10.1002/qj.828.

DeGaetano, A.T., Belcher, B.N., 2007. Spatial interpolation of daily maximum and minimum air temperature based on meteorological model analyses and independent observations. J. Appl. Meteorol. Climatol. 46, 1981–1992, http://dx.doi.org/10.1175/2007JAMC1536.1.

Deublein, M., Schubert, M., Adey, B.T., Köhler, J., Faber, M.H., 2013. Prediction of road accidents: a Bayesian hierarchical approach. Accid. Anal. Prev. 51, 274–291, http://dx.doi.org/10.1016/j.aap.2012.11.019.

Edwards, J.B., 1996. Weather-related road accidents in England and Wales: a spatial analysis. J. Transp. Geogr. 4, 201–212, http://dx.doi.org/10.1016/0966-6923(96)00006-3.

Eisenberg, D., 2004. The mixed effects of precipitation on traffic crashes. Accid. Anal. Prev. 36, 637–647, http://dx.doi.org/10.1016/S0001-4575(03)00085-X.

Elvik, R., 2008. A survey of operational definitions of hazardous road locations in some European countries. Accid. Anal. Prev. 40, 1830–1835, http://dx.doi.org/10.1016/j.aap.2008.08.001.

Farmer, C.M., 2003. Reliability of police-reported information for determining crash and injury severity. Traffic Inj. Prev. 4, 38–44, http://dx.doi.org/10.1080/15389580309855, PMID: 14522660.

FHWA, 2013. Traffic Monitoring Guide. Federal Highway Administration.

Findley, D.J., Cunningham, C.M., Hummer, J.E., 2011. Comparison of mobile and manual data collection for roadway components. Transp. Res. Part C: Emerg. Technol. 19, 521–540, http://dx.doi.org/10.1016/j.trc.2010.08.002.

Frei, C., 2014. Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances. Int. J. Climatol. 34, 1585–1605, http://dx.doi.org/10.1002/joc.3786.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., Thomsen, L.K., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accid. Anal. Prev. 27, 1–20, http://dx.doi.org/10.1016/0001-4575(94)E0023-E.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22, http://dx.doi.org/10.18637/jss.v033.i01.

FSV, 2002. RVS 11.06.65: Baudurchführung – Grundlagen – Prüfverfahren – Feldprüufungen, Teil V: Griffigkeitsmessungen mit dem Stuttgarter Reibungsmesser (System RoadSTAR) [execution of construction works – basics – testing methods – in-situ testing, part V: grip measurements with the Stuttgart friction meter (RoadSTAR system)]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

FSV, 2004a. RVS 11.06.66: Baudurchführung – Grundlagen – Prüfverfahren – Feldprüufungen, Teil VI: Lasertexturmessungen mit dem System RoadSTAR. [execution of construction works – basics – testing methods – in-situ testing, part V: Laser texture measurements with the RoadSTAR system]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

FSV, 2004b. RVS 11.06.67: Baudurchführung – Grundlagen – Prüfverfahren – Feldprüufungen, Teil VII: Querebenheitsmessungen mit dem System RoadSTAR. [execution of construction works – basics – testing methods – in-situ testing, part VII: Transverse flatness measurements with the RoadSTAR system]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

FSV, 2004c. RVS 11.06.68: Baudurchführung – Grundlagen – Prüfverfahren – Feldprüufungen, Teil VIII: Längsebenheitsmessungen mit dem System RoadSTAR. [execution of construction works – basics – testing methods – in-situ testing, part VIII: Longitudinal flatness measurements with the RoadSTAR system]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

FSV, 2009a. RVS 11.06.69: Qualitätssicherung Bau – Prüfungen – Fahrbahnoberflä: Digitale Hochgeschwindigkeitsbilderfassung der Fahrbahnoberfläche mit dem System RoadSTAR. [quality assurance construction – tests – road surface: Digital high-speed image capturing of road surfaces with the RoadSTAR system]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

FSV, 2009b. RVS 13.01.15: Qualitätssicherung Bauliche Erhaltung – Bauliche Straßenerhaltung – Pavement Management: Beurteilungskriterien für messtechnische Zustandserfassung mit dem System RoadSTAR. [quality assurance for structural maintenance – structural road maintenance – pavement management: Assessment criteria for pavement condition measurements with the RoadSTAR system]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

FSV, 2015. RVS 02.01.12: Verkehrsplanung – Grundlagen – Verkehrsuntersuchungen: Straßenverkehrszählungen [traffic planning – basics – transport analyses: Traffic counting]. Richtlinien und Vorschriften für das Straßenwesen [Guidelines and regulations for road engineering].

Fu, G., Charles, S.P., Timbal, B., Jovanovic, B., Ouyang, F., 2016. Comparison of NCEP-NCAR and ERA-Interim over Australia. Int. J. Climatol. 36, 2345–2367, http://dx.doi.org/10.1002/joc.4499.

Gao, L., Schulz, K., Bernhardt, M., 2014. Statistical downscaling of era-interim forecast precipitation data in complex terrain using LASSO algorithm. In: Adv. Meteorol., http://dx.doi.org/10.1155/2014/472741, Article ID 472741.

Garnowski, M., Manner, H., 2011. On factors related to car accidents on German autobahn connectors. Accid. Anal. Prev. 43, 1864–1871, http://dx.doi.org/10.1016/j.aap.2011.04.026.

Gennarelli, T.A., Wodzin, E., 2007. The Abbreviated Injury Scale 2005. American Association for Automotive Medicine (AAAM), Update 2008.

Giorgi, F., Gutowski Jr., W.J., 2015. Regional dynamical downscaling and the CORDEX initiative. Annu. Rev. Environ. Resour. 40, 467–490, http://dx.doi.org/10.1146/annurev-environ-102014-021217.

GIP.gv.at, 2016. Graphenintegrations-Plattform (Graph Integration Platform). http://www.GIP.gv.at (accessed 27.06.16).

Goulden, T., Hopkinson, C., Jamieson, R., Sterling, S., 2016. Sensitivity of DEM, slope, aspect and watershed attributes to LiDAR measurement uncertainty. Remote Sens. Environ. 179, 23–35, http://dx.doi.org/10.1016/j.rse.2016.03.005.

Gyasi-Agyei, Y., Pegram, G., 2014. Interpolation of daily rainfall networks using simulated radar fields for realistic hydrological modelling of spatial rain field ensembles. J. Hydrol. 519 (Part A), 777–791, http://dx.doi.org/10.1016/j.jhydrol.2014.08.006.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B., Gruber, C., 2011. The integrated nowcasting through comprehensive analysis (INCA) system and its validation over the eastern alpine region. Weather Forecast. 26, 166–183, http://dx.doi.org/10.1175/2010WAF2222451.1.

Harrell, F., 2015. Regression Modeling Strategies. With Applications to Linear Models. Logistic and Ordinal Regression, and Survival Analysis. Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-19425-7.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, second edition. Springer, New York, http://dx.doi.org/10.1007/978-0-387-84858-7.

Haylock, M.R., Hofstra, N., Klein Tank, A.M.G., Klok, E.J., Jones, P.D., New, M., 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. J. Geophys. Res.: Atmos. 113, D20119, http://dx.doi.org/10.1029/2008JD010201.

Hiebl, J., Frei, C., 2016. Daily temperature grids for Austria since 1961 – concept, creation and applicability. Theor. Appl. Climatol. 124, 161–178, http://dx.doi.org/10.1007/s00704-015-1411-4.

Hofstra, N., Haylock, M., New, M., Jones, P., Frei, C., 2008. Comparison of six methods for the interpolation of daily, European climate data. J. Geophys. Res.: Atmos. 113, D21110, http://dx.doi.org/10.1029/2008JD010100.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., Hofner, B., 2010. Model-based boosting 2.0. J. Mach. Learn. Res. 11, 2109–2113.

Hu, Z., Hu, Q., Zhang, C., Chen, X., Li, Q., 2016. Evaluation of reanalysis, spatially interpolated and satellite remotely sensed precipitation data sets in central Asia. J. Geophys. Res.: Atmos. 121, 5648–5663, http://dx.doi.org/10.1002/2016JD024781.

Jakobson, E., Vihma, T., Palo, T., Jakobson, L., Keernik, H., Jaagus, J., 2012. Validation of atmospheric reanalyses over the central Arctic Ocean. Geophys. Res. Lett. 39, http://dx.doi.org/10.1029/2012GL051591.L10802.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accid. Anal. Prev. 42, 213–224, http://dx.doi.org/10.1016/j.aap.2009.07.020.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bull. Am. Meteorol. Soc. 77, 437–471, http://dx.doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kann, A., Pistotnik, G., Bica, B., 2012. INCA-CE: a Central European initiative in nowcasting severe weather and its applications. Adv. Sci. Res. 8, 67–75, http://dx.doi.org/10.5194/asr-8-67-2012.

Kobayashi, S., Ota, Y., Harada, Y., Ebita, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K., Takahash, K., 2015. The JRA-55 reanalysis: general specifications and basic characteristics. J. Meteorol. Soc. Jpn. 93, 5–48, http://dx.doi.org/10.2151/jmsj.2015-001.

Koetse, M.J., Rietveld, P., 2009. The impact of climate change and weather on transport: an overview of empirical findings. Transp. Res. Part D: Transp. Environ. 14, 205–221, http://dx.doi.org/10.1016/j.trd.2008.12.004.

Kumara, S., Chin, H., 2005. Application of Poisson underreporting model to examine crash frequencies at signalized three-legged intersections. Transp. Res. Rec. 1908, 46–50, http://dx.doi.org/10.3141/1908-06.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis. Accid. Anal. Prev. 34, 149–161, http://dx.doi.org/10.1016/S0001-4575(01)00009-4.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transp. Res. Part A: Policy Pract. 44, 291–305, http://dx.doi.org/10.1016/j.tra.2010.02.001.

Lord, D., Persaud, B., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transp. Res. Rec. 1717, 102–108, http://dx.doi.org/10.3141/1717-13.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37, 35–46, http://dx.doi.org/10.1016/j.aap.2004.02.004.

Lowry, M., 2014. Spatial interpolation of traffic counts based on origin–destination centrality. J. Transp. Geogr. 36, 98–105, http://dx.doi.org/10.1016/j.jtrangeo.2014.03.007.

Luhamaa, A., Kimmel, K., Männik, A., R o om, R., 2011. High resolution re-analysis for the Baltic Sea region during 1965–2005 period. Clim. Dyn. 36, 727–738, http://dx.doi.org/10.1007/s00382-010-0842-y.

Ma, J., 2009. Bayesian analysis of underreporting Poisson regression model with an application to traffic crashes on two-lane highways. In: TRB 88th Annual Meeting Compendium of Papers, Presented at the 88th Annual Meeting of the Transportation Research Board. Washington, DC.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: an application to vehicle accident frequencies. Accid. Anal. Prev. 41, 217–226, http://dx.doi.org/10.1016/j.aap.2008.11.001.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22, http://dx.doi.org/10.1016/j.amar.2013.09.001.

Mao, B., Tian, Z., Huang, H., Gao, Z., Zheng, P., Mike, M., 2012. An investigation on the manual traffic count accuracy. Procedia Soc. Behav. Sci. 43, 226–231, http://dx.doi.org/10.1016/j.sbspro.2012.04.095.

Martin, J.-L., 2002. Relationship between crash rate and hourly traffic flow on interurban motorways. Accid. Anal. Prev. 34, 619–629, http://dx.doi.org/10.1016/S0001-4575(01)00061-6.

Matulla, C., 2005. Regional, seasonal and predictor-optimized downscaling to provide groups of local scale scenarios in the complex structured terrain of Austria. Meteorol. Z. 14, 31–45, http://dx.doi.org/10.1127/0941-2948/2005/0014-0031.

Matulla, C., Groll, N., Kromp-Kolb, H., Scheifinger, H., Lexer, M.J., Widmann, M., 2002. Climate change scenarios at Austrian national forest inventory sites. Clim. Res. 22, 161–173, http://dx.doi.org/10.3354/cr022161.

Mitra, S., 2014. Sun glare and road safety: an empirical investigation of intersection crashes. Saf. Science 70, 246–254, http://dx.doi.org/10.1016/j.ssci.2014.06.005.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accid. Anal. Prev. 49, 439–448, http://dx.doi.org/10.1016/j.aap.2012.03.014.

Mooney, P.A., Mulligan, F.J., Fealy, R., 2011. Comparison of ERA-40. ERA-Interim and NCEP/NCAR reanalysis data with observed surface air temperatures over Ireland. Int. J. Climatol. 31, 545–557, http://dx.doi.org/10.1002/joc.2098.

National Safety Council, 2007. Manual on the Classification of Motor Vehicle Traffic Accidents (ANSI D16.1-2007), seventh edition.

Ni, D., 2016. Chapter 1 – traffic sensing technologies. In: Ni, D. (Ed.), Traffic Flow Theory. Butterworth-Heinemann, pp. 3–17, http://dx.doi.org/10.1016/B978-0-12-804134-5.00001-5.

OECD/ITF, 2015. Road Safety Annual Report 2015. International Transport Forum, Paris, http://dx.doi.org/10.1787/irtad-2015-en.

Patil, S., Geedipally, S.R., Lord, D., 2012. Analysis of crash severities using nested logit model – accounting for the underreporting of crashes. Accid. Anal. Prev. 45, 646–653, http://dx.doi.org/10.1016/j.aap.2011.09.034.

Python Software Foundation, 2016. Python Programming Language. Delaware, United States of America. http://www.python.org.

Quaife, T., Cripps, E., 2016. Bayesian analysis of uncertainty in the globcover 2009 land cover product at climate model grid scale. Remote Sens. 8, 314, http://dx.doi.org/10.3390/rs8040314.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Refractions Research Inc., 2016. PostGIS. http://www.postgis.net/.

Rienecker, M.M., Suarez, M.J., Gelaro, R., Todling, R., Bacmeister, J., Liu, E., Bosilovich, M.G., Schubert, S.D., Takacs, L., Kim, G.-K., Bloom, S., Chen, J., Collins, D., Conaty, A., da Silva, A., Gu, W., Joiner, J., Koster, R.D., Lucchesi, R., Molod, A., Owens, T., Pawson, S., Pegion, P., Redder, C.R., Reichle, R., Robertson, F.R., Ruddick, A.G., Sienkiewicz, M., Woollen, J., 2011. MERRA: NASA's modern-era retrospective analysis for research and applications. J. Clim. 24, 3624–3648, http://dx.doi.org/10.1175/JCLI-D-11-00015.1.

Rose, S., Apt, J., 2016. Quantifying sources of uncertainty in reanalysis derived wind speed. Renew. Energy 94, 157–165, http://dx.doi.org/10.1016/j.renene.2016.03.028.

Savolainen, P.T., Mannering, F.L., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives. Accid. Anal. Prev. 43, 1666–1676, http://dx.doi.org/10.1016/j.aap.2011.03.025.

Schmidli, J., Frei, C., Schär, C., 2001. Reconstruction of mesoscale precipitation fields from sparse observations in complex terrain. J. Clim. 14, 3289–3306, http://dx.doi.org/10.1175/1520-0442(2001)014<3289:ROMPFF>2.0.CO;2.

Selby, B., Kockelman, K.M., 2013. Spatial prediction of traffic levels in unmeasured locations: applications of universal kriging and geographically weighted regression. J. Transp. Geogr. 29, 24–32, http://dx.doi.org/10.1016/j.jtrangeo.2012.12.009.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27, 371–389, http://dx.doi.org/10.1016/0001-4575(94)00078-Z.

Sherretz, L.A., Farhar, B., 1978. An analysis of the relationship between rainfall and the occurrence of traffic accidents. J. Appl. Meteorol. 17, 711–715.

Song, C., Ke, L., Richards, K.S., Cui, Y., 2016. Homogenization of surface temperature data in high mountain Asia through comparison of reanalysis data and station observations. Int. J. Climatol. 36, 1088–1101, http://dx.doi.org/10.1002/joc.4403.

Steinacker, R., Mayer, D., Steiner, A., 2011a. Data quality control based on self-consistency. Mon. Weather Rev., 139, http://dx.doi.org/10.1175/MWR-D-10-05024.1.

Steinacker, R., Ratheiser, M., Bica, B., Chimani, B., Dorninger, M., Gepp, W., Lotteraner, C., Schneider, S., Tschannett, S., 2011b. A mesoscale data analysis and downscaling method over complex terrain. Mon. Weather Rev. 134, 2758–2771, http://dx.doi.org/10.1175/MWR3196.1.

Szentimrey, T., Bihari, Z., Lakatos, M., Szalai, S., 2011. Mathematical, methodological questions concerning the spatial interpolation of climate elements. Időjárás 118, 1–11.

Tarko, A., Boyle, L.N., Montella, A., 2013. Emerging research methods and their application to road safety. Accid. Anal. Prev. 61, 1–2, http://dx.doi.org/10.1016/j.aap.2013.07.006.

Teegavarapu, R.S., Meskele, T., Pathak, C.S., 2012. Geo-spatial grid-based transformations of precipitation estimates using spatial interpolation methods. Comput. Geosci. 40, 28–39, http://dx.doi.org/10.1016/j.cageo.2011.07.004.

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256, http://dx.doi.org/10.1016/j.aap.2014.06.017.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc.: Ser. B: Methodol. 58, 267–288.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc.: Ser. B: Stat. Methodol. 73, 273–282, http://dx.doi.org/10.1111/j.1467-9868.2011.00771.x.

Uppala, S.M., Kållberg, P.W., Simmons, A.J., Andrae, U., Bechtold, V.D.C., Fiorino, M., Gibson, J.K., Haseler, J., Hernandez, A., Kelly, G.A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R.P., Andersson, E., Arpe, K., Balmaseda, M.A., Beljaars, A.C.M., Berg, L.V.D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B.J., 2005. The ERA-40 re-analysis. Q. J. R. Meteorol. Soc. 131, 2961–3012, http://dx.doi.org/10.1256/qj.04.176.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Chapman Hall/CRC, Boca Raton, FL.

WHO, 2004. World Report on Road Traffic Injury Prevention. World Health Organization.

WHO, 2010. Data Systems: A Road Safety Manual for Decision-Makers and Practitioners. World Health Organization.

WHO, 2015. Global Status Report on Road Safety 2015. World Health Organization.

Wickham, H., 2014. Tidy data. J. Stat. Softw. 59, 1–23, http://dx.doi.org/10.18637/jss.v059.i10.

Wilby, R., Wigley, T., 1997. Downscaling general circulation model output: a review of methods and limitations. Prog. Phys. Geogr. 21, 530–548, http://dx.doi.org/10.1177/030913339702100403.

Wilson, S.J., Begg, D.J., Samaranayaka, A., 2012. Validity of using linked hospital and police traffic crash records to analyse motorcycle injury crash characteristics. Accid. Anal. Prev. 49, 30–35, http://dx.doi.org/10.1016/j.aap.2011.03.007.

WMO, 2008. WMO Guide to Meteorological Instruments and Methods of Observation (WMO-No. 8, CIMO-Guide). World Meteorological Organization, Geneva, Updated in 2010.

WMO, 2010. Manual on the Global Observing System (WMO-No. 544). World Meteorological Organization, Geneva, Updated in 2013.

WMO, 2011. Technical Regulations (WMO-No. 49). World Meteorological Organization, Geneva, Updated in 2012.

Wu, K.-F., Aguero-Valverde, J., Jovanis, P.P., 2014a. Using naturalistic driving data to explore the association between traffic safety-related events and crash risk at driver level. Accid. Anal. Prev. 72, 210–218, http://dx.doi.org/10.1016/j.aap.2014.07.005.

Wu, W.-Q., Wang, W., Li, Z.-B., Liu, P., Wang, Y., 2014b. Application of generalized estimating equations for crash frequency modeling with temporal correlation. J. Zhejiang Univ. Sci. A: Appl. Phys. Eng. 15, 529–539, http://dx.doi.org/10.1631/jzus.A1300342.

Xia, Y., Shi, X., Song, G., Geng, Q., Liu, Y., 2016. Towards improving quality of video-based vehicle counting method for traffic flow estimation. Signal Process. 120, 672–681, http://dx.doi.org/10.1016/j.sigpro.2014.10.035.

Yamamoto, T., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accid. Anal. Prev. 40, 1320–1329, http://dx.doi.org/10.1016/j.aap.2007.10.016.

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accid. Anal. Prev. 59, 506–521, http://dx.doi.org/10.1016/j.aap.2013.06.040.

Ye, F., Lord, D., 2011. Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. Transp. Res. Rec. 2241, 51–58, http://dx.doi.org/10.3141/2241-06.

Zhao, F., Park, N., 2004. Using geographically weighted regression models to estimate annual average daily traffic. Transp. Res. Rec. 1879, 99–107, http://dx.doi.org/10.3141/1879-12.

Zhou, M., Sisiopiku, V., 1997. Relationship between volume-to-capacity ratios and accident rates. Transp. Res. Rec. 1581, 47–52, http://dx.doi.org/10.3141/1581-06.