Accident Analysis and Prevention 127 (2019) 134–149

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset

Matthias Schlögl[a,b,*], Rainer Stütz[c], Gregor Laaha[a], Michael Melcher[a]

[a] Institute of Applied Statistics and Scientific Computing, University of Natural Resources and Life Sciences, Vienna, Austria
[b] Transportation Infrastructure Technologies, Austrian Institute of Technology, Vienna, Austria
[c] Digital Insight Lab, Austrian Institute of Technology, Vienna, Austria

ARTICLE INFO

MSC:
62-07

Keywords:
Statistical learning
Imbalanced data
Binary classification
Accident analysis
Road safety

ABSTRACT

One of the main aims of accident data analysis is to derive the determining factors associated with road traffic accident occurrence. While current studies mainly use variants of count data regression to achieve this aim, the problem can also be considered as a binary classification task, with the dichotomous target variable indicating events (accidents) and non-events (no accidents). The effects of 45 variables – describing road condition and geometry, traffic volume and regulations, weather, and accident time – are analyzed using a dataset in high temporal (1 h) and spatial (250 m) resolution, covering the whole highway network of Austria over the period of four consecutive years. A combination of synthetic minority oversampling and maximum dissimilarity undersampling is used to balance the training dataset. We employ and compare a series of statistical learning techniques with respect to their predictive performance and discuss the importance of determining factors of accident occurrence from the ensemble of models. Findings substantiate that a trade-off between accuracy and sensitivity is inherent to imbalanced classification problems. Results show satisfying performance of tree-based methods which exhibit accuracies between 75% and 90% while exhibiting sensitivities between 30% and 50%. Overall, this analysis emphasizes the merits of using high-resolution data in the context of accident analysis.

1. Introduction

In spite of continuous improvements made in the area of vehicle safety technology throughout recent decades, motor-vehicle crashes continue to remain one of the leading causes of death and injury across the world, thus imposing a substantial human and economic toll to society (WHO, 2015). Consequently, considerable efforts have been made in the field of accident analysis, particularly as far as injury prevention and accident prediction modeling are concerned.

Traditionally, a vast majority of studies related to the assessment of accident occurrence is based on the application of various types of count-data regression models for modeling accident frequency (Yannis et al., 2017; Mannering and Bhat, 2014; Lord and Mannering, 2010). In these classical approaches, the number of accidents per predefined segment is counted over certain time periods, usually covering several years (Lord and Mannering, 2010). While this is well suited for assessing the impacts of covariates which remain constant over the time periods under consideration (e.g. properties related to road geometry), independent variables exhibiting high variation over time (e.g. traffic volume or weather) can only be considered in an aggregated way (e.g. annual average daily traffic, monthly or annual aggregates of weather data) which leads to loss of information related to specific accidents. Moreover, potentially important time-specific variables – like for instance weekday or accident time – cannot be included in count data regression models operating at an aggregated scale.

One way to overcome these shortcomings is to increase the data resolution both spatially and temporally. Instead of modeling the number of accidents on a given segment over extensive time periods, we therefore propose to consider this as a binary classification problem with the dichotomous outcome states “one or more accidents occurred” (`TRUE`) and “no accidents occurred” (`FALSE`) for short road segments within time periods of one hour. By assessing accident occurrence at such detailed granularity, important nuances in highly variable covariates can be captured, while variables that remain constant over time can still be included naturally.

Studies using high resolution real-time weather and traffic information have just emerged in recent years (e.g. Chen et al., 2018a; Wu et al., 2018; Theofilatos, 2017; Theofilatos et al., 2016; Yu et al., 2013; Abdel-Aty et al., 2007). While these precursor works are carried out at the example of selected freeways, comprehensive studies

* Corresponding author at: Institute of Applied Statistics and Scientific Computing, University of Natural Resources and Life Sciences, Vienna, Austria.
E-mail address: matthias.schloegl@boku.ac.at (M. Schlögl).

https://doi.org/10.1016/j.aap.2019.02.008
Received 10 October 2018; Received in revised form 30 January 2019; Accepted 7 February 2019
Available online 08 March 2019
0001-4575/

providing a country- or nationwide analysis of the whole highway network over an extended period of time are not known to the authors. Furthermore, comparative assessments of different methods and models are not the standard – in many cases, only certain specific methods are presented and discussed in other studies. As shown in the overview by (Mannering and Bhat, 2014), variants of generalized linear models dominate, while studies employing tree-based models or other statistical learning methods are rare.

Special emphasis is put on the assessment of a broad range of weather variables that may constitute potentially contributing factors to accident prevalence. While several efforts have been made to investigate the influence of various weather conditions – most notably rainfall – on road accidents (e.g. Omranian et al., 2018; Mais et al., 2016; Bergel-Hayat et al., 2013; Andrey, 2010; Bijleveld and Churchill, 2009; Koetse and Rietveld, 2009; Brijs et al., 2008; Eisenberg, 2004; Andrey et al., 2003; Edwards, 1996), many existing studies are somewhat limited by restrictions imposed by data availability. This study aims at broadening the perspective of previous studies by employing high-resolution data for ten weather variables obtained via meteorological re-analysis.

In addition, we seek to advocate an expansion of the methodological landscape for obtaining robust findings derived from a steadily growing amount of data in the area of accident research. Albeit algorithms for statistical learning are rapidly gaining popularity in many areas of research, their application in the context of accident analysis and modeling is still dominated by generalized linear models in terms of count data regression (Lord and Mannering, 2010; Maher and Summersgill, 1996). In this study, we apply and compare several state-of-the-art techniques for tackling the binary classification problem using an imbalanced high resolution dataset of road accidents in Austria.

2. Data

As by the end of the year 2016 (i.e. the end of the period under consideration), the Austrian freeway network (Fig. 1) consists of 18 freeways ('Autobahn') comprising a total length of 1719 km (BMVIT, 2017). In Austria, 'Autobahns' are controlled-access highways with distinct properties. They are grade separated, feature two structurally separated directional carriageways and comprise at least two lanes per direction. Austrian 'Schnellstraßen', which are also limited-access road similar to 'Autobahns' but with a lower design standard, are not considered in this study.

The whole network is used as a basis to build up a graph containing multiple features. The evaluation period spans four years from 2013-01-01 00:00 CET to 2017-01-01 00:00 CET (i.e. 35,040 h). The dataset is divided into a training dataset, which covers the years 2013 to 2015 (26,280 h), and a dedicated test dataset (8784 h in 2016) for evaluation purposes. In total, 45 variables are considered as possible predictors in this analysis.

2.1. Infrastructure

Data describing the road infrastructure have been gathered with various mobile measurement systems mounted on a rebuilt truck (RoadSTAR system). These measurement systems deliver high-resolution data on road condition and road surface characteristics, road tracing and geometry, as well as road environment (Maurer et al., 2002). Measurements are performed in a single run per lane 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 and differential GPS with a spatial resolution of 1 m, while the evenness is measured by different laser scanning systems with a resolution up to 5–20 mm. Skid resistance is measured using a modified Stuttgart skiddometer. Several camera systems provide information about surface damage and the road environment (e.g. traffic signs).

30 covariates are derived from these measurements (Table 1).

All variables are calculated as the mean across all lanes for each segment and for both directions separately.

2.2. Weather

High resolution weather data used in this study are derived from VERAflex re-analyses (Steinacker et al., 2011). The gridded datasets for various meteorological variables are available at a temporal resolution of one hour and a spatial grid of 250 m. Eight different meteorological indicators are considered in this study (Table 2).

In addition to real-time weather data, three climate indicators have been used. In order to represent the climatological conditions, annual precipitation totals (`rr_totals`), the number of frost days ($T_{min} < 0\,^{\circ}C$, `frost`) and hot days per year ($T_{max} > 30\,^{\circ}C$, `hot`), are derived from the E-OBS dataset (Haylock et al., 2008).

Weather data at tunnel sections have to be slightly modified for the purpose of representing the limited exposure to weather impacts. Thus, the variables precipitation, temperature and wind speed are corrected. Concerning precipitation, all precipitation events in tunnel sections are set to 0.1 mm/h to indicate a wet road surface in case of rain. The

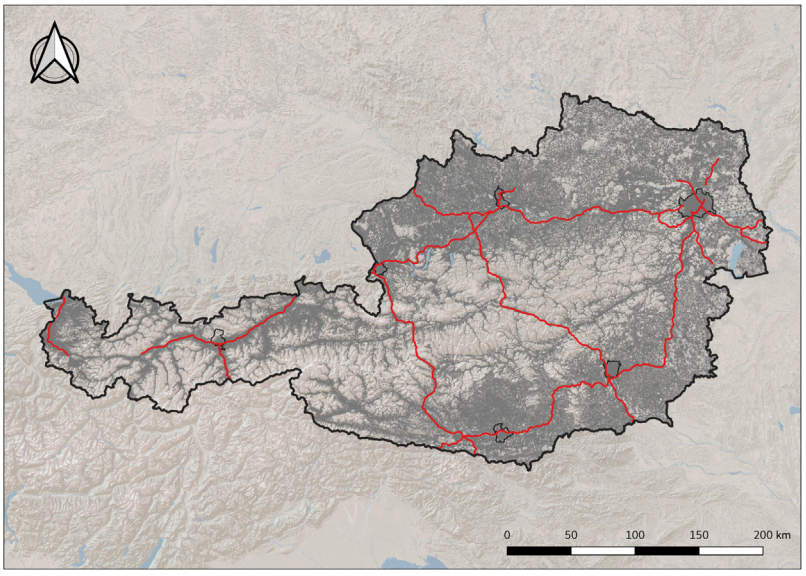

Fig. 1. Overview of the road network of Austria. Bold red lines indicate the highway network ('Autobahns'), which has been used in this study. Cities with more than 100,000 inhabitants are also depicted in the map for orientation purposes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Table 1
Description of variables derived from RoadSTAR measurements.

Variable	Data type	Description
w_tot	Metric	Total width of the road
w_bdl	Metric	Width of breakdown lane
is_bdl	Logical	Indicator for the existence of a breakdown lane
bdl_rel	Metric	Share of breakdown lane existence of total segment length
n_lanes	Factor	Number of lanes, reference class is 2
lane_ind	Factor	Variable indicating changes in the number of lanes within one segment; levels are none, exp (expansion) and red (reduction); reference class is none
q_med	Metric	Median of transverse gradient
q_zerocross	Logical	Indicator for a zero-cross (i.e. change of sign) of the transverse gradient
s_med	Metric	Median of longitudinal gradient
radius_med	Metric	Median of radius
curv_med	Metric	Median of curvature, which is the reciprocal of the radius
bendiness	Metric	Median of bendiness b, which is defined as the mean absolute course angle change, i.e. $b = \frac{1}{L}\sum_{i=1}^{L} \lvert\psi_i\rvert$, with L being the segment length and ψ_i indicating the respective change course angle
friction_q10	Metric	0.1 quantile of friction
friction_diff_q10	Metric	Difference between median and 0.1 quantile of friction, i.e. an indicator for changes in friction within one segment
iri	Metric	Median of international roughness index (Sayers and Karamihas, 1998)
wlp	Metric	Median of changes in the weighted longitudinal profile, i.e. an indicator for longitudinal evenness (Ueckermann and Steinauer, 2008)
wdepth_q90	Metric	Median of waterfilm thickness, i.e. an indicator for rut depth
wdepth_diff_q90	Metric	Difference between 0.9 quantile and median of waterfilm thickness, i.e. an indicator for changes of rut depth within a segment
surface	Factor	Material type of road surface, levels are asphalt, concrete and tms (thin membrane surface); reference class is asphalt
cracks_perc	Metric	Percentage of cracks on total road surface area
cracks_max	Metric	Maximum damage class of cracks
damage_perc	Metric	Percentage of road surface damage on total road surface area
damage_max	Metric	Maximum damage class of road surface damage
speed_limit_car	Factor	Speed limit for cars, levels are 130, 100, 80 and 60, reference class is 130
speed_limit_heavy	Factor	Speed limit for trucks and buses, levels are 80, 60 and 40, reference class is 80
no_overtaking	Logical	Overtaking ban
no_overtaking_heavy	Logical	Overtaking ban for trucks
event	Logical	Indicator for the existence of event sections, i.e. sections that feature acceleration and deceleration lanes (e.g. ramps, exits, motorway stations) or junctions
bridge	Logical	Indicator for the existence of bridges
tunnel	Factor	Indicator for the existence of tunnels (longer than 50 m), levels are 0 (no tunnel), 1 (tunnel) and 2 (tunnel portal area); reference class is 0

Table 2
Description of real-time weather variables.

Variable	Meteorological indicator	Unit
T	Air temperature	[°C]
RR	Accumulated precipitation	[mm]
RR_SA	Percentage of solid precipitation	[%]
RF	Relative humidity of the air	[%]
P	Air pressure	[hPa]
FFX	Wind gusts (maximum wind speed)	[m/s]
DD	Wind direction	[°]
ASD	Absolute sunshine duration	[min]

percentage of solid precipitation (RR_SA) is set to 0. Regarding temperature, temperature is set to 0.1 °C if the air temperature is below zero and truncated to the next integer towards zero otherwise. Wind gusts are set to 2 m/s constantly.

2.3. Traffic volume

Traffic volume is obtained from the traffic management and information system at Austrian highways, 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 average daily traffic estimates, including heavy good vehicles (ASFINAG, 2018). In order to obtain more robust results, annual average daily traffic (AADT) numbers (including proportion of heavy goods vehicles) are used at a highway section level (i.e. sections between two junctions). While AADT is an aggregated value rather than a temporally stratified variable, interactions between time, weekday and AADT are considered to be capable of capturing hourly traffic flow to some extent.

2.4. Accidents

Accident data (on road accidents with personal injuries) are derived from the official annual Austrian crash statistics, which are provided by the national bureau of statistics (Statistics Austria) based on police crash reports (Statistik Austria, 2018).

2.5. Dataset structure

The road graph for the whole highway network was constructed based on the official chainage according to the Austrian highway section register. Road infrastructure data were joined to this graph using the geometry data collected with IMU and differential GPS. The resulting highway graph is available at a spatial resolution of 1 m. In a next step, these 1 m-data were aggregated into sections of equal length. In light of the pitfalls related to segment definition (Schlögl and Stütz, 2017; Shankar et al., 1995), a sensitivity analysis was run in order to empirically determine the most adequate segment length. While short segments lead increase uncertainty regarding accident localization and a preponderance of zeros, large segment lengths tend to water down site-specific effects. Different segment lengths ranging between 150 and 1000 m were therefore considered in determining the optimal spatial resolution. In order to obtain robust segmentations of the whole network across this range of segment lengths, different segment starting points were considered in the in the sensitivity analysis by shifting the starting points of each highway by 25 m. The thereby created datasets featuring various intervals have been used as a basis for a preliminary accident analysis using straightforward count data models. All of the aforementioned data sources have been joined to the different datasets, using spatial aggregates according to the segment lengths and a temporal aggregate across all four years. The number of accidents on each segment has then been analyzed by using negative binomial regression

models. Goodness of fit, significance of the covariates and the variance of model coefficients was analyzed. Based on this analysis, a segment length (spatial resolution) of 250 m was chosen, since this turned out to be the sweet spot where segments feature a certain robust length with respect to accident localization, while local effects do not get diluted. In addition, a segment length of 250 m does also correspond to the spatial resolution of the gridded weather datasets. The resulting segmentation of the Austrian road network results in 13,466 segments (i.e. 6733 segments per direction) with a length of 250 m. Residual distances at the end of each highway have been added to the last segment if the leftover was smaller than 50 m, added as a shorter segment if the leftover was larger than 200 m, and dropped in all other cases. Infrastructure variables are featured as the mean value across all lanes, aggregated as shown in Table 1. Weather data have been extracted from the gridded datasets at a spatial resolution of 125 m (i.e. at the beginning, mid and end of each segment). The mean value of these three locations was then assigned to each road segment. In case of traffic volume, the maximum value was used for each segment, representing maximum traffic exposure. The binary target variable `acc` indicates the occurrence of an accident (i.e. usually one, in extremely rare cases more than one) on a certain segment at a specific hour. The accidents were assigned to the respective interval based on their accident location (highway kilometer) and accident timestamp. Eventually, a factor variable `h_cat` related to the time of the day each hour belongs to was derived from the timestamp, corresponding to homogeneous classes with respect to traffic exposure and user groups. The five levels of this factor variable are `night` (19:00–05:00), `morning` (06:00–09:00), `noon` (10:00–12:00), `afternoon` (13:00–15:00) and `evening` (16:00–18:00). These levels represent characteristic periods of traffic in the daily cycle with distinct user profiles. In addition, the variable `weekday` (a factor variable with 7 levels) contains a weekday classification based on the Austrian guideline for traffic counts (FSV, 2015), which establishes a comparable characterization of weekdays by taking (movable) feasts into account.

Overall, the final datasets feature 353,886,480 rows (training/validation set) and 118,298,810 rows (test set) respectively, as well as 50 columns. The resulting **R** data table (Dowle and Srinivasan, 2018) has a total size of 145.1 GB in memory, thus being in the transition zone to big data analysis.

2.6. Balancing the training dataset

Due to the high temporal and spatial resolution, accidents and non-accidents are highly imbalanced in the dataset. The training dataset comprises 4438 event sections in contrast to 353,882,042 non-event sections, resulting in class imbalance equivalent to an order of 1 to 80,000. The issue of class imbalance, which is in many cases – like the present one – intrinsic to the dataset, has been widely known to heavily compromise the application of methods for statistical learning, since the prevalence of the majority class overshades rare events in the minority class (Cateni et al., 2014; Kuhn and Johnson, 2013; He and Garcia, 2009; Japkowicz and Stephen, 2002).

Several resampling approaches have been proposed in recent years to alleviate this issue (Sáez et al., 2016). A widely applied and well-known supersampling algorithm is the Synthetic Minority Over-sampling Technique (SMOTE) developed by Chawla et al. (2002). SMOTE is a minority over-sampling technique that takes each minority class sample and introduces synthetic instances of the minority class using k nearest neighbors within a bootstrapping approach. Recent applications of this approach in the context of road safety analysis include studies by Basso et al. (2018), Yuan et al. (2019) and Katrakazas et al. (2019).

Following the line of Chawla et al. (2002) a combination of SMOTE and undersampling has been applied in this study in order to give a larger presence to the minority class, thereby creating a more balanced training dataset. In this context, undersampling refers to non-events, while oversampling refers to events (accidents). The oversampling rate denotes to the share of original events that is synthetically added to the data set, while the undersampling rate is defined as percentage of the number of newly generated minority class instances that is added to the data set.

Parameter tuning for SMOTE was accomplished by testing different percentages for SMOTE-oversampling and random undersampling, using oversampling rates ranging from 100% to 500% and undersampling rates in the same range as well. Cross-validation results of the 25 SMOTEd data sets were then used as a basis to select a reasonable parameter combination for SMOTE.

In terms of the undersampling rate, no substantial differences could be observed when increasing the share of minority class instances past 25%, that is, an event to non-event ratio of 1:4. Regarding the oversampling rate, it has to be noted that even though minority class instances are very rare in terms of relative occurrence, more than 4000 accidents entail that there still is a solid number of absolute events present in the dataset. Also, an overly expansive use of SMOTE for creating synthetic samples is detrimental, since the oversampled dataset may contain events that somehow deviate from the actual minority class data.

Based on the results of the cross-validation as well as these additional considerations, an oversampling rate of 300% (16,728 events in the minority class) and an undersampling rate of 400% (50,184 instances of the majority class), resulting in an event rate of 1/3, were chosen for creating the SMOTEd dataset for further analysis.

In addition to the approach proposed by Chawla et al. (2002), we combined SMOTE with stratified maximum dissimilarity sampling (Willett, 1999) instead of simple random sampling in order to achieve a better coverage of the full data space. The stratified maximum dissimilarity sampling was conducted separately for 8760 groups of three consecutive hours of the training dataset by first selecting a random observation from the 44,000 observations of these three hours, which was then used to iteratively add the five most dissimilar observations. This yields 52,560 instances of the majority class, which is a close approximation of the above mentioned optimal undersampling rate. Groups of three hours were chosen, since this setup constitutes a reasonable trade-off between computation time and effectiveness of the maximum dissimilarity sampling. While the use of larger strata (months, weeks, days) was tested, the computational demands in constructing distance/similarity matrices for these setups turned out to be prohibitive for a practical application. The resulting selection yields a structurally diverse set which provides a reasonable representation of the full dataset covering the three years of the training period.

Another option taken to tackle the problem of class imbalance is to adjust the classification according to predicted class probabilities. Usually, predictions with class probabilities $p \geq 0.5$ are classified as `TRUE`. Using a cutoff other than 0.5 has been proposed to reflect unequal misclassification costs (Kuhn and Johnson, 2013). Based on the predicted probabilities, we have tested different thresholds for classification by gradually lowering the decision threshold in steps of 0.1 down to $p = 0.1$.

3. Methods

3.1. Logistic regression

Binary logistic regression models are a long-established method of estimating the probability of a dichotomous response variable based on one ore more covariates (Cox, 1958). One of the biggest advantages of logistic regression for classification of dichotomous response variables is the interpretability of the results, since conclusions about whether or not the presence of a risk factor has an influence on the probability of a given outcome are straightforward to draw. The use of downsampling approaches is not advisable when applying logistic regression, as this will only lower the precision of the estimated odds ratios. However, problems that may occur in this context are biased estimation of

regression coefficients and erroneous estimation of intercepts as well as a phenomenon known as separation, which occurs if the likelihood converges while at least one parameter estimate diverges to ± infinity (Heinze and Schemper, 2002).

In order to overcome these issues, a penalization method known as Firth's penalized likelihood (Firth, 1993; Heinze and Schemper, 2002) has been proposed as a viable option for estimating logistic regression models for unbalanced data sets. The approach penalizes the likelihood by the Jeffreys invariant prior, a non-informative prior distribution that is proportional to the square root of the determinant of the Fisher information matrix. The Firth penalty has been shown to result in good bias and MSE properties for regression coefficients, it solves the problem of separation, is always convergent (i.e. by producing finite, consistent estimates of regression parameters) and rather performant (Heinze, 2006; Heinze and Schemper, 2002). It is thus preferred over exact logistic regression, which is computationally intensive.

In addition, given the large-scale learning problem at hand, a computationally efficient variant of logistic regression that uses a stochastic gradient descent approach to optimize logistic loss was performed (Shalev-Shwartz et al., 2011).

In this study, we have used the GLM implementation in the **R** package `stats` for conducting logistic regression, as well the implementation of bidirectional elimination based on AIC for stepwise regression from the `MASS` package (Venables and Ripley, 2002). In addition to maximum likelihood estimation, bias reduced binomial-response GLMs employing the Firth penalty were estimated using `brglm` (Kosmidis, 2017). Logistic regression with Pegasos updates was performed by means of `sofia-ml`, a suite of fast incremental algorithms for machine learning by Sculley (2009), using the wrapper-functions from the `RSofia` package in **R**.

3.2. Binary quantile regression

While conventional linear regression relies on least squares for estimating the mean of the response variable conditional on the values of the independent variables, the idea behind quantile regression is to estimate conditional quantiles (e.g. the median) of the response variable. Even though quantile regression is not an obvious choice for modeling a dichotomous response variable (since a binary dependent variable does not yield continuous quantiles), efforts have been undertaken to explore the potential benefits of quantile regression for settings with a binary response as well, including both frequentist approaches – based on maximum score estimator introduced by Manski (1975) – and Bayesian approaches (Benoit and Van den Poel, 2012).

In this study we follow the line of Benoit and Van den Poel (2012) and conduct a Bayesian variant of binary quantile regression based on the asymmetric Laplace distribution, additionally employing adaptive lasso variable selection as described in Benoit et al. (2013). We make use of the **R** implementation by Benoit and Van den Poel (2017), which is available through the `bayesQR` package.

3.3. Multivariate adaptive regression splines

Multivariate adaptive regression splines (MARS) is a flexible, non-parametric regression technique for modeling high-dimensional data (Friedman, 1991). It can be viewed as a combination of recursive partitioning, stepwise linear regression and spline fitting that is capable of automatically modeling nonlinearities and interactions between variables. The key part of MARS models is the usage of hinge functions as basis functions, whose weighted sums constitute the final model.

The algorithm can also be used to handle classification problems with a logical response by treating the problem as a regression (Hastie et al., 2009). Applying a MARS-GLM with logit link to a binary response allows for estimating response probabilities.

Three variable importance measures are available for MARS. First, the number of model subsets that include the variable are reported. Second, decreases in the generalized cross-validation (GCV) value are considered. The GCV criterion is obtained in a three-step-procedure by (i) calculating the decrease in GCV for each subset relative to the previous subset during the pruning pass, (ii) summing the decreases over all subsets that include the variable and (iii) scaling the result so that the largest summed decrease is 100 (Milborrow, 2018). Third, the same procedure as for the GCV criterion can be applied using the residual sum of squares (RSS) instead of the GCV.

The **R** implementation of MARS is available in a package called `earth` (Milborrow, 2017).

3.4. Random forest

Random forests (Breiman, 2001) are ensembles of de-correlated decision trees. The algorithm extends the idea of bootstrap aggregating (bagging) by using a random selection of features to determine the best variable/split-point within the process of growing trees to the bootstrapped samples from the training data (*feature bagging*). When random forests are used for classification problems, the resulting classification is based on the majority vote derived from all class votes from each tree (Hastie et al., 2009).

Feature importance is reported as permutation accuracy importance (also referred to as mean decrease accuracy) for random forest and totally randomized trees. It is based on the difference in prediction accuracy before and after replacing single predictor variables with random noise, which is drawn from the same distribution as original feature values (Strobl et al., 2007; Breiman, 2001). A straightforward way to achieve this is to simply shuffle the values for a feature, thereby using other instances feature values.

The random forest implementation of the **R** package `ranger` (Wright and Ziegler, 2017) is used in this study. It includes an implementation of probability forests for estimating individual probabilities for binary responses according to Malley et al. (2012), where the forest probability estimate is obtained as the average of all probability estimates for each single tree.

3.5. Extremely randomized trees

Introducing an additional randomization step into the random forest procedure yields extremely randomized trees. In this algorithm, values for the cut-points are selected fully at random – i.e., independently of the target variable – instead of being derived as the locally optimal feature/split combination. In the most extreme case this algorithm builds *totally randomized trees* by randomly selecting a *single* attribute and cut-point at each node (Geurts et al., 2006).

Extremely randomized trees were also fitted using the `ranger` package in **R**.

3.6. Model-based boosting

The basic principle of boosting is to combine the output of many 'weak' classifiers in order to create a strong learner by iteratively improving the already trained ensemble (Hastie et al., 2009). This is usually done in a stage-wise fashion by adding an estimator to the residuals of the respective predecessor models. We concur with Friedman et al. (2000) as well as Bühlmann and Hothorn (2007), who pointed out the benefits of minimizing the negative binomial log-likelihood to estimate the probability parameter of a binary response. Therefore, we decided to give preference to BinomialBoosting – which is in fact closely related to LogitBoost – instead of e.g. the classical AdaBoost algorithm described by Freund and Schapire (1997), which employs an exponential loss function. As far as trees are concerned, xgboost is given preference over applying LogitBoost to conditional inference trees (Hothorn et al., 2006). Tuning of the number of iterations for obtaining the final boosting estimate was done through 25-fold boostrap cross-validation.

Model boosting is performed for generalized linear models (Schmid and Hothorn, 2008), using the implementation available through `mboost` by Hofner et al. (2014).

3.7. XGBoost

Extreme Gradient Boosting (XGBoost) is a variant of the gradient boosting machine (also known as gradient tree boosting or gradient boosted regression trees) described by Friedman (2002, 2001) and (Mason et al., 1999). Gradient boosting is also based on combining an ensemble of weak learners into a single strong model by iteratively improving the ensemble learner. By employing the concept of gradient descent, gradient boosting allows to enable the optimization of a more complex (in fact any differentiable) loss function. Thus, instead of training new learners on the residuals of the previous model, gradient boosting relies on training new models to the gradient of the loss function. Due to a number of optimizations (most notably the use of second-order gradient descent), XGBoost is a very fast, efficient and accurate tree boosting algorithm (Chen and Guestrin, 2016).

Feature importance for the xgboost model is assessed through the relative contribution of the corresponding feature to the model, referred to as gain. For the boosted tree model used in this study, each gain of each feature of each tree is taken into account, then average per feature to give a vision of the entire model (Chen et al., 2018; Chen and Guestrin, 2016).

The **R** interface to XGBoost, available through package `xgboost` by T. Chen et al. (2018) is used. Hyperparameter-tuning was performed through model-based optimization using the package `mlrMBO` (Bischl et al., 2017). The Kriging-implementation of the package `DiceKriging` is used as a surrogate model for optimizing hyperparameters (Roustant et al., 2012).

3.8. Support vector machine

In its original form, a support vector machine (SVM) is a non-probabilistic binary linear classifier that can be used to solve a classification problem by constructing hyperplanes in a way that the resulting gaps between classes exhibit margins that are as large as possible (Cristianini and Shawe-Taylor, 2000; Vapnik, 2000, 1998). While SVMs are linear learning algorithms, they can also be used to perform non-linear classification by obtaining a decision boundary that is linear in a high-dimensional space (Kernel trick). In this study we use Pegasos, a primal estimated sub-gradient descent algorithm for SVM that is particularly suited for learning from large datasets (Shalev-Shwartz et al., 2011). The Pegasos implementations for support vector machines available through the **R** packages `Rsofia` and `ssvm` were used.

3.9. Bayesian regularized neural network

Artificial neural networks are directed, weighted graphs consisting of a network of artificial neurons. The output of each neuron within the network depends on the input and activation (i.e. the change of the internal state of a neuron). A two layer neural network, whose network parameters are estimated through employing a Bayesian approach – thus performing shrinkage toward some prior distribution – is used in this study (Makridakis et al., 2018; Burden and Winkler, 2009). The Bayesian penalization method follows suggestions provided by (MacKay, 1992; Foresee and Hagan, 1997). Initial weights are assigned according to the Nguyen and Widrow algorithm (Nguyen and Widrow, 1990), and optimization is performed using the Gauss–Newton algorithm. The Bayesian regularized neural network is fitted using the **R** package `brnn` (Pérez-Rodríguez et al., 2013).

3.10. Model quality assessment

Several commonly used performance measures are used to assess the quality of the fitted models. All of them are derived from the confusion matrix, a two-dimensional contingency table that displays the agreement between predicted and actual class of the test dataset. The confusion matrix features counts for correctly predicted outcomes (true positive and true negative) as well as incorrectly predicted outcomes (false positive and false negative, respectively). A number of different performance metrics can be obtained from this error matrix, including the following:

- **Accuracy** indicates the ability of a binary classification test to correctly identify or exclude an outcome. It is defined as the proportion of correct predictions among all predictions.
- **Sensitivity**, also known as true positive rate (TPR) or recall, is defined as the proportion of actual true positives among all positive predictions, i.e. the proportion of correctly classified positives. Since the main focus is of course the correct classification of the rare instances of the accident class, sensitivity is a particularly important indicator of classifier performance in this case.
- **Specificity**, also called true negative rate (TNR), analogously is the proportion of actual negatives among all negative predictions, i.e. the proportion of correctly identified negatives.
- The false positive rate (**FPR**) is the proportion of false positives among all negative predictions, thereby representing the percentage of 'false alarms'.
- **AUC** denotes the area under the receiver operating characteristic (ROC) curve. The ROC curve is obtained by plotting the true positive rate (ordinate) against the false positive rate (abscissa) across different discrimination thresholds. The AUC measures discrimination, i.e. the ability of the predictor to correctly classify an outcome. Thus, AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

Results as presented in the next section are based on an evaluation of model quality by assessing the performance of classifiers fitted on the training data (2013–2015) on the – hitherto unseen – test data of the holdout (2016). Standard statistical performance metrics for binary classification tests are used to assess model quality.

3.11. Hardware specifications

All data processing and model fitting was done in **R** using two servers running Ubuntu 16.04 LTS on an Intel® Xeon® CPU E5-2667 v2 @ 3.30 GHz with 48 cores and 320 GB RAM and an Intel® Xeon® CPU E5-2687W v4 @ 3.00 GHz with 32 cores and 128 GB RAM.

4. Results

4.1. Model performance

Model performance is evaluated for the two different undersampling strategies employed. The classical approach proposed by Chawla et al. (2002), which uses simple random undersampling, serves as a baseline.

The use of SMOTE and random undersampling shows underwhelming results (Table 3). Using stepwise logistic regression, only

Table 3
Overview of selected scalar classification performance metrics for hourly data using SMOTE with random undersampling as training data.

Method	Accuracy	Sensitivity	Specificity	FPR	AUC
Naive	1	0	1	0	0.50
Stepwise logistic regression	0.94	0.18	0.94	0.06	0.66
MARS	0.99	0.04	0.99	0.01	0.65
Random forest	0.99	0.06	0.99	0.01	0.65

Table 4
Overview of selected scalar classification performance metrics for hourly data using SMOTE with maximum dissimilarity sampling as training data.

Method	Threshold	Accuracy	Sensitivity	Specificity	FPR	AUC
Naive	–	1	0	1	0	0.50
Stepwise logistic regression	0.5	0.61	0.54	0.61	0.39	0.60
logit w/ Firth penalty	0.5	0.60	0.54	0.60	0.40	0.60
logit w/ Pegasos projection	0.5	0.33	0.81	0.33	0.67	0.61
Bayesian quantile regression	0.5	0.33	0.76	0.33	0.67	0.56
BinomialBoost GLM	0.5	0.60	0.54	0.60	0.40	0.60
MARS	0.5	0.81	0.30	0.81	0.19	0.60
Random forest	0.5	0.85	0.32	0.85	0.15	0.65
Extremely randomized trees	0.5	0.78	0.40	0.78	0.22	0.65
Totally randomized trees	0.5	0.75	0.46	0.75	0.25	0.65
xgboost	0.5	0.87	0.27	0.87	0.13	0.62
BRNN	0.5	0.65	0.52	0.65	0.35	0.61
Pegasos SVM	0.5	0.55	0.60	0.55	0.45	0.60
Stepwise logistic regression	0.4	0.48	0.64	0.48	0.52	0.60
logit w/ Firth penalty	0.4	0.48	0.64	0.48	0.52	0.60
logit w/ Pegasos projection	0.4	0.18	0.90	0.18	0.82	0.61
Bayesian quantile regression	0.4	0.28	0.78	0.28	0.45	0.53
BinomialBoost GLM	0.4	0.48	0.65	0.48	0.52	0.60
MARS	0.4	0.72	0.41	0.72	0.28	0.60
Random forest	0.4	0.68	0.54	0.68	0.32	0.65
Extremely randomized trees	0.4	0.59	0.63	0.59	0.41	0.65
Totally randomized trees	0.4	0.56	0.66	0.56	0.52	0.65
xgboost	0.4	0.81	0.36	0.81	0.19	0.64
BRNN	0.4	0.48	0.66	0.48	0.52	0.61

about 18% of road accidents in the test dataset are classified correctly as such, given an overall accuracy of around 94%. Albeit both MARS and random forest perform better in terms of overall accuracy (by classifying more cases correctly as false), this is at the expense of sensitivity, which is as low as approximately 5%.

Results using a combination of SMOTE and maximum dissimilarity undersampling as training data show clearly better results in terms of the true positive rate at the expense of overall accuracy (Table 4). Our analyses show that the performance of statistical learning models employed to model the dichotomous outcome state is a trade-off between overall accuracy and sensitivity. As far as accuracy in the standard case (threshold = 0.5) is concerned, xgboost yields the best results (87% correctly classified observations), closely followed by random forest (85%). However, these two classifiers show a sensitivity of around 30%, indicating that about a third of all accident instances is predicted as such. Classifiers exhibiting the highest sensitivity (including all four types of logistic regression, the BRNN and the Pegasos SVM) show very low accuracy values of below 60%. Consequently, these models yield a high number of false positives. This issue becomes even more pronounced when lowering the threshold for classification according to class probabilities. Already at a threshold of 0.4, accuracy drops notably in most cases. While a majority of classifiers show an increase in sensitivity of about 10%, this is offset by an accuracy loss to the same extent. Only xgboost seems to benefit from this procedure, gaining an additional 10% sensitivity while showing an accuracy drop of only 5%. However, the golden mean seems to be provided by totally randomized trees, which achieve a sensitivity of more than 45% while still featuring a decent accuracy of just above 75%.

A closer look at the predicted class probabilities provides interesting insights into the underlying prediction data basis that is eventually reclassified into a dichotomous prediction (Fig. 2). The desired outcome of predicted class probability would be a bimodal distribution displaying the two classes of the dichotomous response. Given an ideal predictor, the first, large peak of the distribution is close to 0, and a second, minor peak close to 1. Results show a consistent picture between related methods. All types of logistic regression are rather uniformly distributed across the whole range, with a slight decrease towards 1. Logistic regression with Pegasos projection differs from this general pattern and is shifted to the right. Random forest and the two types of extremely randomized trees show a bulk of highest density at predicted probabilities between 0.15 and 0.60. xgboost shows a distinct peak close to 0, followed by an exponentially decreasing density towards 1. Finally, BRNN exhibit a totally different behavior, with a small peak around 0, a steep increase at 0.25 followed by a continuous decrease towards 1.

Class discrimination is ambiguous across all models (Fig. 3). While observed accidents do feature higher probabilities of being actually predicted as accident in every classifier, distributions of predicted probabilities are vastly overlapping. This is also illustrated by very shallow discrimination slopes (Table 5), i.e. the slopes of a simple linear regression of predicted probabilities of accidents on the binary accident status (Pencina et al., 2017).

4.2. Variable importance

Albeit variable importance varies across the different models, certain patterns emerge. These are visible both in single feature importance rankings for each variable (Fig. 4) as well as the mean feature importance rank considering the full ensemble of models (Fig. 5).

Findings across all models exhibit the importance of certain groups of variables:

- Traffic volume (`aadt`) is among the top five variables in all models. The volume of heavy goods vehicles (`hgv`) is of lesser importance, still achieving top ranks in random forest and xgboost.
- Findings related to road geometry and road condition are less consistent. The existence of bridges (`bridge`), the existence of a zero-cross of the transversal gradient (`q_zerocross`) and maximum damage of the road surface (`damage_max`) prevail as important features across all models. Other features that exhibit higher importance at least in several models include surface roughness (`iri`), the number of lanes (`n_lanes`), the width of the breakdown lane (`w_bdl`) and the existence of event sections (`event`). In certain cases, longitudinal evenness (`wlp`) and surface type (`surface`) are amongst the most important features as well. Features related to curvature (`curv_med`), radius (`radius_med`), bendiness

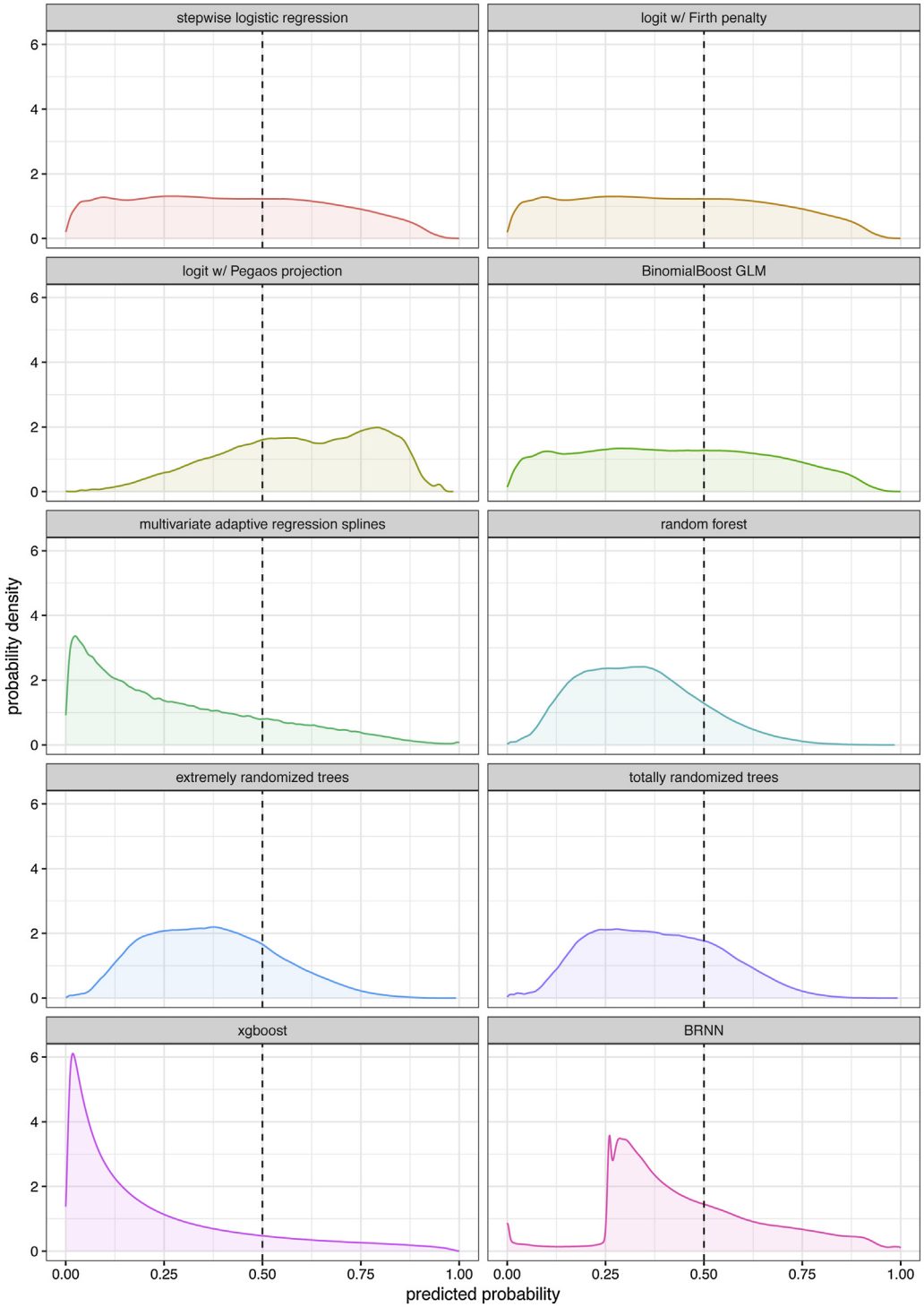

Fig. 2. Density plot of predicted class probabilities for class "accident" for hourly data using SMOTE with maximum dissimilarity sampling as training data. The dashed vertical line indicates the class discrimination threshold, i.e. the probability at which the positive class ("accident") is chosen over the negative class ("no accident").

(`bendiness`) as well as friction (`mu`) are of less importance.

- No overtaking and speed limits: restrictions on overtaking are important, particularly if overtaking is prohibited for heavy goods vehicles (`no_overtaking_heavy`). Even though speed limits are less important, the same applies for these features: speed limits for heavy good vehicles (`speed_imit_heavy`) more important than speed limits for cars (`speed_limit_car`).
- Information related to the time of the day (`h_cat`) as well as the weekday (`weekday`) do not seem to be of major importance. With some exceptions these are usually located towards the end of the variable importance scale.
- Throughout all models, climate data (i.e. long-term averages related to both temperature and precipitation) seem to be important, while most of high resolution weather data are located at the end of the importance ranking across all models.

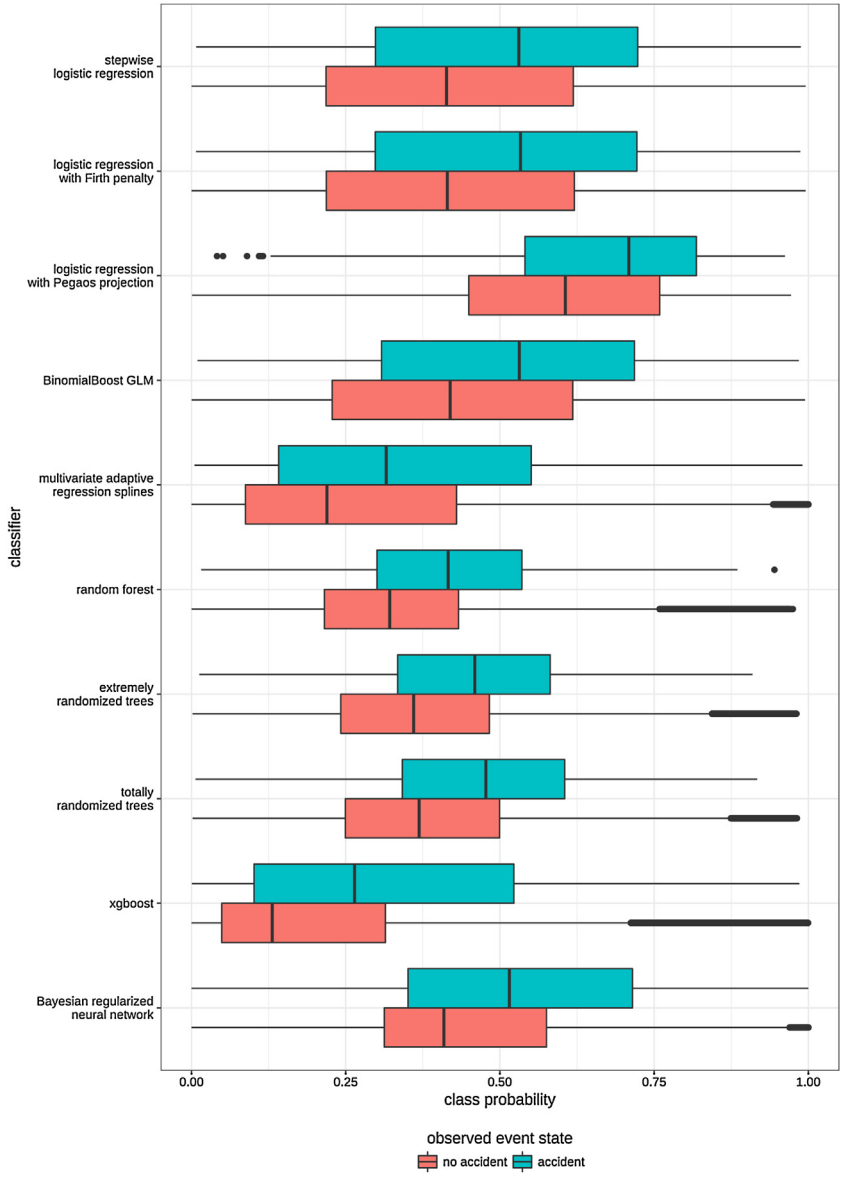

Fig. 3. Boxplots of predicted class probabilities for class "accident" for hourly data using SMOTE with maximum dissimilarity sampling as training data grouped by model and observed outcome.

Table 5
Resulting discrimination slopes, i.e. the slopes of a simple linear regression of predicted probabilities of accidents on the dichotomous accident status.

Model	Discrimination slope
Stepwise logistic regression	0.0836
logit w/ Firth correction	0.0842
logit w/ Pegasos projection	0.0743
Bayesian quantile regression	0.0797
BinomialBoost GLM	0.0825
Multivariate adaptive regression spline	0.0778
Random forest	0.0881
Extremely randomized trees	0.0893
Totally randomized trees	0.0939
xgboost	0.1185
BRNN	0.0818

5. Discussion

5.1. Methodology

Findings show that tree-based ensemble methods seem to outperform more classical approaches such as logistic regression. Both bagging (i.e. random forests and extremely randomized trees) as well as boosting methods (xgboost) show remarkably good results given the difficult classification task at hand.

To some extent, results might indicate that there is not enough variation of the response between the levels of categorical predictors or within the range of continuous predictor variables to unambiguously predict the outcome of the target variable. This entails that it is cumbersome to infer marginal effects of some variables to the outcome. That being said this is hardly the case in any properly conducted accident data analysis (i.e. using separate training, validation and test datasets to avoid overfitting and biased results) due to the rare event nature of accidents and the large variability in possible accident causes and accident-contributing factors. In this context it is important to emphasize that contradictory findings are often reported in the field of

Fig. 4. Overview over feature importance ranks across all models. Barplots display the number of models ("count") containing the respective feature within each binned rank class. For reasons of clarity, variable ranks are aggregated into 5 classes. The last class comprises all variable ranks between 25 and 45, thus being the largest class.

accident research (Theofilatos, 2017; Theofilatos and Yannis, 2014). This is most likely attributable to high uncertainty and large variance inherent to many accident prediction models, which can hardly be avoided due to small sample sizes.

A common drawback that is common to basically all accident frequency and accident occurrence models is that one misses potentially important predictors. Certain features related to the person-level (and not the section level) – including for instance distraction or alcoholization – cannot straightforwardly be incorporated in this type accident frequency analysis.

As illustrated in detail in Schlögl and Stütz (2017), all features that may be considered as potentially relevant input data for such models are subject to considerable uncertainties. This is especially the case for time and location of the accidents in such a high-resolution setting. Since these uncertainties are related to the target variable, any errors in segment or timeslot assignment due to misspecifications of accurate accident time and location may lead to potentially considerably different covariates.

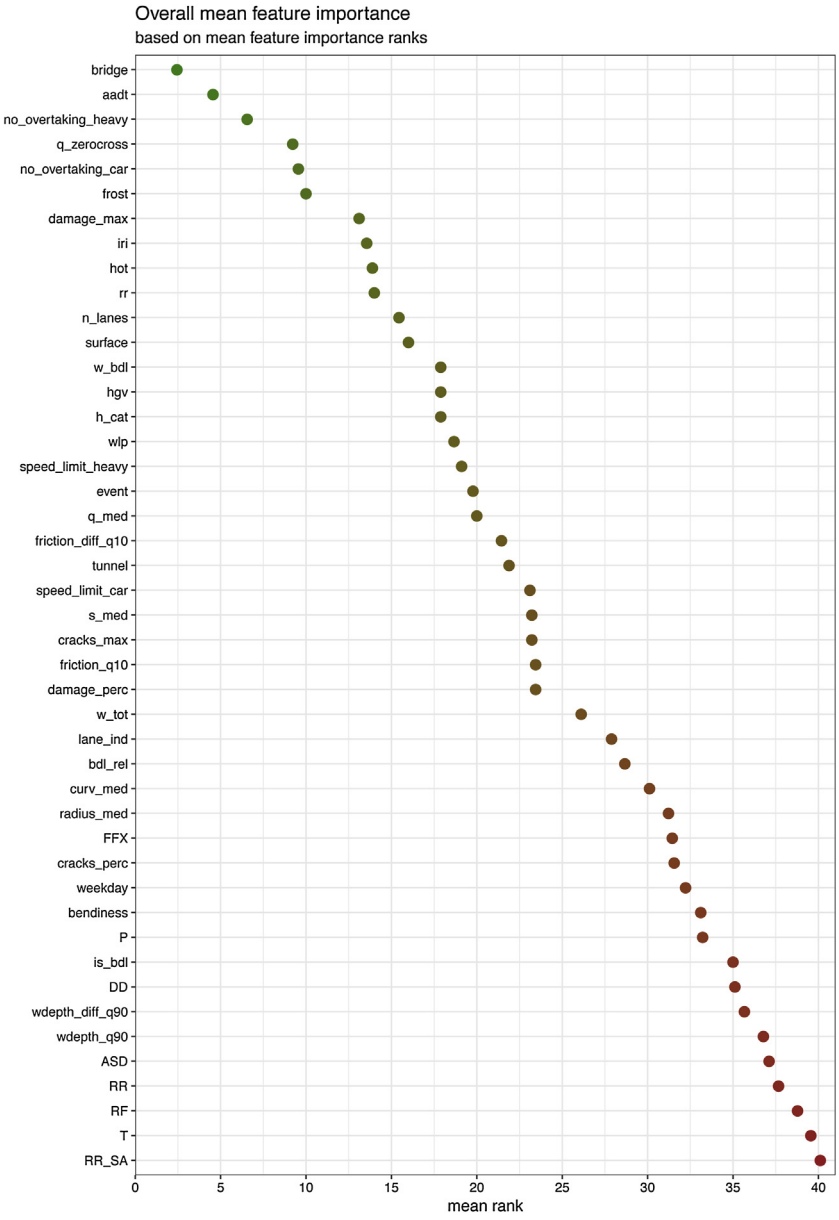

Fig. 5. Overview of feature importances across all models. Results are based on the mean importance rank of each variable in the respective models.

Albeit SMOTE is a commonly recommended approach to tackle the problem of class imbalance in machine learning models, this might not necessarily be the best solution to approach this problem. While we have shown that replacing random undersampling with maximum dissimilarity sampling can be used to improve the coverage of the feature space, thus leading to potentially better performance of any model applied on this dataset, there are several other options to cope with the problem of class imbalance that might even further improve results. Sophisticated methods that might be used instead to derive a training dataset include the use of generative adversarial networks (Goodfellow et al., 2014) or the nested sampling algorithms MultiNest (Buchner et al., 2014; Feroz et al., 2009) and PolyChord (Handley et al., 2015a,b).

Concerning the assessment method, efforts have been made to validate the models based on independent data, but the results might be sensitive to the (short) observation period. All models were trained and validated on data for the three years 2013 to 2015, and tested on the dataset for 2016. This approach seems feasible from a theoretical point of view, since a full year can be used as a test dataset. However, it has to be noted that these results are not perfectly robust. Arguably, different partitions might lead to different results. Albeit, the approach taken in this study entails that it is computationally prohibitive to try a sufficiently large number of training samples, bagging classifiers induced over balanced bootstrap training samples is an promising alternative approach that reduce variances in the predictive performances of the models (Wallace et al., 2011). This is explored further in the companion paper by Schlögl (2018).

5.2. General approach

Against the background of the highly imbalanced nature of the datasets, which makes the analysis of such datasets challenging and cumbersome, it can be concluded that model quality assessment is a trade-off between accuracy, sensitivity, and false positive rate. Even though the underlying dataset used in this study is severely imbalanced, overall model performance is perfectly comparable to similar unbalanced classification problems in other domains that have a more favorable event to non-event ratio (cf. Kuhn and Johnson, 2013).

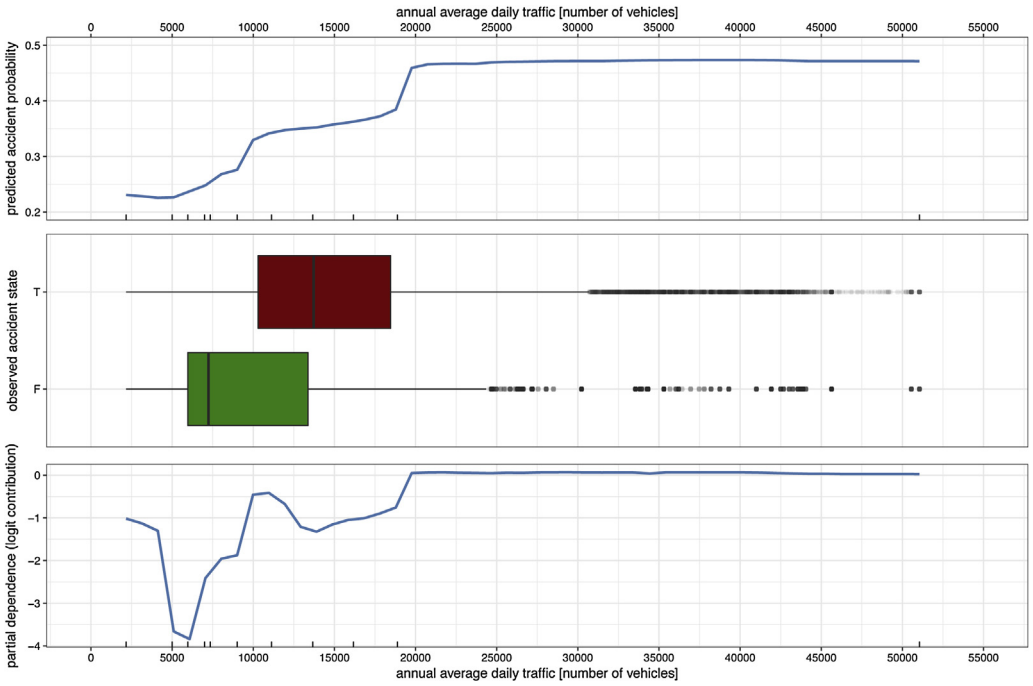

Fig. 6. Relationship between traffic volume and observed accident state displayed as (a) partial dependence of accident probability on traffic volume, (b) empirical distributions of traffic volume grouped by observed accident state in the training data set, and (c) partial dependence displayed as logit contribution of traffic volume on class probability from the perspective of the model. In the last plot, negative values of partial dependence indicate that the positive class is less likely for the corresponding value of the independent variable. Rugs in the partial dependence plot indicate the minimum and maximum values as well as the deciles of the predictor distribution. Both partial dependence plots are based on the random forest model.

Slightly lower AUC values compared to those reported in other studies on hourly crash data analysis (e.g. Chen et al., 2018a) are attributable to the different underlying settings and approaches. In some studies, it is not clear whether the AUC was mistakenly calculated on the training data set (consequently depicting results of severe overfitting) or correctly on a dedicated holdout. Hence, such results are not directly comparable to the findings presented in this study.

Among the variety of models employed in this study, tree-based models yield the best results. xgboost shines in terms of both overall accuracy as well as false positive rate, and shows the best discrimination between the two classes. In addition, xgboost benefits most from lowering the threshold of classifying as `TRUE`. Random forests perform similar to xgboost, featuring similar accuracy and false positive rate, and even slightly higher sensitivity. Totally randomized trees seem to perform best in terms of hitting the sweet spot in the trade-off between accuracy and sensitivity, featuring an accuracy of 75% and a sensitivity of 45%. While the false positive rate is higher compared to xgboost, this is outweighed by the better sensitivity. Given the cost imbalance inherent to this problem, a false alarm is considered to be less severe than a missed event. Multivariate adaptive regression splines yield high accuracy and comparably low false positive rate as well, but fall behind in terms of sensitivity. Albeit results obtained by any variant of logistic regression, the Pegasos support vector machine and the Bayesian regularized neural network yield comparably high sensitivity values, these go along with rather low accuracies and high false positive rates. These models tend to be biased towards predicting `TRUE` outcomes, thus leading to both more true positives (higher sensitivity) and false positives (lower accuracy, higher FPR). These results are also reflected in Fig. 2: In terms of predicted probabilities, MARS, xgboost and the three variants of random forest display more distinct discriminative capabilities than the other models. At the same time, it is apparent that resulting distributions of predicted probabilities are different across all models – even if they exhibit similar classification metrics. In this context, it should therefore always be kept in mind that metrics obtained from a confusion matrix lead to loss of information, since the entire range of predicted probabilities is classified into two classes, thereby exhibiting an abrupt break at some point in this continuous range. The added information contained in (continuous) probability estimates is therefore lost when using discontinuous scoring metrics, which is particularly undesirable when analysing a highly imbalanced data set.

5.3. Variable importance

Findings related to the most important variables are largely consistent with other studies and our initial expectations (Fig. 4). Two important aspects have to be kept in mind, though: First, high goodness-of-fit metrics of a model do not necessarily entail causality between the explanatory variables used in the model and the target variable, but merely indicate a relationship between the independent variables and the outcome. See (Mannering, 2018) for a discussion on the trade-offs between predictive capability and causality/inference capability Second, it has to be noted, that diagnostics for most statistical learning techniques are targeted towards the information content of the variables rather than their directionality. As opposed to linear regression, no simple parametric descriptions corresponding to regression coefficients are available in more complex models such as the majority of models used in this study. Albeit the relationship between a single feature and predicted values can be assessed, the analysis of multivariate interactions is cumbersome. Specifically, the use of partial dependence plots – which show the marginal effect of a feature on the class probability of the predicted outcome of a previously fit model, considering the average affect of all other variables (Friedman, 2001) – can be useful for interpreting the causal relationship between the feature and the model outcome. Yet, it has to be noted that the maximum number of features to look at jointly is limited to three. In the following section, two partial dependence plots are explored exemplarily.

The high importance of traffic volume seems obvious, since a higher number of vehicles passing a section naturally increases the probability that one or more accidents will occur on this section. However, this relationship is not straightforwardly linear, as indicated by the partial dependence plot (Fig. 6). The importance of restrictions on overtaking (both for heavy good vehicles and all vehicles) suggests that no overtaking zones might be safer than sections without any restrictions.

The importance of various other road geometrics and features is plausible as well. Characteristics of the transverse gradient (median, existence of a zero-crossing) might be related to an increased risk of hydroplaning. The road surface type and its characteristics (in

particular roughness, evenness and friction changes within the segment) as well as the maximum damage to the road surface indicate adverse infrastructure characteristics that contribute to an increased accident risk. The existence of event sections, the number of lanes and the width of the breakdown lane are not surprising either, since these features are all related to possible disturbances in traffic flow.

While vehicle speed has been found to be of high importance for both accident severity and accident frequency, speed limits are only of medium importance in our model. This is most likely attributable to the fact that speed limits do not necessarily reflect the actual speed of vehicles, but rather a mere regulation according to the traffic code. The speed limit categories chosen are probably not sufficiently capable of representing the actual vehicle traveling speeds. In addition, using interaction effects between time, weekday and AADT might not be sufficiently capable to mimic high resolution traffic flow data.

The somewhat low influence of bendiness (as well as curvature and radius) could be explained by considering the segment length. Given the strict regulations on highway curve radii to avoid needlessly winding freeways, short highway segments of 250 m only do not exhibit large course angle changes since related curve radii are appropriately large, entailing that values for bendiness are comparably low.

The importance of bridge-sections is somewhat unexpected and difficult to reason. Specifically, bridge sections seem to be saver than sections without bridges. It has to be noted that no minimum bridge length was specified. Therefore, bridge sections may comprise a broad range of different bridge sizes, including large Alpine bridges whose length exceeds one kilometer and heights up to 150 m as well as small bridges with only several meters height and length. Consequently, not all bridges do stretch across valleys or rivers, but may also be overpasses over some smaller roads. Overall, the share of accidents that happen on bridges is slightly lower than the share of bridges on the whole highway network, indicating that bridge sections are seemingly safer than other sections.

Influence of climate variables (average number of hot days, freeze-thaw cycles and annual precipitation totals) is largely consistent with other studies (Bergel-Hayat et al., 2013; Koetse and Rietveld, 2009). Interaction-effects and nonlinear relationships to accident occurrence also prevail with respect to climate variables. However, partial dependence of accident occurrence on the number of frost days and annual rainfall amount illustrates nicely that modelling results reported for single features are not necessarily grounded in underlying causal relationships (Fig. 7). Rather, regional effects caused by the Austrian Alps (decrease of accident occurrence in areas with an average amount of 135–185 frost days per year and average annual precipitation totals between 1100 and 1450 mm) and the eastern lowlands (hot and dry basin areas south-east of Vienna, which seem to have no average impact on class probability) prevail in the plot.

The seemingly low influence of weather effects on road crashes may be attributable to three different causes: First, time and location of accidents are subject to uncertainties. Therefore, the assignment to a certain segment and hour is likely erroneous in several cases. Second, weather data might not be as accurate as the precision of the VERA reanalysis implies. While some weather measurements are considered to be rather robust (e.g. temperature, pressure), other phenomena such as wind or precipitation are governed by small-scale processes that are extremely difficult to capture in reanalysis models. Third, the influence of weather on accidents is actually negligible compared to the other covariates under consideration. Most other studies have focused on weather influence only without considering a comparably large number of other features. Overall, we argue that the low importance of weather variables used in this study is a combination of all three aspects.

Finally, it has to be noted that it is virtually impossible to include all the data that could potentially determine the likelihood of a traffic accident into a statistical model. This lack of information, which is commonly known as the problem of 'unobserved heterogeneity', may lead to biased and inconsistent parameter estimates and, consequently, erroneous inference and prediction. Mannering et al. (2016) provide an extensive discussion of this problem, including a summary of methodological approaches to account for this issue. While consideration of unobserved heterogeneity is particularly important when using count data models (operating at a temporally aggregated scale), this issue does apply to temporally disaggregated models using a binary outcome as well. Albeit several variables with possible heterogeneous effects such as temporal instability, weather effects and roadway characteristics are – at least to some extent – covered by in this work due to (i) the large number of covariates and (ii) the disaggregation of accident data to a high temporal and spatial resolution, potentially important yet unobserved explanatory variables may for instance be related to roadway characteristics, environmental conditions and traffic characteristics, as well as driver-specific features. Possible approaches that

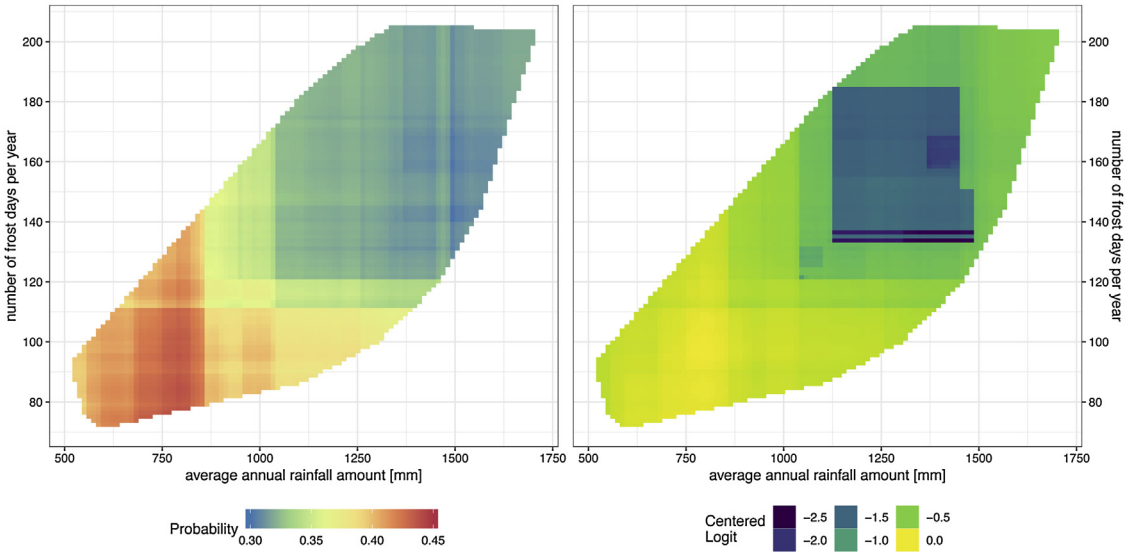

Fig. 7. Multi-predictor partial dependence plot illustrating the individual effects as well as the interactions between annual average rainfall amount and average air temperature. The left panel shows partial dependence of accident probability on the two selected variables, while the right panel displays partial dependence as logit contribution of the two climate variable on class probability from the perspective of the model. Note that the figure is restricted to the region of the data by plotting only values within the convex hull of their training values (i.e., no extrapolation).

could be explored in the context of the present setup to tackle the issue of unobserved heterogeneity include for instance random parameter logit models or Markov regime-switching models. While this potentially important issue is beyond the scope of the current paper, the detailed assessment of unobserved heterogeneity on large high-resolution imbalanced data sets is considered a worthwhile aspect for further research.

6. Conclusion and outlook

This paper presents a novel approach of assessing the importance of potentially accident-causing covariates by employing state-of-the art methods for statistical learning to analyze a high-resolution dataset.

Given the problems inherent to the assessment of temporal and spatial high resolution data in this context, we have outlined a methodological blueprint of how such an analysis may be conducted successfully. Resulting models are pretty well capable of providing consistent results with respect to determining important regressors. Nevertheless, it should be kept in mind that results presented in this study are of course subject to uncertainties and might not be straightforwardly generalizable.

Having described the modeling approach with a methodological focus, further work should be targeted at a more detailed assessment of the results from a traffic-safety point of view. Therefore, next steps should focus on investigating whose sections' outcome is captured well, and shed some light on the why. In addition, further analysis featuring variants of bootstrap aggregating could be useful for improving the robustness of the results. We propose several concrete analysis steps for this empirical assessment:

Further temporal aggregation: Given the assumption that results obtained from any learners applied to the dataset featuring hourly values are subject to uncertainty, the temporal binning size could be adjusted in order to create coarser, yet more robust aggregates. These aggregated data could be used to test the hypothesis that the significance of results would increase with increasing binning level. While some information is lost, since variables related to some sort of timestamp (i.e. hour and weekday classification, respectively) have to be dropped, a more robust assessment might prove to be conclusive.

Assessing model performance using a meta variable: In order to further investigate contributing factors to model quality, several approaches featuring a new binary meta target variable, which is derived from the confusion matrices of the existing model results, could be tested. Multiple definitions of how to derive such a meta-variable are possible. Machine learning models for binary classification could again be trained to assess variable importance for this new meta model.

Balanced bagging: Following the line of Wallace et al. (2011), bagging an ensemble of classifiers induced over balanced bootstrap training samples and predicting the outcome state by using a majority vote could be a valuable approach to obtain more robust results.

Correlation issues: Further insights might be gained by considering collinearity in variables and (spatio-temporal) autocorrelation effects.

Unobserved heterogeneity: Since it is impossible to include all the data that could potentially determine the likelihood of a traffic accident into a statistical model, future work might focus on model formulations accounting for unobserved heterogeneity (Mannering, 2018).

Knowledge-extraction and expert assessment: Tools for further assessment of black-box models, including – among others – *Local Interpretable Model-Agnostic Explanations* [LIME, Ribeiro et al. (2016)] and *Descriptive mAchine Learning EXplanations* [DALEX, Biecek (2018)] could be used for an in-depth assessment of model quality. In addition, the case-specific random forests (Xu et al., 2016), which are tailored to specific points of interest in the regressor space, could be employed to specifically assess certain road sections of interest.

In addition, a comparison with similar analysis conducted in other countries might provide substantial further insights into the applicability of the proposed methodology.

Overall, we hope that our findings will contribute to opening up new methodological applications of statistical learning methods in the field of road safety research.

Acknowledgments

The authors gratefully acknowledge ASFiNAG for providing traffic volume data and for approving the use of RoadSTAR measurement data.

References

Abdel-Aty, M., Pande, A., Lee, C., Gayah, V., Santos, C.D., 2007. Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways. J. Intell. Transportation Syst. 11, 107–120. https://doi.org/10.1080/15472450701410395.

Andrey, J., 2010. Long-term trends in weather-related crash risks. J. Transport Geogr. 18, 247–258. https://doi.org/10.1016/j.jtrangeo.2009.05.002.

Andrey, J., Mills, B., Leahy, M., Suggett, J., 2003. Weather as a chronic hazard for road transportation in Canadian cities. Nat. Hazards 28, 319–343. https://doi.org/10.1023/A:1022934225431.

ASFINAG, 2018. Verkehrsentwicklung. https://www.asfinag.at/verkehr/verkehrszaehlung/. Accessed: 4 May 2018.

Basso, F., Basso, L.J., Bravo, F., Pezoa, R., 2018. Real-time crash prediction in an urban expressway using disaggregated data. Transport. Res. Part C: Emerg. Technol. 86, 202–219. https://doi.org/10.1016/j.trc.2017.11.014.

Benoit, D., Van den Poel, D., 2017. bayesQR: a Bayesian approach to quantile regression. J. Stat. Softw. 76, 1–32. https://doi.org/10.18637/jss.v076.i07.

Benoit, D.F., Alhamzawi, R., Yu, K., 2013. Bayesian lasso binary quantile regression. Comput. Stat. 28, 2861–2873. https://doi.org/10.1007/s00180-013-0439-0.

Benoit, D.F., Van den Poel, D., 2012. Binary quantile regression: a Bayesian approach based on the asymmetric Laplace distribution. J. Appl. Econometr. 27, 1174–1188. https://doi.org/10.1002/jae.1216.

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

Biecek, P., 2018. DALEX: Descriptive mAchine Learning Explanations. https://pbiecek.github.io/DALEX/. Accessed 26 June 2018.

Bijleveld, F., Churchill, T., 2009. The influence of weather conditions on road safety. An assessment of the effect of precipitation and temperature. R-2009-9.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., Lang, M., 2017. mlrMBO: a modular framework for model-based optimization of expensive black-box functions. arXiv, 1703.03373. URL: http://arxiv.org/abs/1703.03373.

BMVIT, 2017. Statistik Strae & Verkehr [Statistics on road and traffic]. Austrian Federal Ministry for Transport, Innovation and Technology. https://www.bmvit.gv.at/service/publikationen/verkehr/strasse/statistik_strasseverkehr.html.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

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. https://doi.org/10.1016/j.aap.2008.01.001.

Buchner, J., Georgakakis, A., Nandra, K., Hsu, L., Rangel, C., Brightman, M., Merloni, A., Salvato, M., Donley, J., Kocevski, D., 2014. X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue. Astron. Astrophys. 564, A125. https://doi.org/10.1051/0004-6361/201322971.

Bühlmann, P., Hothorn, T., 2007. Boosting algorithms: regularization, prediction and model fitting. Stat. Sci. 22, 477–505. https://doi.org/10.1214/07-STS242.

Burden, F., Winkler, D., 2009. Bayesian regularization of neural networks. In: Livingstone, D.J. (Ed.), Artificial Neural Networks: Methods and Applications. Humana Press, Totowa, NJ, pp. 23–42. https://doi.org/10.1007/978-1-60327-101-1_3.

Cateni, S., Colla, V., Vannucci, M., 2014. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. Neurocomputing 135, 32–41. https://doi.org/10.1016/j.neucom.2013.05.059.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chen, F., Chen, S., Ma, X., 2018a. Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data. J. Safety Res. 65, 153–159. https://doi.org/10.1016/j.jsr.2018.02.010.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., 2018. xgboost: Extreme Gradient Boosting. https://CRAN.R-project.org/package=xgboost, r package version 0.6.4.1.

Chen, T., Guestrin, C., 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16. New York, NY, USA, ACM. pp. 785–794. https://doi.org/10.1145/2939672.2939785.
Cox, D.R., 1958. The regression analysis of binary sequences. J. R. Stat. Soc. Ser. B: Methodological 20, 215–242.
Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Presshttps://doi.org/10.1017/CBO9780511801389.
Dowle, M., Srinivasan, A., 2018. data.table: Extension of 'data.frame'. https://CRAN.R-project.org/package=data.table, r package version 1.11.4.
Edwards, J.B., 1996. Weather-related road accidents in England and Wales: a spatial analysis. J. Transport Geogr. 4, 201–212. https://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. https://doi.org/10.1016/S0001-4575(03)00085-X.
Feroz, F., Hobson, M.P., Bridges, M., 2009. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. Mon. Not. R. Astron. Soc. 398, 1601–1614. https://doi.org/10.1111/j. 1365-2966.2009.14548.x.
Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80, 27–38.
Foresee, F.D., Hagan, M.T., 1997. Gauss-Newton approximation to Bayesian learning. Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 3 1930–1935.
Freund, Y., Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. 55, 119–139. https://doi.org/10.1006/jcss.1997.1504.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: a statistical view of boosting. Ann. Statist. 28, 337–407. https://doi.org/10.1214/aos/1016218223.
Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Statist. 19, 1–67. https://doi.org/10.1214/aos/1176347963.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Statist. 29, 1189–1232. https://doi.org/10.1214/aos/1013203450.
Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Stat. Data Anal. 38, 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2. Nonlinear Methods and Data Mining.
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].
Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1.
Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks. arXiv, 1406.2661. URL: https://arxiv.org/abs/1406.2661.
Handley, W.J., Hobson, M.P., Lasenby, A.N., 2015a. PolyChord: nested sampling for cosmology. Mon. Not. R. Astron. Soc. Lett. 450, L61–L65. https://doi.org/10.1093/mnrasl/slv047.
Handley, W.J., Hobson, M.P., Lasenby, A.N., 2015b. PolyChord: next-generation nested sampling. Mon. Not. R. Astron. Soc. 453, 4384–4398. https://doi.org/10.1093/mnras/stv1911.
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd ed. Springer, New York. https://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. https://doi.org/10.1029/2008JD010201.
He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. https://doi.org/10.1109/TKDE.2008.239.
Heinze, G., 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. Stat. Med. 25, 4216–4226. https://doi.org/10.1002/sim.2687.
Heinze, G., Schemper, M., 2002. A solution to the problem of separation in logistic regression. Stat. Med. 21, 2409–2419. https://doi.org/10.1002/sim.1047.
Hofner, B., Mayr, A., Robinzonov, N., Schmid, M., 2014. Model-based boosting in R: a hands-on tutorial using the R package mboost. Comput. Stat. 29, 3–35. https://doi.org/10.1007/s00180-012-0382-5.
Hothorn, T., Hornik, K., Zeileis, A., 2006. Unbiased recursive partitioning: a conditional inference framework. J. Comput. Graph. Stat. 15, 651–674. https://doi.org/10.1198/106186006X133933.
Japkowicz, N., Stephen, S., 2002. The class imbalance problem: a systematic study. Intell. Data Anal. 6, 429–449.
Katrakazas, C., Antoniou, C., Yannis, G., 2019. Time series classification using imbalanced learning for real-time safety assessment. In: TRB (Ed.), Transportation Research Board 98th Annual Meeting. TRB. pp. 19-04457.
Koetse, M.J., Rietveld, P., 2009. The impact of climate change and weather on transport: an overview of empirical findings. Transport. Res. Part D: Transport Environ. 14, 205–221. https://doi.org/10.1016/j.trd.2008.12.004.
Kosmidis, I., 2017. brglm: Bias Reduction in Binary-Response Generalized Linear Models. r package version 0.6.1. http://www.ucl.ac.uk/ucakiko/software.html.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer, New York. https://doi.org/10.1007/978-1-4614-6849-3.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transport. Res. Part A: Policy Pract. 44, 291–305. https://doi.org/10.1016/j.tra.2010.02.001.
MacKay, D.J., 1992. Bayesian interpolation. Neural Comput. 4, 415–447. https://doi.org/10.1162/neco.1992.4.3.415.
Maher, M.J., Summersgill, I., 1996. A comprehensive methodology for the fitting of predictive accident models. Accid. Anal. Prev. 28, 281–296. https://doi.org/10.1016/0001-4575(95)00059-3.
Mais, D., Lloyd, D., Davies, J., 2016. Modelling weather effects on road casualty statistics. Significance 13, 28–31. https://doi.org/10.1111/j. 1740-9713.2016.00880.x.
Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and machine learning forecasting methods: concerns and ways forward. PLoS ONE 13. https://doi.org/10.1371/journal.pone.0194889.
Malley, J.D., Kruppa, J., Dasgupta, A., Malley, K.G., Ziegler, A., 2012. Probability machines: consistent probability estimation using nonparametric learning machines. Methods Inf. Med. 51, 74–81. https://doi.org/10.3414/ME00-01-0052.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Anal. Methods Accid. Res. 17, 1–13. https://doi.org/10.1016/j.amar.2017.10.002.
Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Anal. Methods Accid. Res. 1, 1–22. https://doi.org/10.1016/j.amar.2013.09.001.
Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16. https://doi.org/10.1016/j.amar.2016.04.001.
Manski, C.F., 1975. Maximum score estimation of the stochastic utility model of choice. J. Econom. 3, 205–228. https://doi.org/10.1016/0304-4076(75)90032-9.
Mason, L., Baxter, J., Bartlett, P., Frean, M., 1999. Boosting algorithms as gradient descent. Proceedings of the 12th International Conference on Neural Information Processing Systems NIPS'99. MIT Press, Cambridge, MA, USA, pp. 512–518.
Maurer, P., Meissner, M., Fuchs, M., Gruber, J., Foissner, P., 2002. Straßenzustandserfassung mit dem RoadSTAR – Messsystem und Genauigkeit.
Milborrow, S., 2017. earth: Multivariate Adaptive Regression Splines. https://CRAN.R-project.org/package=earth, r package version 4.6.0.
Milborrow, S., 2018. Notes on the earth package. http://www.milbo.org/doc/earth-notes.pdf.
Nguyen, D., Widrow, B., 1990. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. Proceedings of the IJCNN, vol. 3 21–26.
Omranian, E., Sharif, H., Dessouky, S., Weissmann, J., 2018. Exploring rainfall impacts on the crash risk on Texas roadways: a crash-based matched-pairs analysis approach. Accid. Anal. Prev. 117, 10–20. https://doi.org/10.1016/j.aap.2018.03.030.
Pencina, M.J., Fine, J.P., D'Agostino, R.B., 2017. Discrimination slope and integrated discrimination improvement – properties, relationships and impact of calibration. Stat. Med. 36, 4482–4490. https://doi.org/10.1002/sim.7139.
Pérez-Rodríguez, P., Gianola, D., Weigel, K.A., Rosa, G.J.M., Crossa, J., 2013. Technical note: an R package for fitting Bayesian regularized neural networks with applications in animal breeding. J. Anim. Sci. 91, 3522–3531. https://doi.org/10.2527/jas.2012-6162.
Ribeiro M.T. Singh S. Guestrin C. 2016. Why Should I Trust You?: explaining the predictions of any classifier. 1602.04938.
Roustant, O., Ginsbourger, D., Deville, Y., 2012. DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. J. Stat. Softw. 51, 1–55. https://doi.org/10.18637/jss.v051.i01. https://www.jstatsoft.org/v051/i01.
Sáez, J.A., Krawczyk, B., Woźniak, M., 2016. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. Pattern Recogn. 57, 164–178. https://doi.org/10.1016/j.patcog.2016.03.012.
Sayers, M., Karamihas, S., 1998. Little Book of Profiling.
Schlögl, M., 2018. A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach. Transp. Res. D, to be submitted.
Schlögl, M., Stütz, R., 2017. Methodological considerations with data uncertainty in road safety analysis. Accid. Anal. Prev. https://doi.org/10.1016/j.aap.2017.02.001.
Schmid, M., Hothorn, T., 2008. Boosting additive models using component-wise P-splines. Comput. Stat. Data Anal. 53, 298–311. https://doi.org/10.1016/j.csda.2008.09.009.
Sculley, D., 2009. Large scale learning to rank. NIPS Workshop on Advances in Ranking.
Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A., 2011. Pegasos: primal estimated sub-gradient solver for SVM. Math. Program. 127, 3–30. https://doi.org/10.1007/s10107-010-0420-4.
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. https://doi.org/10.1016/0001-4575(94)00078-Z.
Statistik Austria, 2018. Unfälle mit Personenschaden. https://www.statistik.at/web_de/statistiken/energie_umwelt_innovation_mobilitaet/verkehr/strasse/unfaelle_mit_personenschaden/index.html. Accessed: 4 May 2018.
Steinacker, R., Ratheiser, M., Bica, B., Chimani, B., Dorninger, M., Gepp, W., Lotteraner, C., Schneider, S., Tschannett, S., 2011. A mesoscale data analysis and downscaling method over complex terrain. Mon. Weather Rev. 134, 2758–2771. https://doi.org/10.1175/MWR3196.1.
Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 8, 25. https://doi.org/10.1186/1471-2105-8-25.
Theofilatos, A., 2017. Incorporating real-time traffic and weather data to explore road accident likelihood and severity in urban arterials. J. Safety Res. 61, 9–21. https://doi.org/10.1016/j.jsr.2017.02.003.
Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accid. Anal. Prev. 72, 244–256. https://doi.org/10.1016/j.aap.2014.06.017.
Theofilatos, A., Yannis, G., Kopelias, P., Papadimitriou, F., 2016. Predicting road accidents: a rare-events modeling approach. Transp. Res. Proc. 14, 3399–3405. https://

doi.org/10.1016/j.trpro.2016.05.293. Transport Research Arena TRA2016.
Ueckermann, A., Steinauer, B., 2008. The weighted longitudinal profile. Road Mater. Pavement Des. 9, 135–157. https://doi.org/10.1080/14680629.2008.9690111.
Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience.
Vapnik, V.N., 2000. The Nature of Statistical Learning Theory. Springer New York, New York, NY. https://doi.org/10.1007/978-1-4757-3264-1.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, 4th ed. Springer, New York , ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.
Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A., 2011. Class imbalance, redux. 2011 IEEE 11th International Conference on Data Mining 754–763.
WHO, 2015. Global Status Report on Road Safety 2015. World Health Organization.
Willett, P., 1999. Dissimilarity-based algorithms for selecting structurally diverse sets of compounds. J. Comput. Biol. 6, 447–457. https://doi.org/10.1089/106652799318382.
Wright, M.N., Ziegler, A., 2017. ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01.
Wu, Y., Abdel-Aty, M., Lee, J., 2018. Crash risk analysis during fog conditions using real-time traffic data. Accid. Anal. Prev. 114, 4–11. https://doi.org/10.1016/j.aap.2017.05.004. Road Safety on Five Continents 2016 - Conference in Rio de Janeiro, Brazil.
Xu, R., Nettleton, D., Nordman, D.J., 2016. Case-specific random forests. J. Comput. Graph. Stat. 25, 49–65. https://doi.org/10.1080/10618600.2014.983641.
Yannis, G., Dragomanovits, A., Laiou, A., Torre, F.L., Domenichini, L., Richter, T., Ruhl, S., Graham, D., Karathodorou, N., 2017. Road traffic accident prediction modelling: a literature review. Proc. Inst. Civil Eng. Transp. 170, 245–254. https://doi.org/10.1680/jtran.16.00067.
Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accid. Anal. Prev. 50, 371–376. https://doi.org/10.1016/j.aap.2012.05.011.
Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-time crash risk prediction using long short-term memory recurrent neural network. In: TRB (Ed.), Transportation Research Board 98th Annual Meeting. TRB. pp. 19-03414.