Accident Analysis and Prevention 136 (2020) 105398

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach

Matthias Schlögl[a,b,*]

[a] Institute of Statistics, University of Natural Resources and Life Sciences (BOKU), Vienna, Austria
[b] Transportation Infrastructure Technologies, Austrian Institute of Technology (AIT), Vienna, Austria

ARTICLE INFO

MSC:
62-07

Keywords:
Adverse weather effects
Imbalanced data
Binary classification
Balanced bagging
Accident analysis
Road safety
Random forest
xgBoost

ABSTRACT

Determining and understanding the environmental factors contributing to road traffic accident occurrence is of core importance in road safety research. In this study, a methodology to obtain robust and unbiased results when modeling imbalanced, high-resolution accident data is described. Based on a data set covering the whole highway network of Austria in a fine spatial (250 m) and temporal (1 h) scale, the effects of 48 covariates on accident occurrence are analyzed, with a special emphasis on real-time weather variables obtained through meteorological re-analysis. A balanced bagging approach is employed to cope with the issue of class imbalance. By fitting different tree-based classifiers to a large number of bootstrapped training samples, ensembles of binary classification models are established. The final prediction is achieved through majority vote across each ensemble, resulting in a robust prediction with reduced variance. Findings show the merits of the proposed approach in terms of model quality and robustness of the results, consistently displaying accuracies around 80% while exhibiting sensitivities of approximately 50%. In addition to certain features related to roadway geometrics, surface condition and traffic volume, a number of weather variables are found to be of importance for predicting accident occurrence. The proposed methodological take may not only pave the way for further analyses of high-resolution road safety data including real-time information, but can also be transferred to any other imbalanced classification problem.

1. Introduction

Along with geometric characteristics, road surface conditions, traffic status and driver behavior, weather conditions are among the most important variables potentially influencing the occurrence of highway crashes. Consequently, the effects of weather on road accidents have been in the focus of research efforts for several decades, and it is broad consent that weather does have an effect on road safety (Theofilatos and Yannis, 2014; Koetse and Rietveld, 2009).

The most commonly considered meteorological variable that is analyzed in terms of road safety effects is precipitation. Empirical evidence on the adverse effects of rainfall on accident frequency is abundant (Theofilatos and Yannis, 2014; Bergel-Hayat et al., 2013; Koetse and Rietveld, 2009; Brijs et al., 2008; Eisenberg, 2004; Edwards, 1996; Fridstrøm et al., 1995; Shankar et al., 1995). Albeit studies on other types of precipitation are less widespread, a handful of studies suggest that snowfall does have a negative effect on traffic safety as well (Heuel et al., 2014; Eisenberg and Warner, 2005; Andrey et al., 2003; Fridstrøm et al., 1995). The effects of other meteorological variables are explored far less extensively in literature. Several studies focusing on various different weather variables have found relationships between fog (Hassan and Abdel-Aty, 2013; Abdel-Aty et al., 2012, 2011; Al-Ghamdi, 2007; Hermans et al., 2006), wind (Brijs et al., 2008; Hermans et al., 2006; Baker and Reynolds, 1992) or air temperature (Antoniou et al., 2013; Brijs et al., 2008) and traffic safety.

A majority of existing studies are based on monthly (e.g. Bergel-Hayat et al., 2013) or daily (e.g. Brijs et al., 2008) aggregates obtained from weather stations or large gridded data sets. However, the use of real-time weather data has gained center stage in recent years (Wu et al., 2018; Theofilatos, 2017; Yu et al., 2013). While these studies have made use of data from weather radars (Jaroszweski and McNamarau, 2014) or the closest weather station (Chen et al., 2018a; Wu et al., 2018; Theofilatos, 2017) to obtain estimates for certain weather variables, Becker et al. (2018) and Schlögl et al. (2019) were the first ones to use nationwide high-resolution re-analysis data providing location-specific weather variables in time steps of one hour for modeling accident probability.

In this context it is important to emphasize that contradictory

[*] Corresponding author.
E-mail address: matthias.schloegl@boku.ac.at.

https://doi.org/10.1016/j.aap.2019.105398
Received 3 June 2019; Received in revised form 6 September 2019; Accepted 5 December 2019
Available online 17 December 2019

results and mixed evidence are often reported in the area of accident data analysis, entailing that findings related to weather effects are only partially conclusive (Theofilatos, 2017; Theofilatos and Yannis, 2014; Koetse and Rietveld, 2009; Maze et al., 2006). It can be argued that this is most likely attributable to two different main causes. First, there is high uncertainty of police-reported accident data and various covariates (Schlögl and Stütz, 2017). In addition, underreporting and unobserved heterogeneity are common problems in accident data analysis (Mannering and Bhat, 2014). Due to error propagation and in combination with small sample sizes, resulting accident prediction models are likely to exhibit uncertainties which can hardly be avoided. Second, statistically sound model formulation is a key to obtain robust, unbiased results. Unfortunately, this is not always the case in the field of accident research. Particular care should be taken that fitted models do not fall victim to overfitting. Therefore, the whole data set used should be splitted into training, validation and test data partitions. Model quality should always be assessed on a holdout test data, since goodness-of-fit metrics reported for the accordance between the fitted model and the observations used to fit the respective model will underestimate the true error rate. Using different random seeds for creating multiple random splits of the full data set can be used to further enhance outcome robustness.

Against this background, the study pursues two main objectives. First, a methodology for obtaining robust and unbiased results when modeling imbalanced, high-resolution accident data is described. Given the high spatial (250 m) and temporal (1 h) resolution of the data set, the dichotomous outcome variable (indicating accident/no accident) is severely imbalanced. After introducing balanced bagging as an alternative approach to cope with the issue of class imbalance when performing accident data analysis at such a fine scale, an ensemble of binary classification forests is employed, and the final prediction is obtained through majority vote. The resulting outcome is robust with reduced variance.

Second, special emphasis is put on the impact of weather effects on road accident occurrence when using high-resolution weather data derived through re-analysis. Weather effects are not assessed solitary but rather in combination with other important covariates known to affect traffic safety. Overall, a total number of 48 explanatory variables are considered.

2. Data and methods

2.1. Data

The time period under consideration spans exactly one calender year from 2016-01-01 00:00 CET to 2017-01-01 00:00 CET at a temporal resolution of 1 h (i.e. 8784 hours in total). The whole highway network of Austria, which featured a total length 1719 km in the year 2016 (BMVIT, 2017), is considered in this study.

In total, 48 explanatory variables related to weather (11) and climate (3), road geometry (15) and road condition (11), traffic code (4), traffic volume (2) and time (2) are analyzed.

High resolution weather data used in this study are obtained through a VERAflex re-analyses (Steinacker et al., 2011). Gridded data sets for the meteorological variables under consideration are available at a temporal resolution of one hour and a spatial grid of 250 m. An overview of real-time weather variables considered is provided in Table 1.

Data on road condition and geometrics are derived from the most recent measurement campaign (conducted in 2017) of the RoadSTAR laboratory, which is a rebuilt truck carrying numerous mobile measurement systems (Schlögl and Stütz, 2017; Maurer et al., 2002). Raw data are collected separately for each lane in both directions at a spatial resolution of 1 m (using an inertial measurement unit and a differential GPS), and are subsequently aggregated to the segment length of 250 m. The segment length was determined empirically by conducting a sensitivity analysis using count data regression models across different segment lengths (ranging between 150 and 1000 m) as well as different segment starting points. An overview and description of all covariates used can be found in the appendix, and more detailed information on both road safety data collection in Austria and related uncertainties can be found in the companion paper by Schlögl et al. (2019) as well as Schlögl and Stütz (2017).

Table 1
Description of real-time weather variables obtained through meteorological re-analysis based on VERAflex runs. Data are available on a 250 m raster covering Austria at a temporal resolution of 1 h.

Variable name	Meteorological indicator	Type	Unit
T	Air temperature	Numeric	[° C]
RR	Accumulated precipitation	Numeric	[mm]
RR_lead	Precipitation (one hour ahead)	Numeric	[mm]
RR_lag	Precipitation (one hour behind)	Numeric	[mm]
RR_SA	Percentage of solid precipitation	Numeric	[%]
SCI	Surface condition index	Factor	[][a]
RF	Relative humidity of the air	Numeric	[%]
P	Air pressure	Numeric	[hPa]
FFX	Wind gusts (maximum wind speed)	Numeric	[m/s]
DD	Wind direction	Integer	[°]
ASD	Absolute sunshine duration	Integer	[min]

[a] Factor levels are: dry, damp, wet, snowy and glaze.

2.2. Method

From a methodological point of view, the biggest challenge in analyzing high resolution accident occurrence data sets is the issue of class imbalance that is inherent to the underlying target variable under consideration. Due to the high temporal and spatial resolution, the resulting data set comprising 111,812,424 rows contains 111,810,864 non-event instances and ‘only’ 1560 accident instances.

Imbalanced classification problems are common across many scientific disciplines, ranging from medicine over behavior analysis to text and video mining (Haixiang et al., 2017; Krawczyk, 2016), and several methods to tackle this issue have been proposed. As opposed to other studies in the road safety domain, which used a matched case–control design (e.g. Chen et al., 2018a; Abdel-Aty et al., 2012), synthetic minority oversampling (SMOTE; c.f. Basso et al., 2018; Katrakazas et al., 2019; Yuan et al., 2019), or a combination of minority oversampling and maximum dissimilarity undersampling (Schlögl et al., 2019) to obtain a balanced training data set, a method grounded in bootstrap aggregating (Hastie et al., 2009; Breiman, 1996) is proposed in this study.

While SMOTE is based on introducing new synthetic event class instances, balanced bagging does solely rely on actual observations. Arguably, this is beneficial in case of severe class imbalance present within large data sets, where sophisticated over- and undersampling strategies provide little added value with respect to model quality and are costly in terms of computational performance. Specifically, the main criticism of SMOTE is due to the interpolative method of pseudo-minority point generation based on nearest neighbours in the minority class, which entails that the synthetic set is generated entirely within the convex hull enclosing the minority instances (Bellinger et al., 2018, 2016; Wallace et al., 2011). Corresponding negative effects stemming from the resulting generative bias (i.e., some instances may be either synthesized inside the majority class or not synthesized close enough to the majority class) grow quickly with absolute imbalance and dimensionality, therefore rendering SMOTE rather suitable for well-sampled low-dimensional data sets than severely imbalanced high-dimensional data (Bellinger et al., 2016). In addition, Wallace et al. (2011) demonstrated that a combination of bagging and undersampling consistently outperformed other strategies (including SMOTE and weighted SVM) in terms of predictive performance, using synthetically

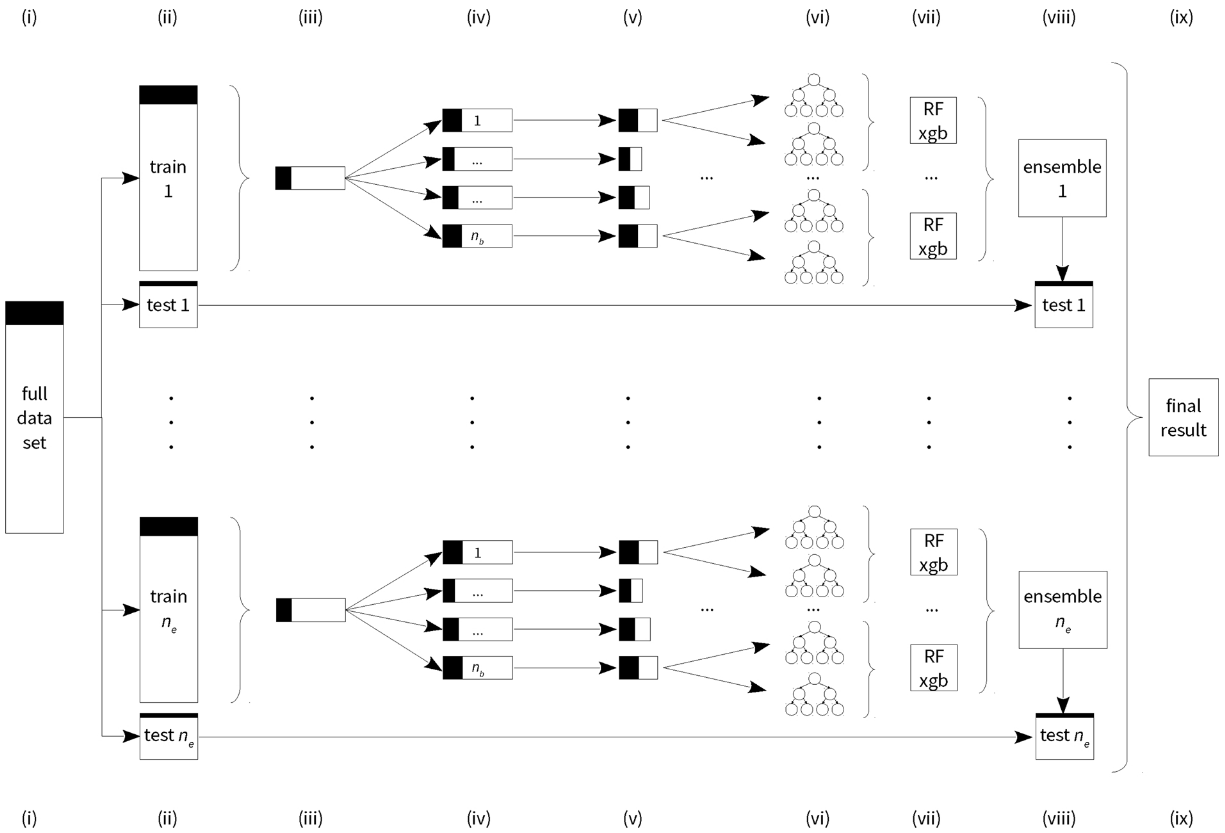

Fig. 1. Schematic overview of the methodological approach: (i) denotes the original population, which is split into train and test data set n_e times, using different random seeds (ii). Each training data set (iii) is used to create n_b bootstrap samples (iv), which are then balanced using random downsampling (v). n_t decision trees (vi) are learned to each of the balanced samples, resulting in n_b random forest or gradient boosted tree models (vii). This n_b-member ensemble available for each of the n_s different splits into training and test data is used to predict the outcome on the respective test data sets by employing majority vote (viii). Final variable importance and model performance metrics are eventually derived as averages of all n_s models (ix). Black portions of the data sets represent accidents, whereas the larger white portions illustrate non-events.

generated imbalanced data. In addition, they showed that this approach is particularly effective in cases when the training data set was not separable.

Following the line of Wallace et al. (2011), the approach in this paper is based on bagging an ensemble of classifiers induced over balanced bootstrap training samples (Fig. 1).

The strategy can be summarized as follows:

1 Split the data set randomly into a training data set (80%) and a test data set (20%)[1] .
2 Take n_b bootstrap samples from the training data set.
3 Balance each of the samples through random downsampling.
4 Train a classifier to the balanced samples.
5 Predict the outcome on the test data set.
6 Use majority vote to derive final decision.
7 Obtain model performance metrics by comparing results of the majority vote with actual observations in the holdout.

In order to obtain robust results, this process is repeated n_e times, thus bearing some similarity to a nested cross-validation design (Schratz et al., 2019). In the present case, $n_e = 5$ iterations (i.e. ensembles) featuring different randomly sampled training and test data sets were carried out. In each iteration, $n_b = 50$ bootstrapped samples were used. Two different tree-based methods were used as classifiers. First, random forests (Hastie et al., 2009; Breiman, 2001) consisting of $n_t = 5000$ de-correlated binary classification trees are grown using the **R**-package `ranger` (Wright and Ziegler, 2017). Second, $500 \leq n_t \leq 5000$ boosted trees are fitted to all balanced bootstrap samples using extreme gradient boosting (Chen and Guestrin, 2016) via `xgboost` in **R** (Chen et al., 2018b). This results in a total number of 1,250,000 decision trees in random forest models and 606,598 boosted trees that are grown using various different samples from the original population and tested on the corresponding holdouts. These two tree-based methods have been chosen due to their model performance and computational efficiency (Schlögl et al., 2019; Wright and Ziegler, 2017; Chen and Guestrin, 2016; Hastie et al., 2009; Breiman, 2001).

Hyperparameter-tuning is done for each single model independently. Tuning of the random forest models is straightforwardly achieved using a grid search with five-fold cross-validation. Tuning of the xgboost-models is computationally more expensive due to the higher number of hyperparameters. Therefore, model-based (Bayesian) optimization using an efficient global optimization algorithm with Kriging as surrogate (Roustant et al., 2012) is employed for optimizing hyperparameters (Bischl et al., 2017).

Variable importance of random forests is obtained as permutation accuracy importance, which is derived from the differences in prediction accuracy before and after randomly shuffling all values for each single predictor separately.

Gain, which represents the relative contribution of a feature to the model based on the average contributions of the corresponding features to each tree in the model (Chen et al., 2018b; Chen and Guestrin, 2016), is reported as the main measure of feature importance for the xgboost

[1] Note that the 80-20 split applied in this case is based on a common rule of thumb (e.g. Hastie et al., 2009) governed by the Pareto principle. Given the large number of instances, the ratio for dividing data into training and test sets is of subordinate importance.

Table 2
Overview of selected scalar classification performance metrics for the five ensembles of random forests and boosted trees. In order to illustrate the trade-off between accuracy and sensitivity, results when applying different class discrimination thresholds and an adjusted voting rule are reported for xgboost ensembles. Results underline the robustness of the approach, as indicated by the similarity of metrics across all iterations.

Method (ensemble)	Threshold	Vote	Accuracy	Sensitivity	Specificity	FPR	Precision	F_1	AUC	*H*
Naive	-	-	1	0	1	0	-	-	0.50	0
Random forest (1)	0.50	0.5	0.80	0.51	0.80	0.20	< 0.1	< 0.1	0.66	0.11
Random forest (2)	0.50	0.5	0.80	0.51	0.80	0.20	< 0.1	< 0.1	0.66	0.11
Random forest (3)	0.50	0.5	0.80	0.46	0.80	0.20	< 0.1	< 0.1	0.63	0.08
Random forest (4)	0.50	0.5	0.79	0.49	0.79	0.22	< 0.1	< 0.1	0.64	0.10
Random forest (5)	0.50	0.5	0.80	0.48	0.80	0.20	< 0.1	< 0.1	0.64	0.09
xgboost (1)	0.50	0.5	0.82	0.48	0.82	0.18	< 0.1	< 0.1	0.65	0.11
xgboost (2)	0.50	0.5	0.82	0.51	0.82	0.18	< 0.1	< 0.1	0.67	0.13
xgboost (3)	0.50	0.5	0.82	0.45	0.82	0.18	< 0.1	< 0.1	0.63	0.09
xgboost (4)	0.50	0.5	0.82	0.48	0.82	0.18	< 0.1	< 0.1	0.65	0.11
xgboost (5)	0.50	0.5	0.82	0.47	0.82	0.18	< 0.1	< 0.1	0.64	0.10
xgboost (1)	0.45	0.5	0.78	0.52	0.78	0.22	< 0.1	< 0.1	0.65	0.10
xgboost (2)	0.45	0.5	0.78	0.55	0.78	0.22	< 0.1	< 0.1	0.67	0.12
xgboost (3)	0.45	0.5	0.78	0.51	0.78	0.22	< 0.1	< 0.1	0.65	0.09
xgboost (4)	0.45	0.5	0.78	0.52	0.78	0.22	< 0.1	< 0.1	0.65	0.10
xgboost (5)	0.45	0.5	0.78	0.53	0.78	0.22	< 0.1	< 0.1	0.65	0.10
xgboost (1)	0.40	0.5	0.75	0.59	0.75	0.25	< 0.1	< 0.1	0.67	0.11
xgboost (2)	0.40	0.5	0.74	0.61	0.74	0.26	< 0.1	< 0.1	0.67	0.12
xgboost (3)	0.40	0.5	0.74	0.55	0.74	0.26	< 0.1	< 0.1	0.65	0.09
xgboost (4)	0.40	0.5	0.74	0.57	0.74	0.26	< 0.1	< 0.1	0.65	0.09
xgboost (5)	0.40	0.5	0.74	0.57	0.74	0.26	< 0.1	< 0.1	0.65	0.09
xgboost (1)	0.50	0.4	0.78	0.52	0.78	0.22	< 0.1	< 0.1	0.65	0.10
xgboost (2)	0.50	0.4	0.78	0.55	0.78	0.22	< 0.1	< 0.1	0.67	0.12
xgboost (3)	0.50	0.4	0.78	0.52	0.78	0.22	< 0.1	< 0.1	0.65	0.09
xgboost (4)	0.50	0.4	0.78	0.53	0.78	0.22	< 0.1	< 0.1	0.65	0.10
xgboost (5)	0.50	0.4	0.78	0.52	0.78	0.22	< 0.1	< 0.1	0.65	0.10
xgboost (1)	0.45	0.4	0.74	0.58	0.74	0.26	< 0.1	< 0.1	0.66	0.11
xgboost (2)	0.45	0.4	0.74	0.60	0.74	0.26	< 0.1	< 0.1	0.67	0.11
xgboost (3)	0.45	0.4	0.74	0.56	0.74	0.26	< 0.1	< 0.1	0.65	0.09
xgboost (4)	0.45	0.4	0.74	0.56	0.74	0.26	< 0.1	< 0.1	0.65	0.09
xgboost (5)	0.45	0.4	0.74	0.57	0.74	0.26	< 0.1	< 0.1	0.66	0.10
xgboost (1)	0.40	0.4	0.70	0.66	0.70	0.30	< 0.1	< 0.1	0.68	0.12
xgboost (2)	0.40	0.4	0.69	0.63	0.69	0.31	< 0.1	< 0.1	0.66	0.10
xgboost (3)	0.40	0.4	0.70	0.61	0.70	0.30	< 0.1	< 0.1	0.65	0.09
xgboost (4)	0.40	0.4	0.69	0.59	0.69	0.31	< 0.1	< 0.1	0.64	0.07
xgboost (5)	0.40	0.4	0.69	0.62	0.69	0.31	< 0.1	< 0.1	0.66	0.09

models.

Thereby, the initial approach by Wallace et al. (2011) is extended with respect to multiple aspects. First, instead of training single regression trees, sophisticated tree-based ensemble learning methods are employed. Second, the robustness of the approach is demonstrated by using multiple random train-test data splits. Third, the effects of using different class discrimination thresholds and different voting rules are explored. Fourth, instead of using a synthetic data set, an application of the balanced bagging methodology in the transport domain is presented.

2.3. Model quality assessment

Model performance assessment for imbalanced data can be cumbersome, since many commonly used performance metrics used to describe the quality of the trained models tend to induce misleading conclusions. In order to provide a more balanced view of overall predictive performance, a number of different measures are reported. Most of these measures are directly derived from the confusion matrix, a two-dimensional contingency table that displays the agreement between the predicted and the actual class of the test data set in terms of counts.

- **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. In the case of severe class imbalance considering overall accuracy alone is elusive, since very high values of accuracy can be achieved by simply classifying everything as the majority class.
- **Sensitivity**, also known as true positive rate 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, analogously is the proportion of actual negatives among all negative predictions, i.e. the proportion of correctly identified negatives. Due to the low number of events, specificity is almost identical to accuracy.
- **Precision** or positive predictive value, is defined as the proportion of all positive predictions which are true positives.
- **Fallout**, also known as false positive rate, is the proportion of false positives among all negative predictions, thereby representing the percentage of 'false alarms'. Fallout is the complementary rate to specificity.
- **F_1score** is the harmonic mean of precision and sensitivity.
- **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.
- ***H*-measure** is an alternative measure for classifier performance proposed by Hand (2009). It seeks to alleviate the incoherence that AUC uses different misclassification cost distribution (and thus

different metrics) for evaluating different classifiers. As proposed by Hand and Anagnostopoulos (2014) the severity ratio, which provides an indication on the expected value of the cost ratio, is set to the reciprocal of relative class frequency. This entails that misclassifying the rare class is considered a substantially graver mistake than misclassifying the majority class.

2.4. 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.

3. Results

Two main objectives are pursued in this study. On the one hand, an assessment of the methodological approach for obtaining robust and unbiased results when modeling imbalanced, high-resolution accident data is aspired. This is reported in the first section related to model performance. On the other hand, the contributions of the numerous covariates under consideration are of interest, with a particular focus on the impact of real-time weather variables derived through meteorological re-analysis. This second aspect is described in the feature importance section.

3.1. Model performance

Given the severe class imbalance the proposed approach leads to decent results, with both random forests and boosted trees performing similarly well (Table 2). Using a threshold of 0.5 for class discrimination and majority vote, average results yield an overall accuracy of slightly above 80%, with every second accident being classified correctly. xgboost performs slightly better in terms of accuracy at the expense of sensitivity, but the differences are almost negligible. The extent of the severity of class imbalance is illustrated nicely by the fact that specificity is practically identical to accuracy. Values below 0.1 for precision and F_1-score stem from the vanishingly low proportion of event instances in the data set as well. Average AUC-values around 0.65 indicate moderate performance. However, the AUC is not overly informative in the face of substantial class and cost imbalance. Resulting *h*-measures are just above 0.1, indicating reasonable performance.

In terms of discriminative quality findings show that gradient boosted trees are superior to random forests (Fig. 2) as indicated by steeper discrimination slopes. This is mainly attributable to the better identification of true negative instances, which is clearly visible in the class probability distributions of the non-accident outcomes.

In the pursuit of exploring the trade-off between accuracy and sensitivity, different thresholds for class discrimination and the voting procedure are tested for the xgboost models. Since costs for false negatives are higher than for false positives in the task at hand (since overlooking an accident has substantially worse implications than inadvertently classifying a non-event as accident), sensitivity is considered a particularly decisive metric. Results nicely illustrate the trade-off between accuracy and sensitivity (Table 2). Both lowering the probability threshold for classifying as `TRUE` as well as changing the voting rule (i.e. replacing majority vote by an adjusted voting rule that classifies as `TRUE` if at least 40% of all ensemble members predict the outcome as `TRUE`) naturally has positive effects on sensitivity, but leads to decreasing accuracy. For instance, lowering both the class discrimination threshold and the threshold for the voting rule to 40% leads to an average sensitivity of way beyond 60%, with average accuracy still exceeding 70%.

3.2. Feature importance

Since tree-based models such as random forests and gradient boosted trees are nonlinear by nature, isolated feature effects cannot be properly assessed independently of the other variables. As opposed to additive models, feature importance measures do not signify whether variables have a positive or negative effect on the classification. Instead, they assess how much discriminatory information each variable[2] yields.

Overall feature importance metrics averaged over all single models illustrate a consistent picture between the two different statistical learning methods employed (Figs. 3 and 4). This becomes obvious when considering not only the mean ranking, but also confidence intervals of mean feature importance.

Unsurprisingly, traffic volume [`aadt`] clearly prevails as most important feature in both methods. Naturally, the number of accidents is proportional to the total number of cars. In addition, the share of heavy good vehicles [`hgv`] is also a very important variable in this context, particularly in random forest models.

Results show that certain weather variables are of high importance, including air temperature, air pressure, relative humidity, wind speed and direction, sunshine duration and precipitation totals (including the rainfall at the time of the accident as well as precipitation during the hour before and after the accident). If the three variants of precipitation are considered as one precipitation variable, this feature is the second most important variable after traffic volume. The proportion of solid precipitation as well as the surface condition (from a meteorological point of view) are are only of medium importance.

In addition, three variables related to climate conditions (the average number of hot days per year [`hot`], the average number of freeze-thaw-days per year [`frost`], and the average annual precipitation totals within the last 30 years [`rr`]) are ranked amongst the most important features.

Multiple features related to road geometrics (road bendiness [`bendiness`] and curvature [`curv_med`], medians of longitudinal [`s_med`] and transversal gradients [`q_med`], total width of the road [`w_tot`] as well as width of the breakdown lane [`w_bdl`]) and surface condition (from an engineering point of view, i.e. friction [`mu`], longitudinal evenness [`wlp`], surface roughness [`iri`]) manifest themselves to be of high importance. However, related features such as the surface construction type, the existence of tunnels and bridges, surface damage, water film depth and the existence of zero-crossings of transversal gradient have been found to be of marginal importance.

Features related to the time of the accident (weekday [`weekday`] and time of the day [`h_cat`]) seem to be only of medium importance, as are the number of lanes [`n_lanes`] and the existence of event sections (i.e. acceleration or deceleration lanes, junctions) [`event`]. With the exception of the maximum crack size [`cracks_max`], which is of medium importance as well, all features related to road surface damage are of minor importance.

As far as traffic regulations are concerned, only speed limits for cars [`speed_limit_car`] prevail as important feature, while overtaking bans [`no_overtaking_car`, `no_overtaking_heavy`] and speed limits for heavy goods vehicles [`speed_limit_heavy`] are of minor importance.

4. Discussion

Findings are partially consistent with results presented in literature. Specifically, the companion paper by Schlögl et al. (2019) does serve as a reference, since a very similar data set from the same country was used. In order to provide a comparative assessment of model quality,

[2] Please note that all feature names in this section are quoted in square brackets for better interpretability of the variable importance plots.

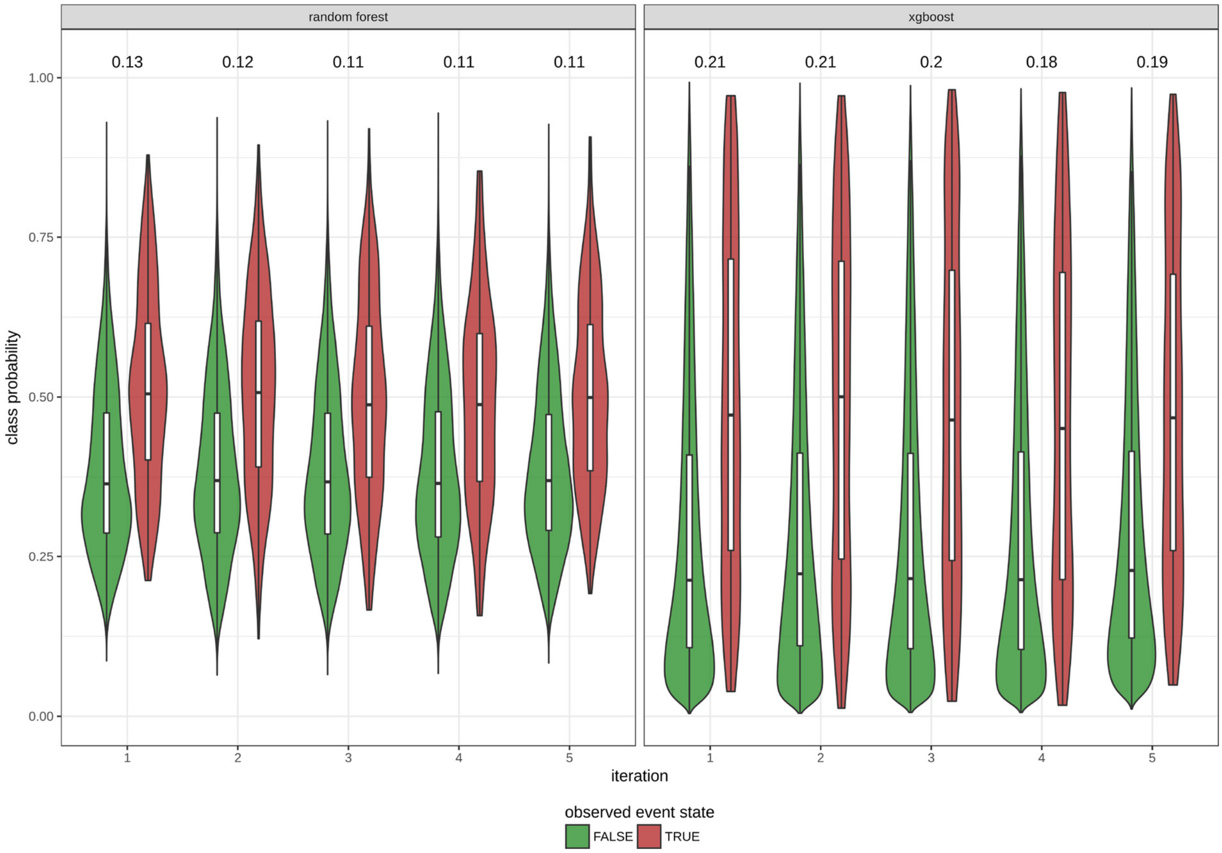

Fig. 2. Illustration of class discrimination. Results are based on the arithmetic mean of the predicted class probability calculated for each of the 22,362,484 test data instances across the 50-member ensembles. The number atop represents the discrimination slope.

results are contrasted to this congenial study and findings and discussed with respect to two different aspects, namely model performance and variable importance.

It is reasoned that the approach used in the study at hand is more robust for three main reasons. First, Schlögl et al. (2019) have used one fixed split into training (covering the years 2013 to 2015) and testing (comprising the year 2016) data. In the present study, a number of different splits were used to create data partitions for training and testing, thus leading to more robust results when averaging the resulting ensembles over all different splits. Second, Schlögl et al. (2019) trained a dozen different statistical learning methods once, with some of them performing quite well, and others exhibiting only mediocre performance. Here, the focus is on two selected tree-based methods which have been shown to perform particularly well when analysing high-dimensional, nonlinear data sets (Hastie et al., 2009; Schlögl et al., 2019). Third, instead of running every model just once, five ensembles featuring 50 model runs each were trained, hence reducing the variance of the resampling approach through bagging and thereby fostering the robustness of the results. Altogether, the repeated cross-validation procedure applied here is a robust approach that helps to avoid overoptimistic performance results (Schratz et al., 2019; Cawley and Talbot, 2010).

4.1. Model performance

First of all, results underline the importance of proper model performance evaluation. Conclusions solely drawn based on the training error (or other goodness-of-fit metrics obtained when evaluating the model on the training data set) are essentially meaningless. Both random forest and xgboost exhibit extremely low training errors across all ensembles (i.e. accuracies and AUC close to 1), which is somewhat inherent to the design of these methods. Sound model performance assessment needs to be grounded in model generalization capabilities, which are evaluated independently using out-of-sample error metrics. Since the bias-variance trade-off (and particularly the problem of overfitting) has been addressed with an extensive hyperparameter tuning step (based on 5-fold cross-validation) prior to model fitting, the seemingly mediocre performance metrics illustrate the formidable challenge of separation in the given feature space. It is presumed that this is attributable to (i) the extreme class imbalance ratio and (ii) the lack of discriminative information contained in certain features. Due to the fine spatial and temporal resolution, features of positive and negative instances might be very similar, if not nearly identical. In light of the complexity of the classification task at hand, the resulting model performance is respectable.

Due to the high spatial and temporal resolution of the data set and the inclusion of explanatory variables reflecting temporal and local spatial conditions, temporal and spatial autocorrelation was not explicitly considered in this modelling approach. Arguably, applying nested spatial cross-validation (Lovelace et al., 2019; Schratz et al., 2019) instead of repeated random *k*-fold cross-validation might lead do further bias-reduction. However, since autocorrelation does only become noticeable for some highways if data are aggregated substantially, this should be investigated further together with an assessment of (particularly temporal) aggregation levels.

Regarding model performance of the two tree-based methods under consideration, both random forests and gradient boosted trees are basically on par. There are only marginal differences visible in the resulting performance metrics. In this context it is argued that a combined consideration of accuracy and sensitivity constitutes reliable choice of performance metrics for imbalanced data. Other metrics such as precision, F_1-score and AUC should be treated with caution in the case of severe class and cost imbalance, as they might be of limited informative value. For the same reason, precision should be treated with care, as

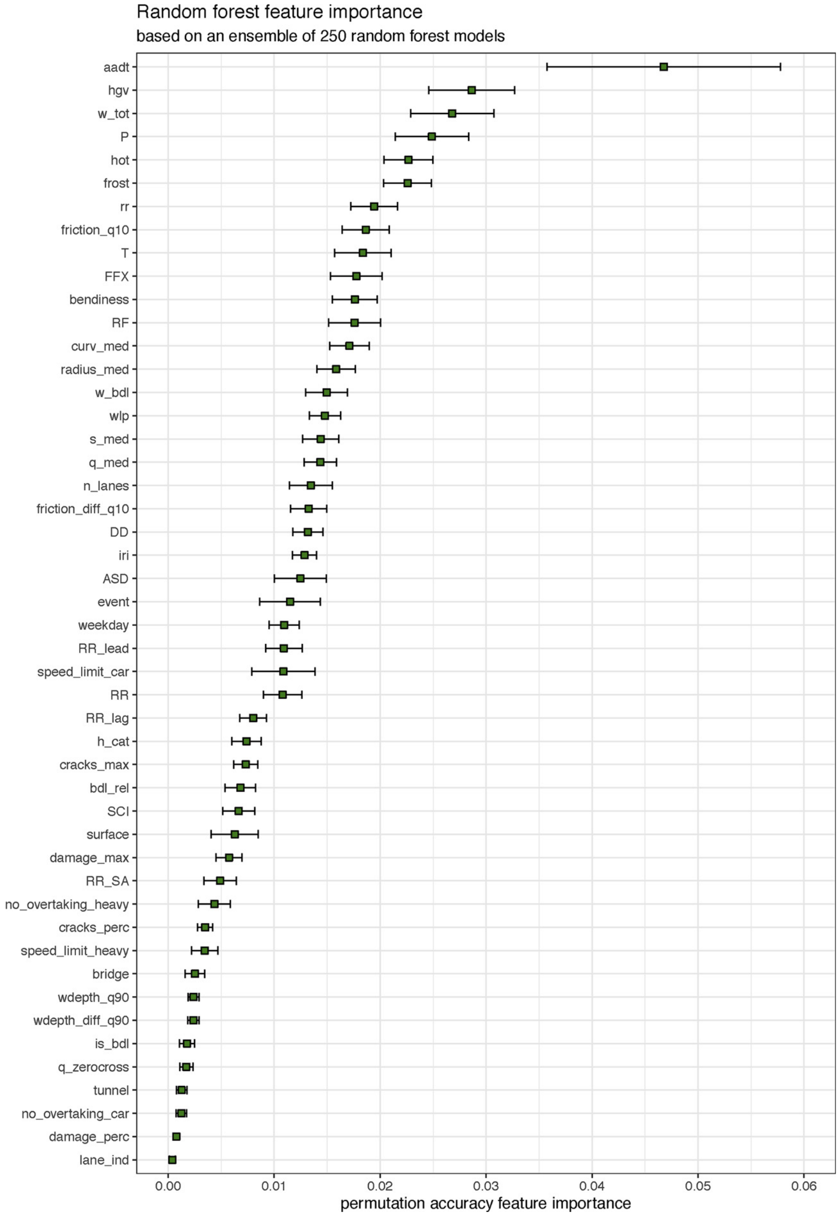

Fig. 3. Feature importance (permutation importance) averaged over 250 random forest models. Dots represent the ensemble mean feature importance, and errorbars indicate the corresponding two-sigma (≈ 95%) confidence interval.

this metric is highly sensitive to false positives.

The choice of voting rule and threshold selection for class discrimination is somewhat subject to personal preference and the task at hand. The choice of how to handle the trade-off between accuracy and sensitivity depends on the judgment of misclassification costs. If false negatives are considered a substantially severe outcome than false positives, lowering both values can prove to be beneficial for increasing sensitivity at the expense of overall accuracy. All instances predicted as `TRUE` can be considered as potentially accident-prone sections. If cost imbalance is not a major issue, default majority voting and a class discrimination threshold of 0.5 lead to reasonable results.

Due to the distribution of the predicted class probabilities, which indicate a more distinct separation between the two classes in the xgboost models, lowering the class discrimination threshold is comparably safe when employing gradient boosted trees. Caution is advised when changing the class discrimination threshold of random forest models, though (Fig. 5).

Concerning computational efficiency, random forests shine in terms of tuning and training, while xgboost excels in prediction. Overall computation time for tuning, training and prediction as implemented in this study is approximately on par as well.

Compared to Schlögl et al. (2019), model quality is slightly better in the present case across all performance metrics, which is especially remarkable in spite of the fact that the other study draws on a four times larger data basis (namely 5998 events versus 1560 accidents in the present case, at a comparable class imbalance ratio).

As opposed to the combination of synthetic minority oversampling and maximum dissimilarity undersampling, the proposed resampling approach using balanced bootstrap training samples is comparably simple, fast, and computationally less expensive, while showing no negative ramifications regarding model validity. Instead of introducing synthetic event class instances – a procedure which is subject to the condition that the synthetic set is generated entirely within the convex hull formed by the the minority class training points, which in turn has several implications on the prediction bias of the classifier (Bellinger et al., 2018) – balanced bagging does solely rely on actual observations. There are no prerequisites in terms of assumptions regarding distribution or properties of underlying data. This is particularly beneficial in

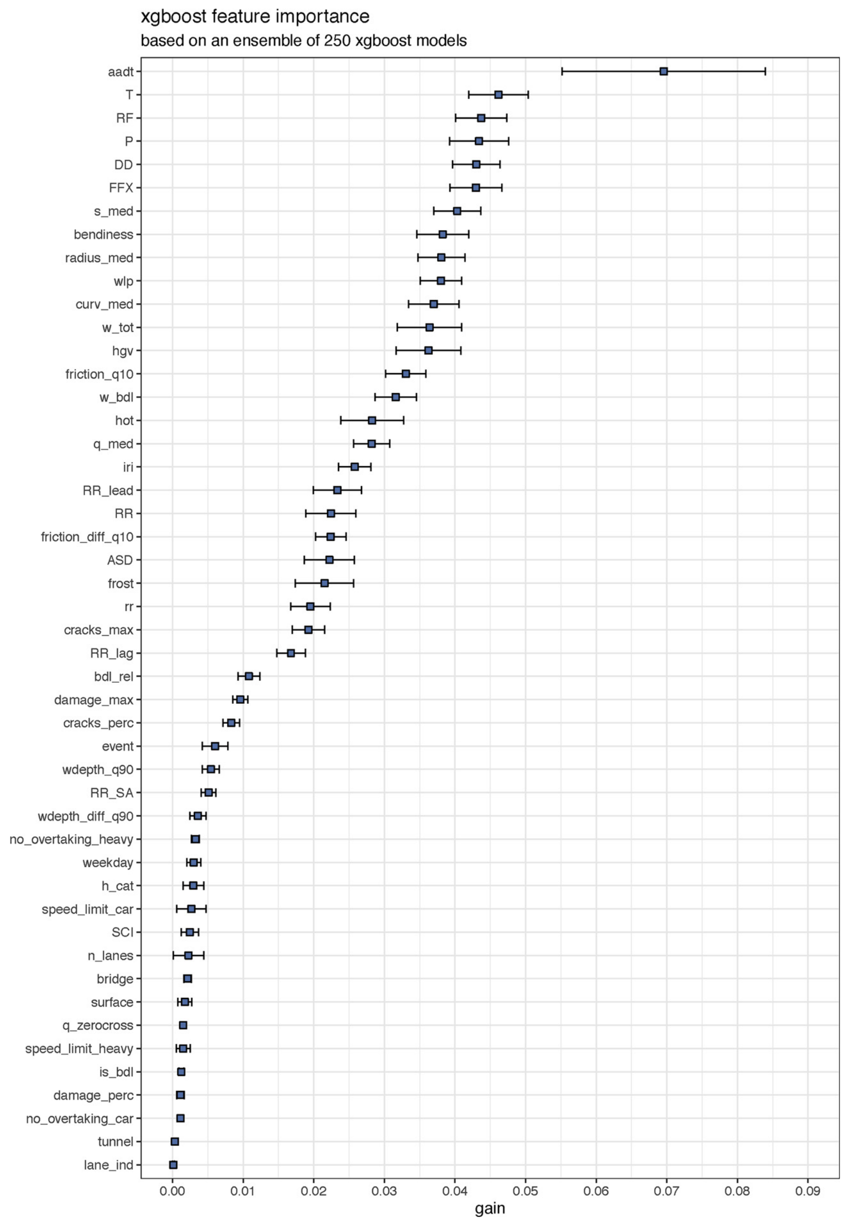

Fig. 4. Feature importance (gain) averaged over 250 xgboost models. Dots represent the ensemble mean feature importance, and errorbars indicate the corresponding two-sigma (≈ 95%) confidence interval.

the problem at hand, which is characterized by an extremely severe class imbalance ratio within a large data set. At the same time, the balanced bagging resampling approach clearly beats the combination of SMOTE and maximum dissimilarity undersampling in terms of computational efficiency. Consequently, obtaining results through majority vote across ensembles of models fitted to balanced, bootstrapped training data comes at comparably low cost, while offering to make use of the reduced variance properties inherent to bootstrap aggregating.

In addition, the approach is hallmarked by remarkable robustness, as indicated by almost identical results across all ensembles. This is illustrated both in the resulting performance metrics (Table 2) as well as the distributions of class discrimination (Fig. 2) and predicted probabilites (Fig. 5).

4.2. Feature importance

Results with respect to variable importance have to be interpreted with care, also when contrasting them to findings presented in Schlögl et al. (2019). The focus of both studies is on presenting a methodological framework for analysing severely imbalanced accident data sets, rather than on deriving general policy implications. Despite the size and comprehensiveness of the data set at hand, obtained results might not be universally transferable, especially when taking the current spatial and temporal coverage of the data set as well as the problem of unobserved heterogeneity into account.

In this context, it is important to emphasize that the focus of the companion paper by Schlögl et al. (2019) was primarily on comparing and contrasting different statistical learning methods for analysing imbalanced accident data sets using SMOTE. Therefore, the study focused on the predictive performance of different models rather than on inference. The major finding was that tree-based methods outperform more classical statistical approaches. In terms of assessing effects of environmental variables on road accident occurrence, only a mean variable importance derived by averaging the individual models' variable importance was presented in the congenial study. This is suboptimal for two main reasons. First, also models with low discriminatory power and subpar performance are included, thereby negatively affecting and distorting average feature importance. Second,

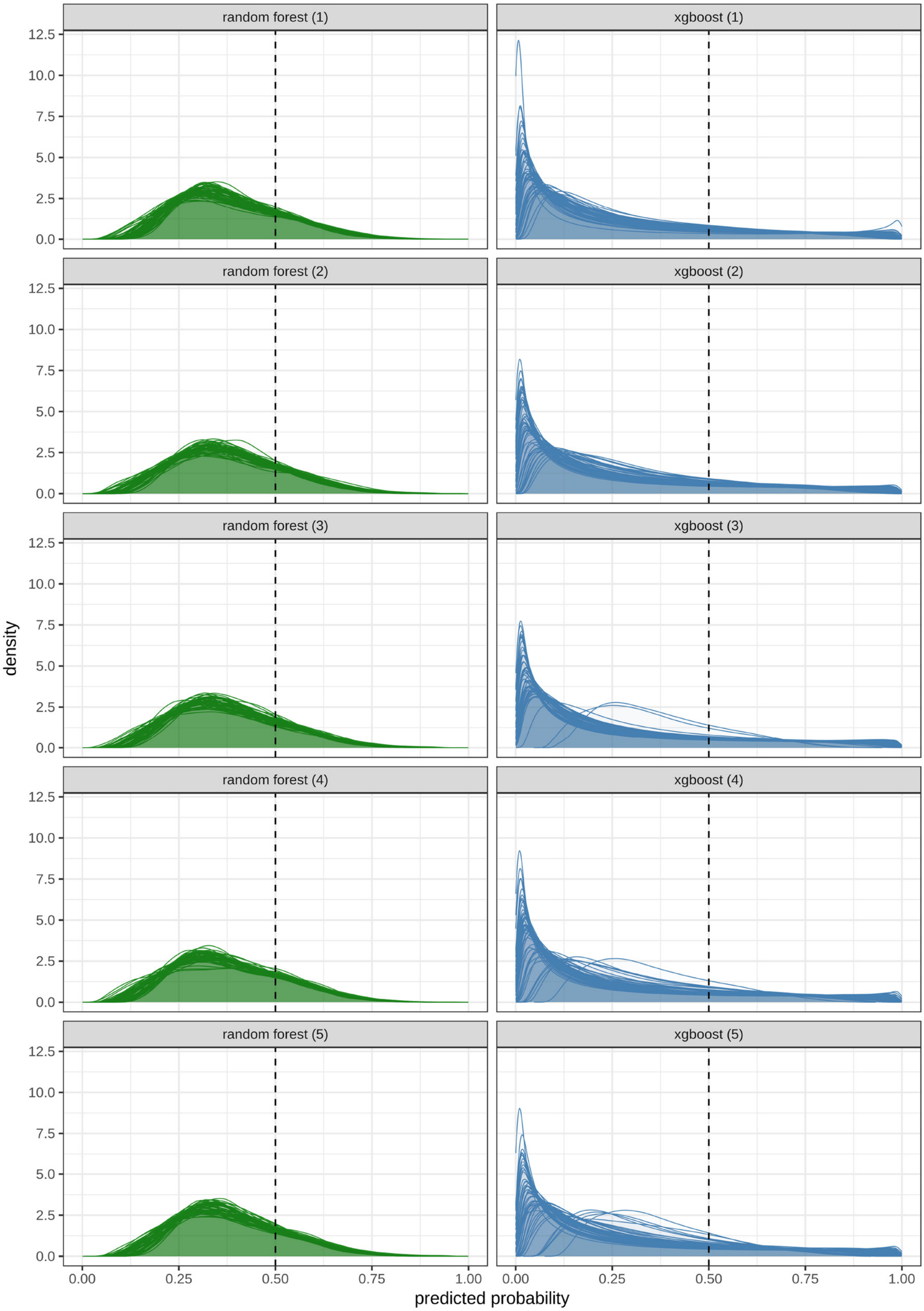

Fig. 5. Density plot of predicted class probabilities for class 'accident'. The dashed vertical line indicates the class discrimination threshold, i.e. the probability at which the positive class ('accident') is chosen over the negative class ('no accident').

no metric of uncertainty was provided, which might imply overoptimistic accuracy. These two aspects have been improved upon in this study.

While the focus of the current paper is also mainly methodological (in terms of assessing the performance of the balanced bagging approach), results with respect to feature importance show an important extension towards inference. The resampling procedure allows to compute confidence intervals based on the standard deviation of the feature importance metrics using all underlying ensemble members. Thereby, the uncertainty of the mean overall feature importance can be quantified. Results illustrate that the final feature importance ranking should not be assessed solely based on ensemble means, but also by considering the confidence intervals of these means as well. This entails that specific ranking should not be over-interpreted. Instead, groups of variables exhibiting similar ranks and confidence intervals should be contrasted.

This quantification of uncertainty highlights one major strength of the employed bagging approach. Knowledge about the uncertainty of the impacts of certain environmental effects on road accident occurrence is of major importance for interpretability and practical applicability.

Against this background, seemingly contradictory results between the two studies have to be reassessed.

Many features are consistent with respect to their importance (e.g. traffic volume, climate variables, a majority of geometric features), especially when scale and confidence intervals of the variable importance metrics are considered. Taking into account the associated uncertainty, results exhibit practically equal feature importance for many variables.

A number of features show strikingly different importance in both studies. To some extent, this may be attributable to the calculation of mean feature importance in Schlögl et al. (2019). Yet, three conspicuously diverging features protrude in the comparative assessment: Bridges and overtaking bans are of high importance in the other study, whereas they seem to be among the most unimportant variables under consideration in this study. The opposite is true for weather variables, which prevail as important features in this study, but are found to be of secondary importance in Schlögl et al. (2019).

Possible reasons for this divergence are threefold. First, models trained in this study are inherently more consistent and perform comparably. Since the main focus of the other study was providing a comparative assessment of the performance of statistical learning models in the context of unbalanced accident data, also models exhibiting subpar results are included. Second, variable importance is reported for models fitted only to one possible training data set. Albeit the use of multiple different models guarantees some validity, they are all trained on the same data basis. While this is practicable for a comparative assessment of different statistical learning techniques, the overall model quality cannot be assessed genuinely. Despite the use of a sophisticated resampling strategy, there is no guarantee that the selected training sample is representative of the whole data set. This is where the merits of using model ensembles trained on multiple different training partitions come into play, as single suboptimal training samples leading to inferior results are simply averaged out. Third, both inference and prediction from severely imbalanced data sets is subject to high uncertainty. Considering the lack of variation within predictor variables, the task of fitting models is extremely difficult. Therefore, the possibility that certain differences arise simply due to the use of two different time periods for training (2013–2015 vs 2016) cannot be ruled out. Once again, this supports the usefulness of the balanced bagging approach.

In addition, discrimination between different target accident groups could facilitate deriving important features more distinctly and foster gaining a deeper understanding of accident contributing factors. Since different explanatory variables contribute to different accident patterns (e.g. run-off-road accidents, rear-end collisions, involvement of powered two wheelers, etc.) in different ways, a differentiation might be beneficial. However, this obviously further aggravates the problem of class imbalance.

It should be noted that the existence of collinear features might lead to an underestimation of variable importance of affected features. However, since the number of variables sampled at each split can be some considered as a sort of regularization parameter (which is tuned in the cross-validation process), tree models are robust in the face of correlated predictors, especially if the main focus of the work is prediction rather than inference. Naturally, multi-collinearity was checked beforehand, with a majority of variables being uncorrelated. The only noteworthy correlation in this respect is related to precipitation and its lagged or leading variants respectively (Pearson's correlation coefficient: $r = 0.45$). Albeit this has no effect on the predictive performance of the model, the importance of rainfall reported in the variable importance plots might be slightly underestimated, which could affect the ranking of given the close values for feature importance. Nevertheless, removing both shifted precipitation values has no effect in terms of variable importance, after all.

Finally, despite the inclusion of a large number of explanatory variables, not all features that could potentially determine the probability of traffic accident occurrence are included in the presented models. Lack of information caused by this ‘unobserved heterogeneity’ (Mannering et al., 2016) could entail biased model results. Even though this study seeks to minimize possible heterogeneous effects by (i) including a large number of potentially relevant explanatory variables and (ii) using highly disaggregated input data (both temporally and spatially), it has to be kept in mind that potentially relevant characteristics affecting accident occurrence that are not covered in the present data set do certainly exist.

5. Conclusion

This paper presents a robust approach for analyzing high-resolution road-safety data. In addition to classical data on traffic volume, road geometrics and road surface condition, special emphasis was put on weather variables obtained through meteorological re-analysis. The methodological approach on how to overcome the problem of severe class imbalance inherent to such binary classification tasks is illustrated at the example of a data set featuring 48 covariates, covering the whole highway network of Austria.

This study also lays the foundation for future research in this area. In particular, spatial and temporal autocorrelation of accidents could be interesting to explore in future work. The benefits of nested spatial cross validation over repeated random *k*-fold cross-validation could be assessed together with an assessment of different (temporal) aggregation levels.

All things considered, it is argued that the balanced bagging approach is a statistically sound machine learning ensemble meta-algorithm that is preferable to potentially more complex approaches to handle class imbalance. The robustness of the results that can be obtained through this approach is of particular importance in the face of very small shares of event instances.

Acknowledgments

I would like to thank my colleague Rainer Stütz for providing insight and expertise that supported this research. Thanks to Claus Aichinger, Astrid Berg, Michael Kammer, Christoph Matulla and Michael Melcher for fruitful statistical discussions. In addition, I gratefully acknowledge ASFiNAG for providing traffic volume data and for approving the use of RoadSTAR measurement data.

Appendix A. Overview of variables

Table A.1

Table A.1
Overview of explanatory variables used in this study. The *variable name* denotes the variable tag used in the data set, *data type* indicates the data type from a computer science point of view, and *variable description* provides a short explanation of the respective variable. The last column *unit* contains the physical unit of the variable. Note that the unit [1] indicates dimensionless scalar quantities (e.g. dimensionless indices), while – indicates factor variables where no units apply, and # indicates count data (i.e. number of items).

Variable name	Data type	Variable description	Unit
T	Numeric	Air temperature	[° C]
RR	Numeric	Accumulated precipitation	[mm]
RR_lead	Numeric	Precipitation (one hour ahead)	[mm]
RR_lag	Numeric	Precipitation (one hour behind)	[mm]
RR_SA	Numeric	Percentage of solid precipitation	[%]
SCI	Factor	Surface condition index, with factor levels dry, damp, wet, snowy and glaze	–
RF	Numeric	Relative humidity of the air	[%]
P	Numeric	Air pressure	[hPa]
FFX	Numeric	Wind gusts (maximum wind speed)	[m/s]
DD	Integer	Wind direction	[°]
ASD	Integer	Absolute sunshine duration	[min]
rr	Double	Average precipitation totals per year	[mm]
hot	Double	Average number of hot days per year	[d]
frost	Double	Average number of freeze-thaw-days per year	[d]
weekday	Factor	Weekday, reference class is sunday	–
h_cat	Factor	Time of day, separated into five levels: 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); reference class is night	–
aadt	Integer	Annual average daily traffic volume	[#]
hgv	Integer	Annual average traffic volume of heavy goods vehicles	[#]
w_tot	Double	Total width of the road	[m]
w_bdl	Double	Width of breakdown lane	[m]
is_bdl	Logical	Indicator for the existence of a breakdown lane	–
bdl_rel	Double	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	Double	Median of transverse gradient	[%]
q_zerocross	Logical	Indicator for a zero-cross (i.e. change of sign) of the transverse gradient	–
s_med	Double	Median of longitudinal gradient	[%]
radius_med	Double	Median of radius	[m]
curv_med	Double	Median of curvature, which is the reciprocal of the radius	[m^{-1}]
bendiness	Double	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	[gon]
friction_q10	Double	0.1 quantile of the coefficient of friction	[1]
friction_diff_q10	Double	Difference between median and 0.1 quantile of the coefficient of friction, i.e. an indicator for changes in friction within one segment	[1]
iri	Double	Median of the international roughness index (Sayers and Karamihas, 1998)	[1]
wlp	Double	Median of changes in the weighted longitudinal profile, i.e. an indicator for longitudinal evenness (Ueckermann and Steinauer, 2008)	[mm]
wdepth_q90	Double	Median of waterfilm thickness, i.e. an indicator for rut depth	[mm]
wdepth_diff_q90	Double	Difference between 0.9 quantile and median of waterfilm thickness, i.e. an indicator for changes of rut depth within a segment	[mm]
surface	Factor	Material type of road surface, levels are asphalt, concrete and tms (thin membrane surface); reference class is asphalt	–
cracks_perc	Double	Percentage of cracks on total road surface area	[%]
cracks_max	Double	Maximum damage class of cracks	[m]
damage_perc	double	percentage of road surface damage on total road surface area	[%]
damage_max	Double	Maximum damage class of road surface damage	[m^2]
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 busses, levels are 80, 60 and 40, reference class is 80	–
no_overtaking	Logical	General 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	–

References

Abdel-Aty, M., Ekram, A.-A., Huang, H., Choi, K., 2011. A study on crashes related to visibility obstruction due to fog and smoke. Acc. Anal. Prevent. 43, 1730–1737. https://doi.org/10.1016/j.aap.2011.04.003.

Abdel-Aty, M.A., Hassan, H.M., Ahmed, M., Al-Ghamdi, A.S., 2012. Real-time prediction of visibility related crashes. Transport. Res. Part C: Emerging Technol. 24, 288–298. https://doi.org/10.1016/j.trc.2012.04.001.

Al-Ghamdi, A.S., 2007. Experimental evaluation of fog warning system. Acc. Anal. Prevent. 39, 1065–1072. https://doi.org/10.1016/j.aap.2005.05.007.

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

Antoniou, C., Yannis, G., Katsohis, D., 2013. Impact of meteorological factors on the number of injury accidents. In: Proceedings of the 13th World Conference on Transportation Research. 15-18 July, 2013, Rio de Janeiro, Brazil.

Baker, C., Reynolds, S., 1992. Wind-induced accidents of road vehicles. Acc. Anal. Prevent. 24, 559–575. https://doi.org/10.1016/0001-4575(92)90009-8.

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: Emerging Technol. 86, 202–219. https://doi.org/10.1016/j.trc.2017.11.014.

Becker, N., Pardowitz, T., Ulbrich, U., 2018. Modelling probabilities of weather-related road accidents. Geophys. Res. Abstracts 20 EGU2018-4497.

Bellinger, C., Drummond, C., Japkowicz, N., 2016. Beyond the boundaries of smote. In: Frasconi, P., Landwehr, N., Manco, G., Vreeken, J. (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, Cham, pp. 248–263.

Bellinger, C., Drummond, C., Japkowicz, N., 2018. Manifold-based synthetic oversampling with manifold conformance estimation. Mach. Learn. 107, 605–637.

https://doi.org/10.1007/s10994-017-5670-4.

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

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. URL: https://www.bmvit.gv.at/service/publikationen/verkehr/strasse/statistik_strasseverkehr.html.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655.

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. Acc. Anal. Prevent. 40, 1180–1190. https://doi.org/10.1016/j.aap.2008.01.001.

Cawley, G.C., Talbot, N.L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079–2107.

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., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16. New York, NY, USA: ACM, doi: 10.1145/2939672.2939785. pp. 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., 2018b. xgboost: extreme gradient boosting. URL: https://CRAN.R-project.org/package = xgboost r package version 0.6.4.1.

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. Acc. Anal. Prevent. 36, 637–647. https://doi.org/10.1016/S0001-4575(03)00085-X.

Eisenberg, D., Warner, K.E., 2005. Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. Am. J. Public Health 95, 120–124. https://doi.org/10.2105/AJPH.2004.048926.

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. Acc. Anal. Prevent. 27, 1–20. https://doi.org/10.1016/0001-4575(94)E0023-E.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G., 2017. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. 73, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035.

Hand, D., Anagnostopoulos, C., 2014. A better Beta for the H measure of classification performance. Pattern Recognit. Lett. 40, 41–46. https://doi.org/10.1016/j.patrec.2013.12.011.

Hand, D.J., 2009. Measuring classifier performance: a coherent alternative to the area under the ROC curve. Mach. Learn. 77, 103–123. https://doi.org/10.1007/s10994-009-5119-5.

Hassan, H.M., Abdel-Aty, M.A., 2013. Predicting reduced visibility related crashes on freeways using real-time traffic flow data. J. Safety Res. 45, 29–36. https://doi.org/10.1016/j.jsr.2012.12.004.

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.

Hermans, E., Brijs, T., Stiers, T., Offermans, C., 2006. The impact of weather conditions on road safety investigated on an hourly basis. In: Proceedings of the 85th Annual meeting of the Transportation Research Board. January 22-26, 2006, Washington, D.C.

Heuel, S., Straumann, R., Schüller, H., Keller, U., 2014. Einflüsse des Wetters auf das Strassenunfallgeschehen [influences of weather on traffic accidents]. https://www.astra.admin.ch/astra/de/home/dokumentation/unfalldaten/publikationen/forschungspaket-vespa.html. Accessed: 2018-08-10.

Jaroszweski, D., McNamarau, T., 2014. The influence of rainfall on road accidents in urban areas: A weather radar approach. Travel Behav. Soc. 1, 15–21. https://doi.org/10.1016/j.tbs.2013.10.005. Advances in Spatiotemporal Transport Analysis.

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, 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.

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. Progr. Artificial Intelligence 5, 221–232. https://doi.org/10.1007/s13748-016-0094-0.

Lovelace, R., Nowosad, J., Muenchow, J., 2019. Geocomputation with R. CRC Press.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analyt. Methods Acc. 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. Analyt. Methods Acc. Res. 11, 1–16. https://doi.org/10.1016/j.amar.2016.04.001.

Maurer, P., Meissner, M., Fuchs, M., Gruber, J., Foissner, P., 2002. Straßenzustandserfassung mit dem RoadSTAR - Messsystem und Genauigkeit.

Maze, T., Agarwai, M., Burchett, G., 2006. Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. Transport. Res. Record: J. Transport. Res. Board 1948, 170–176. https://doi.org/10.3141/1948-19.

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. Software 51, 1–55. https://doi.org/10.18637/jss.v051.i01.

Sayers, M., Karamihas, S., 1998. Little Book of Profiling.

Schlögl, M., and Stütz, R. (2017). Methodological considerations with data uncertainty in road safety analysis. Accident Analysis & Prevention in press. doi: 10.1016/j.aap.2017.02.001.

Schlögl, M., Stütz, R., Laaha, G., Melcher, M., 2019. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. Acc. Anal. Prevent. 127, 134–149. https://doi.org/10.1016/j.aap.2019.02.008.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., Brenning, A., 2019. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecol. Model. 406, 109–120. https://doi.org/10.1016/j.ecolmodel.2019.06.002.

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

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. Monthly Weather Rev. 134, 2758–2771. https://doi.org/10.1175/MWR3196.1.

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. Acc. Anal. Prevent. 72, 244–256. https://doi.org/10.1016/j.aap.2014.06.017.

Ueckermann, A., Steinauer, B., 2008. The weighted longitudinal profile. Road Mater. Pavement Design 9, 135–157. https://doi.org/10.1080/14680629.2008.9690111.

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.

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. Acc. Anal. Prevent. 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.

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. Acc. Anal. Prevent. 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, 19-03414.