Risk mapping of wildlife–vehicle collisions across the state of Montana, USA: a machine-learning approach for imbalanced data along rural roads

Matthew Bell, Yiyi Wang and Rob Ament

Transportation Safety and Environment, 2024, **6**: tdad043
DOI: 10.1093/tse/tdad043
Advance access publication date: 30 November 2023
Research Article

Risk mapping of wildlife–vehicle collisions across the state of Montana, USA: a machine-learning approach for imbalanced data along rural roads

Matthew Bell[1], Yiyi Wang[2,*] and Rob Ament[1]

[1]Western Transportation Institute, Montana State University, 2327 University Way, Bozeman, MT 59717, USA;
[2]School of Engineering , San Francisco State University, 1600 Holloway Ave, San Francisco, CA 94132, USA.
*Corresponding author. E-mail: yiyiwang@sfsu.edu

Abstract

Wildlife–vehicle collisions (WVCs) with large animals are estimated to cost the USA over 8 billion USD in property damage, tens of thousands of human injuries and nearly 200 human fatalities each year. Most WVCs occur on rural roads and are not collected evenly among road segments, leading to imbalanced data. There are a disproportionate number of analysis units that have zero WVC cases when investigating large geographic areas for collision risk. Analysis units with zero WVCs can reduce prediction accuracy and weaken the coefficient estimates of statistical learning models. This study demonstrates that the use of the synthetic minority oversampling technique (SMOTE) to handle imbalanced WVC data in combination with statistical and machine-learning models improves the ability to determine seasonal WVC risk across the rural highway network in Montana, USA. An array of regularized variables describing landscape, road and traffic were used to develop negative binomial and random forest models to infer WVC rates per 100 million vehicle miles travelled. The random forest model is found to work particularly well with SMOTE-augmented data to improve the prediction accuracy of seasonal WVC risk. SMOTE-augmented data are found to improve accuracy when predicting crash risk across fine-grained grids while retaining the characteristics of the original dataset. The analyses suggest that SMOTE augmentation mitigates data imbalance that is encountered in seasonally divided WVC data. This research provides the basis for future risk-mapping models and can potentially be used to address the low rates of WVCs and other crash types along rural roads.

Keywords: wildlife–vehicle collisions; machine learning; SMOTE; negative binomial; risk mapping

Key messages

- Wildlife carcass locations were aggregated along road segments along with the environmental attributes across Montana.
- The synthetic minority oversampling technique (SMOTE) mitigates imbalanced data and produces consistent estimates between the models.
- SMOTE is more effective at increasing model accuracy for seasonal data than for full-year data.

1. Introduction

For more than two decades, increasing attention has been paid to the number of wildlife–vehicle collisions (WVCs) in the USA, and their consequences and potential solutions [1–3]. The most recent nationwide WVC study showed that over one million crashes with large-bodied animals occurred every year in the USA, costing an annual sum of 8.4 billion USD in property damage, human injury/death, economic loss of animal life and other maintenance costs (e.g. carcass removal) [2]. WVCs are prevalent on rural highways and particularly along low-volume rural roads where the WVC crash rate is eight times higher than on other road types [4].

WVC data are unique from other crash data: they are not systematically collected by transportation agencies and suffer from under-reporting [5]. Two types of data are commonly used to identify road segments with high rates of WVCs in the USA—data on animal carcasses, which are collected by transportation maintenance personnel, sometimes with the assistance of citizen scientists, and data on traffic collisions that are reported to law enforcement [6]. The Montana Department of Transportation (MDT) counted 60,858 animal carcasses along system routes within the state from 2008 to 2017, in stark contrast to the 23,718 WVCs reported by Montana Highway Patrol during the same period [7]. Regardless of the data type, it is estimated that WVCs are severely underreported across the entire USA [8]. This issue is more prevalent in rural areas where there is even less data [9] and WVCs are more likely to occur.

Mitigation measures reduce WVC risk to varying degrees, including wildlife over- or underpasses, wildlife warning signage, fencing and animal detection driver warning systems. Transportation engineers seek to reliably predict high-risk locations of WVCs to decide where to implement mitigation measures to increase safety for the travelling public. Many examples for identifying road segments with high WVC rates or 'hot spots' are reviewed in Pagany [10]. However, rural roads face unique challenges when it comes to hot-spot identification because WVCs are not systematically collected, nor are existing methods tailored for practical use by transportation agencies in rural settings.

Received: March 2, 2023. **Revised:** May 31, 2023. **Accepted:** September 10, 2023

Many studies have attempted to portray where WVCs are spatially clustered but confined their analyses to select corridors or small geographic areas [11–15]. These studies use non-parametric techniques such as the Kernel Density Estimation (KDE) and the Nearest Neighbor Index (NNI) to identify the probability density of WVC locations. These analyses suggest that WVCs are not a random occurrence and are spatially clustered based on an assortment of road, landscape and human variables [16].

Predictive models can be used to anticipate high-risk WVC locations while controlling for landscape and infrastructure variables [13, 15, 16]. This method involves a variety of statistical models to identify variables related to WVC prevalence at the road segment or areal level. Logistic regression models have been used to explain WVC likelihood after controlling for pertinent variables [11, 12, 15, 17]. Poisson, negative binomial (NB) and zero-inflated NB regression models have been used to inform the expected WVC counts over a pre-specified time [18–20]. The NB-related approach works particularly well for over-dispersed count outcomes, where most road segments have zero collisions.

Data augmentation techniques focus on improving the balance between the minority and majority class in the dataset. The ensuing, balanced distribution mitigates overfitting issues and reduces biases that are in favour of the majority class because it enables prediction models to adopt a more generalized decision region [21]. Since data augmentation is done before model calibration, it can be used with any prediction models. Among the data augmentation tools, the synthetic minority over-sampling technique (SMOTE) outperforms other variable sampling techniques (e.g. varying the loss ratios, as used in Ripper or Naive Bayes Classifier) [22, 23]. The original SMOTE handles imbalanced binary data by over-sampling the rare, minority class (e.g. WVC) and under-sampling the majority class (i.e. no WVC). SMOTE has enjoyed various applications in traffic safety studies as well as in WVC studies [10, 22, 24].

2. Objectives

Our study used a machine-learning framework to model WVC crash rates when crash and carcass data were weak or lacking, a common issue plaguing rural areas. Crash rate is defined as the expected number of carcasses per 100 million vehicle miles travelled (VMT) in each analysis unit. We employed an extension of SMOTE to anticipate crash risk using carcass data across MDT's system roads in Montana, USA. This research aimed to address weak WVC data by: 1) integrating SMOTE, a data augmentation technique, into classic statistical and machine-learning models; and 2) evaluating the reliability of SMOTE in terms of coefficient estimates and out-of-sample prediction. Our study demonstrated how data augmentation can be used jointly with popular machine-learning methods to improve WVC hot-spot predictability and add to some early explorations to do the same [25–27]. This research offers an easy-to-use alternative for agencies with limited resources to screen their road network for high-risk WVC locations.

3. Methods

This section explains the main features of the methods selected by this study, why they were chosen and how they were used to predict year-round and seasonal risk of WVCs, in terms of expected number of wildlife carcasses per 100 million VMT. The following steps were used to conduct the analysis in this research: 1) create maps (e.g. road network, wildlife carcass data, environmental layers, etc.) and divide Montana on-system routes into analysis units to summarize data variables; 2) create SMOTE data to run in tandem with original data set; 3) use lasso regularization to identify most influential variables, which were then fed into NB models; 4) compare the best-performing NB models with random forest (RF) models to identify most important variables; 5) choose the more predictive model to estimate yearly and seasonal collision rates for all analysis units; 6) plot predicted rates on the state map of Montana. All analyses were performed in R [28].

3.1. Data

We employed ArcGIS Desktop 10.6 to compute the geospatial variables based on open-source shapefiles from state and federal online archives [29]. Mesh grids were defined along Montana's on-system routes as the analysis units. First, intersections were separated from mid-block road segments. A 76 m buffer is commonly used to define intersections and their influence areas [30]. Then, the remaining roads were divided into segments averaging 1 km in length. The choice of a 1 km segmentation is also supported by prior research [31, 32]. Six years of carcass data collected by MDT from 2011 to 2016 were joined to each road segment in the analysis units. Carcass data was collected by MDT's maintenance crew in an opportunistic approach based on staffing and resource. The process is described in more detail in the MDT Maintenance Environmental Best Management Practices Manual. A 400 m buffer was cast off each side of the road segments to calculate the environmental factors around the road segments. The 400 m buffer size was small enough to reflect the changing environment from one segment to another, yet large enough to catch the diversity of the surrounding environment [33].

Variables were aggregated by their averages and the distance to or counts within each analysis unit. For instance, the percentage of area covered by a habitat type was used to normalize the different sizes of analysis units. Distance-based variables (i.e. distance to open water) were computed as the Euclidean (straight-line) distance from the centroid of each analysis unit to the nearest variable of interest, as used commonly in safety and ecological analysis [17, 19, 34]. Distance to water was used to explain potential increase in collision risk because wildlife frequent areas that provide water. The speed limit along road segments was averaged for each analysis unit to represent the average travel speed. Habitat types served good predictors of the wildlife species that use and occupy the area. The nearness of habitat types to the roadway can explain the increase in exposure and was considered here. Landscape features like a flat slope and dense canopy cover may attract wildlife for food and protection, hence increasing their chance encounter with motor vehicles. Most of these variables have been employed in prior research as possible factors influencing WVC risk.

Additional variables were also investigated, such as large mammals' prevalence, defined as the potential number of species present in the analysis area and tabulated by overlapping species range maps from the Montana Department of Fish, Wildlife and Parks [35]. Canopy area measured the percentage of an analysis unit that has an overstory forest canopy cover. Canopy density defined the average forest canopy coverage within an analysis unit. Vegetation diversity counted the total number of plant species in each analysis unit per the United States Geological Survey's (USGS) land use maps [36]. Vegetation complexity measured the mixing among different plant species in an analysis unit. Both diversity and complexity were considered because vegetation complexity tends to increase as plant species become more interspersed in an area even though its vegetation diversity remains the same.

3.2. SMOTE for regression

SMOTE addresses imbalanced data by creating synthetic data points representing the minority response (i.e. an occurrence of a WVC). This technique over-samples the rare, minority response and under-samples the majority response, i.e. zero WVC. To avoid duplicate data points, SMOTE draws a random ratio from 0 to 1 to calculate a synthetic data point as a weighted average between the two neighbouring variables in the feature space. It outperforms other variable sampling techniques (e.g. varying the loss ratios that are used in the Ripper or Naive Bayes Classifier) [22–24]. SMOTE has been used to estimate vehicle collisions by treating collisions as a binary outcome [22, 24, 37].

Prior work largely used SMOTE on binary response (zero or non-zero WVC), but less understood is how that technique performs when used with regression models that explain numerical crash rates (e.g. expected number of wildlife carcasses per VMT) to inform the magnitude of crash risk. Our study applies an extension of SMOTE that generates synthetic data points for such numerical crash rates [21]. Equal weights were applied to all non-zero WVC analysis rather than investigating the appropriate value to associate with each minority event. The event ratio used in SMOTE describes the ratio between majority (i.e. zero WVC) and minority (i.e. non-zero WVC) events. In the augmented SMOTE data, an event ratio resulting in an equal distribution (50:50) between the minority and majority classes was selected, compared to the original event ratio of 3:10. This ratio was selected as an experiment with using SMOTE for WVC analysis. Varying event ratios and weight schemes can be explored in additional research to emphasize sites that have historically higher WVC rates, but this is outside the scope of this research. SMOTE data was created using the following steps: 1) randomly sample the majority event (i.e. zero WVC) to reach the amount in the selected event ratio; 2) select all the rare events (i.e. non-zero WVC); and 3) create additional synthetic data until the event ratio is reached. For each additional synthetic data point, two rare events are selected and a random ratio is drawn between 0 and 1. The value of the created synthetic data is the ratio of the value between the original variable inputs of the two rare events.

3.3. Negative binomial models with L1 regularization

NB regression models are commonly used to analyze collision counts [19, 20]. The NB model handles over-dispersed count outcomes, like the carcass counts in our study, where the standard deviation (3.87) exceeds the mean (1.37). An exposure factor is computed as VMT per analysis unit to control for the size effect.

After standardizing variables and removing those that exhibit severe co-linearity, using Pearson correlation coefficients, the candidate variables were entered into a NB model with the Least Absolute Shrinkage and Selection Operator (Lasso) regularization. The Lasso regularization removed redundant variables, preventing overfitting through a Lagrange multiplier that maximizes the log-likelihood function of the NB model [38]. A string of nested models were created and compared in detail using Pseudo R-squared (R^2) and out-of-sample prediction error expressed as the root mean-squared error (RMSE) and the mean absolute error (MAE).

3.4. Regularized random forest regression

RF was chosen among many machine-learning algorithms because it is known for detecting patterns in complex datasets and is the most frequently used by traffic safety studies [24, 39, 40]. An RF algorithm draws multiple samples with replacement from the data via bootstrapping, or bagging, and builds a decision tree for each sample. Within each tree, a random subset of predictors (i.e. features) is used to find the best predictor to split a tree node until the tree has met the maximum depth of nodes or other termination criteria. When used for prediction, the RF algorithm aggregates the predictions made by the decision trees like an expert opinion system does—either averaging the individual trees' predictions (for regression) or using the majority votes from individual trees (for classification). Each tree uses an out-of-bag sample of data to measure the predictive power of the individual model. Thanks to bootstrapping and random feature selection, an assembly of non-correlated trees improves the overall prediction accuracy.

This study evaluated how much SMOTE-augmented data altered, if at all, coefficient estimates and prediction performance, when applied to popular crash prediction models. RF was selected among a slew of machine-learning algorithms because of its popularity and performance for regression analysis [39, 41].

4. Results

Overall, 80% or 18,684 analysis units of the total observations were randomly drawn and used for the training data, of which 5792 (31%) contained at least one carcass location. This study set a 50:50 event ratio for SMOTE. Under that ratio, half of the data points contain at least one carcass location (*mean* = 2.27, *std. dev.* = 4.7). As shown in Table 1, SMOTE data mimics the distribution of the original data. The largest increase in the mean was noticed in the response variable. This is expected, as SMOTE created more artificial road segments with non-zero WVC counts that resemble neighbouring events. A full summary of the variables used in the analysis is presented in Table 1.

4.1. Lasso regularization

As the lambda (λ) penalty term in the Lasso regularization equation increases, it imposes a larger penalty on redundant coefficients. A λ value of 0.003 was chosen as it represents the tipping point at which removal of a variable causes a sharp rise in RMSE. The top 20 most influential variables identified in the SMOTE-augmented and original datasets by Lasso regularization shared 18 of the same predictor variables. Because the two datasets show almost identical patterns in variable selection, this confirmed that the SMOTE data stay true to the shape and structure of the original data.

4.2. Cross-validation of the year-round WVC models

A K-fold cross-validation method was employed to compare the prediction performance among candidate models. Standard of practice usually assumed 5 or 10 for the *K* value [42]. As seen in Table 2, the NB model had the lowest MAE for both the original and SMOTE data when compared to the linear regression and Bayesian model. This result prompted us to choose the NB model for the subsequent analysis.

4.3. Negative binomial models

The coefficient estimates in Table 3 are from the best-performing model out of all nested models tested with both datasets. The

Table 1. Summary of statistics for original (OR) and SMOTE (SM) data (No. of obs. = 18,684).

Category	Variable	Minimum		Maximum		Median		Mean		SD	
		OR	SM	OR	SM	OR	SM	OR	SM	OR	SM
WVC	Carcass counts	0	0	99	99	0	0	1.37	2.27	3.87	4.7
Traffic	Speed limit (mph)	15	15	80	80	70	70	64.14	65.4	13.97	13.1
	AADT (veh/day)	10	10	44,924	44,924	945	1395	3049	3391	5191	5179
	Road width (ft)	14	14	96	96	29	30	32	32.7	10.82	10.8
Road density	Arterial road (mi)	0	0	2.7	2.7	0	0.28	0.34	0.4	0.37	0.37
	Collector road (mi)	0	0	2.2	2.2	0	0	0.29	0.25	0.37	0.37
	Interstate road (mi)	0	0	2.48	1.96	0	0	0.15	0.19	0.4	0.44
	Local road (mi)	0	0	14.28	14.28	0	0	0.41	0.4	1.23	1.14
	Rural road (mi)	0	0	5.94	5.94	0	0	0.26	0.3	0.46	0.49
Infrastructure	Dist. to bridge (mi)	0	0	31.99	31.99	1.48	1.31	3	2.7	3.9	3.61
	Number of bridges	0	0	7	7	0	0	0.14	0.16	0.46	0.48
	Dist. to rest area (mi)	0	0	108.6	108.3	27.68	26.2	29.72	28.5	18.59	18.3
	Number of rest areas	0	0	2	2	0	0	0.003	0	0.07	0.07
Landscape	Large mammals	1	1	11	11	4	5	4.74	7	2.04	2.07
	Dist. to water (mi)	0	0	7.27	7.27	0.33	0.31	0.5	0.49	0.59	0.6
	Elevation (ft)	1866	1902	10,187	10,187	3367	3450	3605	3679	1071	1078
	Average slope (degree)	0	0	53	53	3	4	5.2	5.66	5.56	5.88
	Canopy density (%)	0	0	84	84	10	12	12.7	14.2	14.72	15.1
	Canopy area (%)	0	0	100	100	3	5	18.2	20.3	27.99	28.5
Habitat type	Agriculture (%)	0	0	98	96	10	10	25.74	24.6	31.25	29.9
	Developed (%)	0	0	100	100	8	10	20.16	20.6	25.33	23.9
	Disturbed (%)	0	0	95	95	0	0	3.78	3.69	10.25	9.67
	Forest (%)	0	0	95	95	0	0	7.25	8.05	17.58	17.9
	Grassland (%)	0	0	96	96	15	16	23.52	23	23.6	22.5
	Riparian (%)	0	0	89	89	2	3	6.25	7.08	9.91	10.4
	Shrubland (%)	0	0	95	95	0	0	9.79	9.3	19.15	18.2
	Vegetation diversity	1	1	9	9	5	5	5.04	5.19	1.6	1.57
	Vegetation complexity	2	2	176	176	29	31	32.21	33.8	21.17	21.1

Notes: OR = original data. SM = SMOTE data. SD = standard deviation.

Table 2. K-fold model performance indicators for original and SMOTE data; *K* is equal to 10.

Model type	Original data			SMOTE data		
	RMSE	R^2	MAE	RMSE	R^2	MAE
Linear regression	3.514	0.2008	1.771	4.136	0.2296	2.272
Negative binomial	4.005	0.1615	1.634	4.435	0.1942	2.181
Bayesian generalized linear	3.515	0.2008	1.770	4.143	0.2297	2.271

Notes: RMSE = Root Mean-squared Error. R^2 = Pseudo R-squared. MSE = Mean Absolute Error.

best-performing model used SMOTE data and 13 variables. Large mammal distribution, riparian area and vegetation diversity all have a positive influence on increased risk of WVCs, after controlling for VMT as exposure. A greater share of riparian area that surrounds the road can attract wildlife animals and increase the chance of an interaction between wildlife and vehicles. Riparian habitat consists of the transition zone between wet and dry ecosystems and provides travel paths during low water levels as well as abundant plant diversity for animal nourishment and hiding places.

In a similar vein, the positive effect of vegetation diversity may indicate an increase in WVC exposure because animals frequently hit by vehicles prefer mixed habitats with multiple plant species and habitat types.

By contrast, distance to open water, percentage of shrubland and average soil slope are associated with lower WVC risk. The farther open water is from a road, the less likely a WVC will occur on that section of road. This is an expected result, since there is a decrease in WVC exposure the further the water source is from the roadway. In addition, an increase in the steepness of a soil slope means a decrease in WVC risk. Although some wildlife species may seek out steep slopes or cliffs for their protection from predators (e.g. bighorn sheep), the ungulate species most common in Montana WVCs—deer, elk, pronghorn, moose—tend to avoid steep slopes because of the difficulty in travel.

Just as landscape variables influence wildlife behaviour and movements, road characteristics also impact driver reactions and motorist safety. Higher speed limits on roads increase the risk of

Table 3. Coefficient estimates from the best performing model (NB) for the full-year datasets.

Variable	Coefficient estimate	Standard error	Z-value	P-value	Practical significance
Intercept	− 9.712	0.111	− 87.764	< 0.0001	NA
Urban area	− 0.628	0.056	− 11.276	< 0.0001	− 0.4663
NHS non-interstate class	− 0.102	0.028	− 3.646	0.0003	− 0.0970
Secondary road class	0.518	0.04	13.019	< 0.0001	0.6787
Number of large mammals	0.245	0.007	36.663	< 0.0001	0.2776
Speed	0.021	0.001	17.369	< 0.0001	0.0212
AADT	− 0.0001	0.000003	− 28.901	< 0.0001	− 0.0001
Arterial road length	0.766	0.044	17.432	< 0.0001	1.1511
Rural road length	0.09	0.021	4.364	< 0.0001	0.0942
Distance to water	− 0.114	0.019	− 6.044	< 0.0001	− 0.1077
Average terrain slope	− 0.018	0.002	− 8.675	< 0.0001	− 0.0178
Riparian area	0.013	0.001	12.619	< 0.0001	0.0131
Shrubland area	− 0.007	0.0007	− 9.942	< 0.0001	− 0.0070
Vegetation diversity	0.094	0.009	10.473	< 0.0001	0.0986

Table 4. The top 10 variables of importance for original and SMOTE data in RF models.

Rank of most influential variable	Increase in MSE (%)	
	Original data	SMOTE data
1	Elevation (43.5)	Elevation (53.1)
2	AADT (37.8)	AADT (51.7)
3	Mammals (36.1)	Mammals (37.5)
4	Agriculture (33.12)	Arterial (37.5)
5	Arterial (29.8)	Agriculture (36)
6	Slope (25.8)	Grassland (34.1)
7	Dist. to rest area (25.5)	Slope (31.7)
8	Road width (25.1)	Road network (30.8)
9	Canopy (24.6)	Diversity (30.2)
10	Grassland (24)	Road width (29.8)

WVCs, where a lower speed allows for more time for the animal and vehicle to react to each other. This study's results capture how rising traffic volume as measured by the average annual daily traffic (AADT) reduces the risk of WVCs, after controlling for the effect of the VMT. Increased traffic volumes create a barrier effect that may act as a deterrent to large wildlife species that are a safety hazard for motorists [43].

Moreover, an increase in secondary, arterial and rural road centreline miles is associated with an increase in WVC risk. Many arterial roads in Montana pass through areas that overlap with wildlife habitats. The lower traffic volumes and shorter roadway width of these arterials may encourage wildlife crossings, compared to Non-interstate National Highway System (NHS) roads which carry higher traffic volumes. Interstate and local roads are removed from the models because their effects are insignificant. The weak coefficient estimate of interstate roads may be attributed to their smaller representation, compared to other road classes. Although local roads have the highest average length across analysis units, they are commonly identified within city limits where WVCs are scant compared to rural and arterial roads. Practical significance measures the change in the response variable for a unit change in the covariate. For a log-linear model like the NB, the percentage change in the WVC risk for a one-unit increase in a covariate is computed as: $e^{\beta} - 1$ (as shown by the practical significance in Table 4). For example, a unit increase (1 mile) in arterial centreline distance is associated with a 115% spike of WVC risk. For each unit increase (1 mile) in distance to water from the road, WVC risk is estimated to drop by 11%, holding all other variables constant.

4.4. Random forest models

The RF models produced similar results when compared to the NB models but provided much higher R^2 values and lower RMSE values. The importance of a variable is measured by the percentage of increase in RMSE when that variable is removed, as shown in Fig. 1. RF models run on SMOTE data flagged more variables as important, just as NB models estimated coefficients at greater significance levels when run on SMOTE data. As shown in Table 4, eight variables remained among the top 10 most important variables when the original data were replaced with SMOTE data.

4.5. Seasonal WVC models

The same process for SMOTE, variable selection and modelling used for the previously described full-year data analysis was applied to the seasonally divided data. The seasons were divided by full months, with each season identified as starting the first day of the month that contains an equinox or solstice. The seasonal datasets expressed greater imbalance: carcasses are spotted for only 15.6%, 13.8%, 14.6% and 20.2% of the analysis units during the winter, spring, summer and fall (autumn) seasons. SMOTE is used to achieve a 50% event ratio for each season. Table 5 shows the K-fold cross-validation results on the seasonal models. MAE is more robust against outliers than RMSE [44]. Like the full-year model, the NB model has the lowest MAE when compared to other models using the same data.

Table 6 summarizes the top-performing NB and RF models for the full-year and seasonal data sets. Seasonal NB models run on the SMOTE-augmented data report an average of 33% reduction in the RMSE compared with the same models run on the original data. Moreover, the SMOTE-augmented data is found to work particularly well with the RF regressions for analysing seasonal WVC risk, improving both prediction accuracy (RMSE) and goodness of fit (R^2). Across the four seasonal models, RF regressions with SMOTE-augmented data are found to shave off 75% of the RMSE from what are reported by their NB counterparts (Table 6). The improvement in data balance via SMOTE can increase the predictive power of both the NB and RF models, as shown by the smaller RMSE when the SMOTE-augmented data is used.

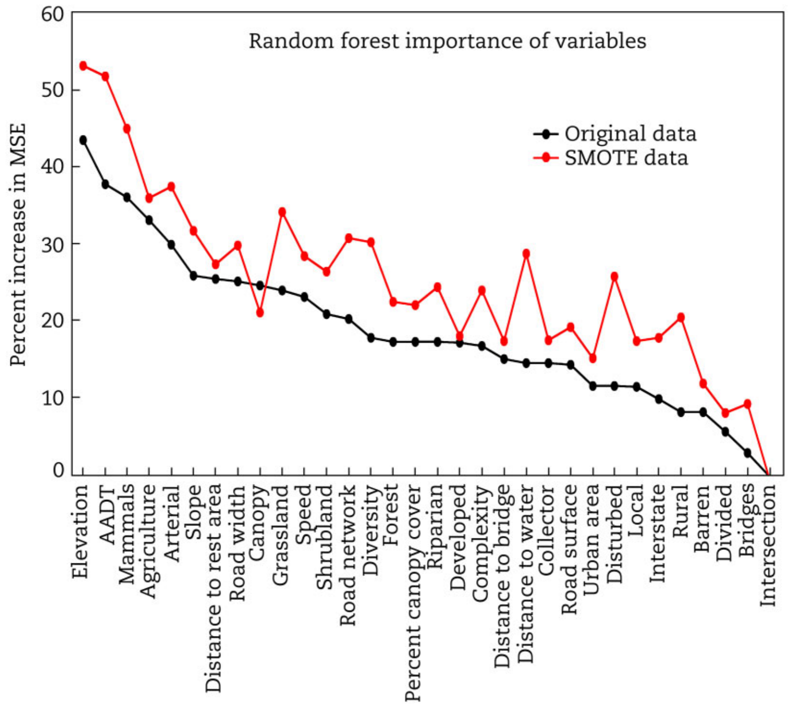

Fig. 1. Comparison of the variables of importance for the original and SMOTE data in RF modelling.

Table 5. K-fold model performance indicators for original and SMOTE data for each season; *K* is equal to 10.

Season	Model type	Original data			SMOTE data		
		RMSE	R^2	MAE	RMSE	R^2	MAE
Winter	Linear regression	1.1947	0.1302	0.5627	1.7647	0.2327	1.0778
	Negative binomial	1.307	0.112	0.5159	1.8158	0.2111	1.0683
	Bayesian generalized linear	1.1947	0.1308	0.5627	1.7634	0.2334	1.0775
Spring	Linear regression	0.9629	0.1448	0.4725	1.4278	0.28	0.9339
	Negative binomial	1.0288	0.1163	0.4202	1.4561	0.2608	0.9171
	Bayesian generalized linear	0.9633	0.1449	0.4723	1.4277	0.28	0.9339
Summer	Linear regression	0.8929	0.148	0.4464	1.3009	0.2598	0.8559
	Negative binomial	0.9448	0.1206	0.4046	1.3349	0.2321	0.8518
	Bayesian generalized linear	0.8919	0.1487	0.4465	1.3007	0.26	0.8559
Fall	Linear regression	1.4692	0.1553	0.7045	2.0124	0.2278	1.1542
	Negative binomial	1.5563	0.1381	0.6411	2.0328	0.2215	1.1149
	Bayesian generalized linear	1.4668	0.1553	0.7044	2.0128	0.2281	1.1543

Notes: RMSE = Root Mean-squared Error. R^2 = Pseudo R-squared. MAE = Mean Absolute Error.

Looking at the results in Table 6, models fitted with SMOTE-augmented data yield smaller RMSE than their counterparts that are run on the original data when they are used to predict seasonal collision risk. This suggests that SMOTE augmentation is able to mimic the data-generating process of the original data and improve out-of-sample prediction. The results also show that RF models outperform NB models for anticipating seasonal risk, as evidenced by the reduction in RMSE throughout the seasonally divided datasets with or without augmentation. However, the RF model fails to outcompete the NB model when predicting year-round collision risk, indicating that seasonal effects may be masked by yearly aggregates and are best treated by running seasonal models separately. Wildlife is known to travel and move through different landscapes depending on the seasons. Trying to predict WVC risk for the full year may overfit the data because the variables used in each seasonal model are not identical.

4.6. Risk maps

The best-performing, full-year model is an NB model with SMOTE data and was used to map state-wide WVC risk (Fig. 2). The predictions represent the average number of collisions expected per 100 million VMT after controlling for road (e.g. speed limit, road functional class), environment (e.g. landcover type, slope, distance to water), species distribution and AADT.

The RF models with SMOTE data yielded the best predictions among the seasonal models, so they were selected to plot seasonal

Table 6. The best performing NB and RF models for full-year and seasonal data.

Data, season No. of variable	Negative binomial			Random forest	
	$RMSE_{NB}$	$R^2_{McFadden}$	$R^2_{Adjusted}$	$RMSE_{RF}$	R^2_{RF}
OR, YR, 9	1.957	0.1392	0.1388	2.917	0.4535
SM, YR, 13	1.685	0.1341	0.1337	2.731	0.6617
OR, WI, 13	2.839	0.1563	0.1552	1.028	0.3614
SM, WI, 14	1.832	0.1378	0.1372	1.032	0.7388
OR, SP, 13	3.189	0.1731	0.1719	0.822	0.3785
SM, SP, 19	1.952	0.1593	0.1586	0.818	0.7683
OR, SU, 18	3.116	0.1774	0.1757	0.801	0.316
SM, SU, 14	1.921	0.1436	0.143	0.814	0.7153
OR, FA, 14	2.554	0.1649	0.1639	1.245	0.4003
SM, FA, 29	2.058	0.1556	0.1545	1.134	0.7511

Notes: OR = original data; SM = SMOTE data; YR = all year; WI = winter; SP = spring; SU = summer; FA = Fall. RMSE = Root Mean-squared Error. R^2 = Pseudo R-squared.

risks across the state. To illustrate the seasonal changes in WVC risks, a stretch of Highway 191 was used to create a zoom-in view in Fig. 3. This highway serves as the only thoroughfare connecting Bozeman to popular recreation sites and the West entrance of Yellowstone National Park. As shown in Fig. 3, this highway section bore significant WVC risks during the fall and winter months. The predicted risks align with state crash reports, correctly marking segments that contain over 80% of the WVCs reported between September and February. Plausibly, WVC risks along that section may rise as snow forces wildlife out of the mountains into the agricultural, low-elevation areas, or as tourists flock to nearby ski resorts and may have trouble driving under inclement weather. Future studies can explore these factors in detail. In sum, the risk maps captured the changes in risk throughout the year and generally matched where WVCs actually happened. These locations coincide with published research on the locations of WVC hotspots during fall migration in the state of Montana [14].

5. Discussion

This paper provides one of the first applications for rural roads to employ SMOTE to confront weak or lacking WVC data. Using a state-wide dataset (No. of road segments = 23,356; No. of variables = 35), we demonstrate that SMOTE retains the shape and structure of the original wildlife carcass data and, when incorporated into standard techniques (e.g. NB and RF), can enhance prediction and estimation for year-round and seasonal WVC risks. We focus on those standard models because they are widely available in R (or Python), have been applied to numerous traffic safety studies and can be readily used by transportation agencies in rural settings.

SMOTE can effectively address data imbalance issues. Dividing the data by season improves prediction accuracy because it allows the model structures to pick up the seasonal effects that may be muted by yearly data. It is important to confront seasonal effect when analysing WVCs due to the seasonal changes in wildlife behaviour (e.g. feeding, breeding or migratory patterns) throughout a year. Inevitably, seasonal division exacerbates data imbalance, which can be effectively addressed by SMOTE.

This study identifies several new directions for future work. First, this study showed the prediction gain by running separate seasonal models with SMOTE data. It is unknown how this approach squares with more advanced models that directly consider the seasonal and regional effects through random parameters and whether the prediction gain, if any, outweighs the increased costs due to the complexity of those models [45, 46]. Although the 1 km segmentation is consistent with prior models, future studies can look at finer spatial scales along homogenous road segments and control for geometric design variables [31, 32]. Secondly, separating WVCs by species has the ability to increase the predictive power of the models, and applying SMOTE on scant species-specific carcass data can effectively identify influential factors and reveal high-risk segments [4]. Lastly, future work can explore varying weights to reflect higher magnitudes of relevance to locations where WVCs have been consistently observed over extended periods of time to increase model inference and prediction accuracy.

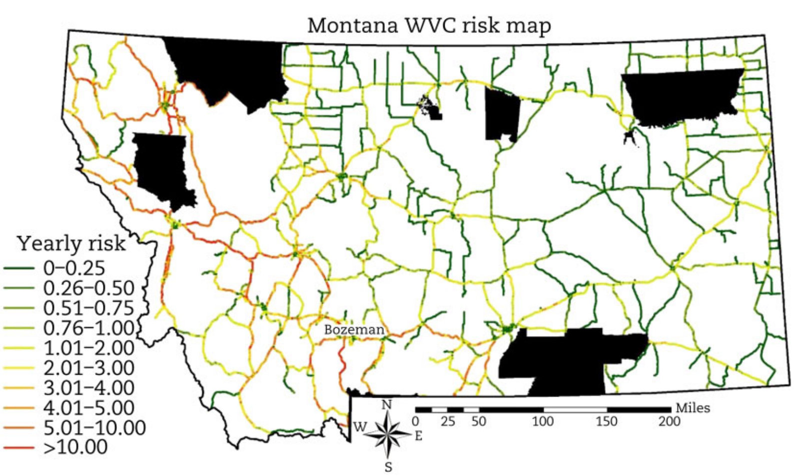

Fig. 2. WVC risk map for the state of Montana using SMOTE data (there are incomplete data for National Parks and Tribal land, so predictions are not made in those areas).

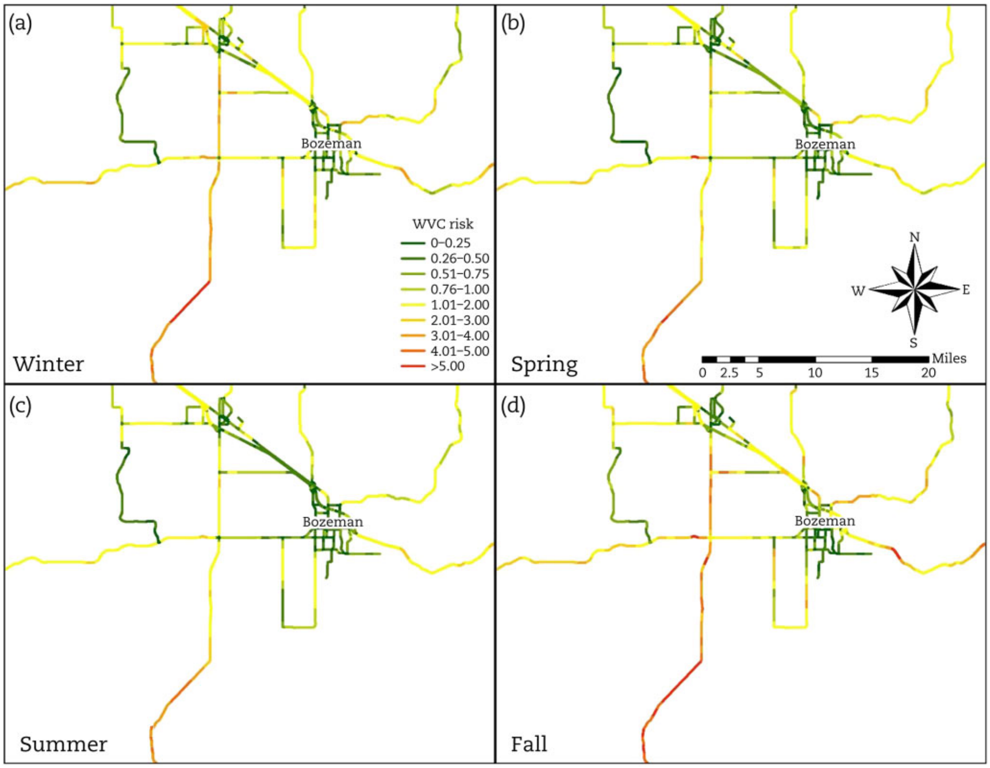

Fig. 3. Zoom-in map of Highway 191, Southwest to Bozeman, MT : (a) winter, (b) spring, (c) summer and (d) fall.

6. Conclusions

This study employs machine-learning techniques to develop high-quality WVC risk maps, by amending sparse carcass data across a vast road network. SMOTE offers a powerful way to augment scant WVC data found along many rural roads, improving the balance between analysis units with no WVCs and those that have WVC data. The balanced data enables statistical models to efficiently identify influential variables associated with WVCs. Combining SMOTE data augmentation with machine learning also allows for more robust detection of at-risk road segments thanks to the synthetically created data that mimic the processes that generate WVC risk. This combined data analysis tool can potentially help transportation agencies and ecologists to make well-informed decisions on where to prioritize mitigation efforts with limited WVC data.

Acknowledgements

We would like to thank Western Transportation Institute and the Montana State University College of Engineering for their support throughout the duration of this research.

Conflict of interest statement

None declared.

References

1. Chen HL, Koprowski JL, Beckmann JP et al. *Safe Passages: Highways, Wildlife, and Habitat Connectivity*. Washington, DC: Island Press, 2010, ISBN:978-1-5972-6654-3.
2. Huijser M, McGowen P, Fuller J et al. *Wildlife-vehicle Collision Reduction Study: Report to congress*. Washington, DC: US Department of Transportation, Federal Highway Administration, 2008, FHWYA-HRT-08-034 254.
3. Sullivan J. *Relationships between lighting and animal-vehicle collisions*. Ann Arbor, MI: University of Michigan, Transportation Research Institute, 2009, UMTRI-2009-35.
4. Ha H, Shilling F. Modelling potential wildlife-vehicle collisions (WVC) locations using environmental factors and human population density: a case-study from three state highways in Central California. *Ecological Informatics* 2017; **43**:212–22.
5. Bíl M, Andrášik R. The effect of wildlife carcass underreporting on KDE+ hotspots identification and importance. *J Environ Manage* 2020; **275**:111254.
6. Huijser MP, Fuller J, Wagner ME et al. *Animal-vehicle Collision Data Collection. A Synthesis of Highway Practice*. Washington DC, USA: Transportation Research Board of the National Academies, 2007, NCHRP Synthesis 370. Project 20-05/Topic 37-12.
7. Bell MA. An investigation modeling risk of wildlife-vehicle collisions in Montana, USA. Master's Thesis. *Montana State University-Bozeman, Norm Asbjornson College of Engineering* 2019.
8. Donaldson BM. Improving animal-vehicle collision data for the strategic application of mitigation. *Virginia Transportation Research Council* 2017, (No. FHWA/VTRC 18-R16).
9. Hall JW, Rutman EW, Brogan JD. *Highway safety challenges on low-volume rural roads*. New Mexico Division: Federal Highway Administration, 2002.
10. Pagany R. Wildlife-vehicle collisions-influencing factors, data collection and research methods. *Biol Conserv* 2020; **251**:108758.
11. Barthelmess EL. Spatial distribution of road-kills and factors influencing road mortality for mammals in Northern New York State. *Biodivers Conserv* 2014; **23**:2491–514.

12. Clevenger AP, Chruszcz B, Gunson KE. Spatial patterns and factors influencing small vertebrate fauna road-kill aggregations. *Biol Conserv* 2003; **109**:15–26.
13. Clevenger AP, Hardy A, Gunson K et al. Analyses of wildlife-vehicle collision data: applications for guiding decision-making for wildlife crossing mitigation and motorist safety. *National Cooperative Highway Research Program*. Utah State University, 2006.
14. Creech T, McClure M, Callahan R. High-risk zones for ungulate-vehicle collisions during Montana's fall migration season. *The Center for Large Landscape Connectivity* 2016.
15. Dank Z, Porter W. Temporal, spatial, and landscape habitat characteristics of moose-vehicle collisions in western Maine. *The Journal of Wildlife Management* 2010; **74**:1229–41.
16. Jensen R, Gonser R, Joyner C. Landscape factors that contribute to animal-vehicle collisions in two northern Utah canyons. *Appl Geogr* 2014; **50**:74–79.
17. Hothorn T, Brandl R, Müller J. Large-scale model-based assessment of deer-vehicle collision risk. *PLoS One* 2012; **7**:e29510.
18. Nelli L, Langbein J, Watson P et al. Mapping risk: quantifying and predicting the risk of deer-vehicle collisions on major roads in England. *Mamm Biol* 2018; **91**:71–78.
19. Stevens B, Dennis B. Wildlife mortality from infrastructure collisions: statistical modeling of count data from carcass surveys. *Ecology* 2013; **94**:2087–96.
20. Visintin C, Ree R, McCarthy M. Consistent patterns of vehicle collision risk for six mammal species. *J Environ Manage* 2017; **201**:397–406.
21. Torgo L, Ribeiro RP, Pfahringer B et al. Smote for regression. *Portuguese Conference on Artificial Intelligence*. Berlin, Heidelberg: Springer, 2013, 378–89.
22. Chen T, Shi X, Wong YD. Key feature selection and risk prediction for lane-changing behaviors based on vehicles' trajectory data. *Accident Analysis & Prevention* 2019; **129**:156–69.
23. Shin J, Yoon S, Kim Y et al. Effects of class imbalance on resampling and ensemble learning for improved prediction of cyanobacteria blooms. *Ecological Informatics* 2020; **61**:101202.
24. Schlögl M, Stütz R, Laaha G et al. A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high-resolution dataset. *Accident Analysis & Prevention* 2019; **127**:134–49.
25. Gonzalez-Velez JC, Torres-Madronero MC, Murillo-Escobar J et al. An artificial intelligent framework for prediction of wildlife vehicle collision hotspots based on geographic information systems and multispectral imagery. *Ecological Informatics* 2021; **63**: 101291.
26. Nguyen HK, Fielding MW, Buettel JC et al. Predicting spatial and seasonal patterns of wildlife–vehicle collisions in high-risk areas. *Wildlife Research* 2022; **49**:428–37.
27. Pagany R, Valdes J, Dorner W. Risk prediction of wildlife-vehicle collisions comparing machine learning methods and data use. In: *10th International Conference on Advanced Computer Information Technologies (ACIT)* IEEE, 2020, 436–40.
28. RStudio Team. *RStudio: Integrated Development for R. RStudio*. Boston, MA: PBC, 2020, http://www.rstudio.com/.
29. ESRI. *ArcGIS desktop: release 10.6 redlands*. Califonia: Environmental Systems Research Institute, 2018.
30. Le T, Gross F, Harmon T et al. *Safety Evaluation of Corner Clearance at Signalized Intersections*. United States: Federal Highway Administration, 2018, (No. FHWA-HRT-17-084).
31. Cook DJ, Souleyrette RR, Jackson J. Effect of road segmentation on highway safety analysis. *Transportation Research Board* 2011; **11-1995**.
32. Zhao X, Lord D, Peng Y. Examining network segmentation for traffic safety analysis with data-driven spectral analysis. *IEEE Access* 2019; **7**:120744–57.
33. Hegland SJ, Hamre LN. Scale-dependent effects of landscape composition and configuration on deer-vehicle collisions and their relevance to mitigation and planning options. *Landsc Urban Plan* 2018; **169**:178–84.
34. Hurley M, Rapaport E, Johnson C. Utility of expert-based knowledge for predicting wildlife-vehicle collisions. *The Journal of Wildlife Management* 2009; **73**:278–86.
35. Montana FWP. *Montana Fish, Wildlife, and Parks Wildlife Maps and GIS Resources*. Helena, Montana: Montana FWP, 2018, https://gis-mtfwp.opendata.arcgis.com/search?tags=wildlife.
36. U.S. Geological Survey. National Land Cover Database. Washington, DC: United States Geological Survey, 2016, https://www.usgs.gov/centers/eros/science/national-land-cover-database.
37. Abou Elassad ZE, Mousannif H, Al Moatassime H. A proactive decision support system for predicting traffic crash events: a critical analysis of imbalanced class distribution. *Knowledge-Based Systems* 2020; **205**:106314.
38. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**:1–2.
39. Iranitalab A, Khattak A. Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention* 2017; **108**:27–36.
40. Scornet E. Random forests and kernel methods. *IEEE Trans Inf Theory* 2016; **62**:1485–500.
41. Mondal AR, Bhuiyan MAE, Yang F. Advancement of weather-related crash prediction model using nonparametric machine learning algorithms. *SN Applied Sciences* 2020; **2**:1–11.
42. James G, Witten D, Hastie T et al. *An Introduction to Statistical Learning: with Applications in R*. New York City, New York: Springer, 2017, ISBN-13: 978-1461471370.
43. Seiler A. Trends and spatial patterns in ungulate-vehicle collisions in Sweden. *Wildlife Biology* 2004; **10**:301–13.
44. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geosci Model Dev* 2014; **7**:1247–50.
45. Geedipally SR, Lord D, Dhavala SS. The negative binomial-Lindley generalized linear model: characteristics and application using crash data. *Accident Analysis & Prevention* 2012; **45**:258–65.
46. Shaon MRR, Qin X, Shirazi M et al. Developing a random parameters negative binomial-Lindley model to analyze highly over-dispersed crash count data. *Analytic methods in accident research* 2018; **18**:33–44.

Received: March 2, 2023. **Revised:** May 31, 2023. **Accepted:** September 10, 2023