



# Analysing and predicting wildlife–vehicle collision hotspots for the Swiss road network

Patrick Laube · Nils Ratnaweera · Anna Wróbel · Ivo Kaelin · Annette Stephani · Martina Reifler-Baechtiger · Roland F. Graf · Stefan Suter

Received: 30 September 2022 / Accepted: 7 April 2023 / Published online: 24 April 2023  
© The Author(s) 2023

## Abstract

**Context** Wildlife–vehicle collisions (WVCs) are a significant threat for many species, cause financial loss and pose a serious risk to motorist safety.

**Objectives** We used spatial data science on regional collision data from Switzerland with the objectives of identifying the key environmental collision risk factors and modelling WVC risk on a nationwide scale.

**Methods** We used 43,000 collision records with roe deer, red deer, wild boar, and chamois from 2010 to 2015 for both midlands and mountainous landscape types. We compared a fixed-length road segmentation approach with segments based on Kernel Density Estimation, a data-driven segmentation method. The segments' environmental properties were derived from land-cover geodata using novel neighbourhood operations. Multivariate logistic regression and random forest classifiers were used to identify and rank the relevant environmental factors and to predict collision risk in areas without collision data.

**Results** The key factors for WVC hotspots are road sinuosity, and two composite factors for browsing/forage availability and traffic noise—a proxy for traffic flow. Our best models achieved sensitivities of 82.5% to 88.6%, with misclassifications of 20.14% and 27.03%, respectively. Our predictions were better in forested areas and revealed limitations in open landscape due to lack of up-to-date data on annual crop changes.

**Conclusions** We illustrate the added value of using fine-grained land-cover data for WVC modelling, and show how such detailed information can be annotated to road segments using spatial neighbourhood functions. Finally, we recommend the inclusion of annual crop data for improving WVC modelling.

**Keywords** Wildlife–vehicle collision · Kernel Density Estimation · Neighbourhood functions · Spatial data science · Random Forests

---

P. Laube (✉) · N. Ratnaweera · A. Stephani · M. Reifler-Baechtiger · R. F. Graf · S. Suter  
Institute of Natural Resource Sciences, ZHAW Zurich  
University of Applied Sciences, Grüental, Postfach,  
8820 Wädenswil, Switzerland  
e-mail: patrick.laube@zhaw.ch

A. Wróbel · I. Kaelin  
Institute of Computational Life Sciences, ZHAW Zurich  
University of Applied Sciences, Grüental, Postfach,  
8820 Wädenswil, Switzerland

## Introduction

Wildlife–vehicle collisions (WVCs) cause financial loss in the millions, can lead to human harm and in severe cases human fatalities. Switzerland alone counts up to 20,000 accidents with medium size and large wildlife every year, resulting in more than 25 million Swiss Francs material damage (Roth 2012). Depending on the various landscape types, the most affected species are roe deer (*Capreolus capreolus*),

red deer (*Cervus elaphus*), wild boar (*Sus scrofa*), and chamois (*Rupicapra rupicapra*). With thriving populations most notably with wild boar and red deer and increasing traffic flow, these figures are expected to grow.

The causes for WVC are manifold and their interplay is complex. Wildlife accidents often occur in situations where roads cross favorable wildlife habitats (Malo et al. 2004; Garrah et al. 2015). Collision predictors include species-specific factors (“the animal”), traffic related factors (“the driver”), and factors related to the road and its embedding in the immediate and broader environment (“the road”). Gunson et al. (2011) categorize a plethora of landscape-related predictors (e.g., proximity to forest, proportion of open area or built-up area along road, Shannon diversity index) and road-related predictors (slope, in-line visibility, speed limit, traffic volume). WVC show furthermore significant spatio-temporal patterns related to diurnal light changes and seasonal changes of the road conditions (Neumann et al. 2012; Ascensão et al. 2019).

#### Research gap and contribution

This article presents results of a multi-year nationwide study in Switzerland aiming at a better understanding of the complex processes leading to WVC aiming at the implementation of better collision prevention measures. Extensive small-scale field studies at collision hotspots were combined with large-scale geospatial analysis and modelling. Six years of collision records were collected from emergency services and hunting associations for three Swiss cantons representing two typical landscape types (Zurich and Fribourg for midlands, Grisons for mountainous areas). These collision records were then associated to segments of the road network and then analysed with respect to their embedding into the natural and built environment. After identifying the most relevant predictors, collision risk was modeled for the road network on a nation-wide scale for all other Cantons. The long-term goal of the overarching project was to identify hotspots, where then locally adapted prevention measures could be realised to finely reduce WVC and thereby reduce socio-economic costs. This article summarizes the results of the large-scale geospatial analysis and modeling part of the project. The specific research questions for this article were:

- What are the key predictors for WVC on Swiss roads and what are differences between the midlands and mountainous areas?
- How can neighbourhood functions operationalize the environmental variables embedding collision hotspots into the natural and built environment?
- How well do these variables predict collision risk using regression models and machine learning?

Our research is in line with a number of closely related studies. Bíl et al. (2016, 2019) studied WVC on Czech roads using a specifically developed network-based Kernel Density Estimation (KDE+) approach on a very similar data set. In their report for the Conference of European Directors of Roads (CEDR), Seiler et al. (2016) used a modified version of KDE+ for a similar study on WVC in Catalonia, Spain and south-central Sweden. Our study aims at comparing our results to these other studies investigating similarities and differences with respect to (i) environmental differences between the study areas, and (ii) spatial and semantic granularities of the geodata and methodologies used.

We argue that our study adds further evidence that the overall workflow put forward by the just mentioned studies (segmentation, annotation, prediction) works well, exemplified in two more landscape and ecosystem types (Swiss midlands and alps). We furthermore advance WVC analysis through the consequent use of spatially fine-grained and semantically detailed geodata mainly from the Swiss National Mapping Agency, and thereby systematically reducing the share of qualitative and semi-manually derived data found in precursory studies. Consequently, the contributions of this article are as follows:

- (1) Development of WVC models for fine-grained and semantically detailed areal landscape geodata.
- (2) Development of geospatial neighbourhood functions for WVC studies for annotating road segments with landscape geodata from topographic landscape models from national mapping agencies (both vector and raster data).
- (3) Comparison between fixed 200 m segmentation and data-driven KDE segmentation. This comparison also allows contrasting workflows with and without the need of coldspot controls.

- (4) Comparison between midlands and mountain models for the same country, identification of (i) variations in the factor ranking, and (ii) differences in model performance depending on landscape types.

### Related work

There is a large body of literature covering various aspects of WVCs. The review article by Gunson et al. (2011) offers a well-structured entry point into the subject. We first give an overview of the most recently used methodologies to analyse WVC and then summarize the most important ecological results relevant for our study.

The most recent WVC studies agree on the general methodology of first *segmenting* the road network into collision hotspots (and where necessary coldspots for comparison), then *annotating* the segments with environmental variables characterizing their embedding into the natural and built environment, and finally using statistical modelling or machine learning for *predicting* collision risk.

**Road network segmentation.** Most WVC studies aim at finding road sections with significantly more collisions than on comparable road sections. Such road segmentation can be done using fixed-length intervals (e.g., 100 m, 200 m, 500 m) or using some form of data-driven clustering. Garrah et al. (2015) use predefined fixed-length segments and then count the number of collision incidents per segments in order to identify hotspots (or coldspots). Bíl et al. (2016, 2019) recommend the clustering method KDE+ for the initial segmentation, a form of Kernel Density Estimation tailored to the identification and ranking of hotspots along networks. Map matching might be necessary to relocate the point locations of the collision reports onto the closest road in the first place (Kubicka et al. 2018).

Many studies on WVC use coldspots, that is road sections where no or only few accidents were recorded, as controls for statistical analysis (Garrah et al. 2015). Several approaches for identifying and constructing coldspots have been proposed. Seiler et al. (2016) selected for every hotspot cluster a closeby coldspot with a predefined upper threshold of collisions (e.g.  $\geq 3$  for hotspots,  $< 3$  for coldspots) and a minimal distance of 1 km. Garrah et al. (2015)

used the Getis–OrdGi\* statistics that compares collision records for given segments with their neighboring segments and an expected overall distribution. Bíl et al. (2019) argue, however, that the use of coldspots as controls might distort the results, as coldspots could be identified and used where animals could physically not enter. In our study we use coldspots but use spatial masks excluding areas not visited by our target animals.

**Environmental annotation of road segments.** Once the collision clusters or hotspot segments are identified, they must be semantically annotated. This typically includes landscape-related and road-related factors that both potentially cause accidents related to the animals' or the motorists' behaviours (Gunson et al. 2011). The semantic annotation requires spatial data science or Geographic Information Systems (GIS) operations and landscape-related geodata. Landscape-related factors mostly involve *distance metrics* to landscape elements (e.g., distance to forest, water, built-up areas) and *neighbourhood functions* (e.g., fraction of open land or a landscape diversity index within a predefined buffer) (Gunson et al. 2011; Seiler et al. 2016; Bíl et al. 2019). Road-related factors can be feature attributes of the roads (e.g., speed limit, traffic intensity), shape descriptors of the road features (e.g., slope, line sinuosity, visibility parameters), or again neighbourhood factors characterizing the infrastructure setting of the road (e.g., guard rails, embankments, shrubs, grass belts) (Seiler et al. 2016; Bíl et al. 2019). The latter factors are often given as categorical or Boolean variables (e.g. presence of guard rails yes/no).

Some studies include also more *complex factors* that have to be modelled in the first place, such as visibility along the roads or linear landscape structures leading the animals towards the roads (Seiler et al. 2016). Visibility can serve here as an example for many other parameters that can be operationalized in many different ways. Laliberté and St-Laurent (2020) use the simple line property sinuosity, whilst others use complex viewshed analysis based on digital surface models and LiDAR data (Castro et al. 2017; Jung et al. 2018). The same holds for leading structures that some studies capture manually from Google Street View™ (Seiler et al. 2016) whilst others model likely passages using complex multi-criteria habitat modeling with least cost paths functions (Gülci and Emin 2015).

The availability and spatio-temporal quality of geodata for annotation largely dictates any WVC analysis, as not always geodata is available for ecologically meaningful factors. Most studies use a mix of available geodata from national mapping agencies, statistical offices, open data providers and aerial imagery. For example, Bíl et al. (2019) used both geodata (road network) and image interpretation of Google Street View™ and orthophoto maps resulting in 50% ratio data and 50% categorical data factors. Seiler et al. (2016) used a similar mix of ratio vs. categorical factors (approx. 50% each) again using interpretation of aerial photographs, satellite imagery, and Google Street View™ combined with geodata from official databases, digital topographic maps and in few cases field data.

Finally, some authors furthermore make a distinction between *local* factors describing the spatial neighbourhood of a road segment or hotspot, and *global* factors influencing WVCs in the entire study area (Neumann et al. 2012; Bíl et al. 2019). Daytime is a typical global factor.

*Statistical modelling and prediction.* Multivariate prediction models acknowledge the interplay of multiple predictors (Malo et al. 2004). Seiler et al. (2016) used univariate tests for exploration of the individual predictor variables followed by correlation models. They developed logistic models for three model approaches: mixed, road and landscape. Furthermore they evaluated expert models based on road section attractiveness and accessibility to wildlife determined by personal, subjective impressions. Similarly, logistic regression models were developed by Bíl et al. (2019) and interpreted in terms of odds ratios. Included parameters were selected in correlation and principal component analysis. Important variables were chosen in bidirectional step-wise procedure based on Akaike information criterion (AIC).

A series of studies closely related to our study have identified key WVC predictors for a range of different landscape types. Collisions are to be expected where roads cross habitats (Malo et al. 2004). Road sections with high traffic loads and poor visibility are particularly prone to collisions (van Langevelde and Jaarsma 2005; Barrientos and Bolonio 2009). Vehicle speed is a further key factor in wildlife accidents (Elvik 2008; Huijser et al. 2015; Seiler et al. 2016). As speed increases, the braking distance and the impact energy increase in proportion to the square

of the speed. Clearly, interspecies ecological differences further complicate analysis and prediction, as the influence of, for example, the road-side vegetation composition or the use of deicing-salt depends heavily on the covered species (Grosman et al. 2009; Gunson et al. 2011). Garrah et al. (2015) also conclude that mortality on roads is strongly seasonal and that this seasonality is strongly species-specific. Laliberté and St-Laurent (2020) confirm WVC seasonality for WVC with moose and deer in southeastern Canada, along with diurnal effects.

Seiler et al. (2016) studied both landscape factors and road factors for their studies in Catalonia, Spain and south-central Sweden. They found that WVC are more likely on busier roads with higher vehicle speeds and an absence of fences, safety rails, or large embankments. Their study, however, also revealed clear differences in predictor relevance between the two landscape types. In the Catalonian model with a dominance of wild boar collisions, the amount of built-up urban area close by and the proximity to water were the strongest landscape predictors. In Sweden, by contrast, the importance of landscape diversity and linear leading structures directing animals towards the roads reflected the different ecology of roe deer and moose being the main collision victims. Bíl et al. (2019) identified the presence of closed habitats and shrubs along roads and the distance to forest as key predictors. Also Garrah et al. (2015) confirmed the correlation of collision hotspots with the presence of suitable habitats, in their example wetland habitat explaining found amphibian and reptile road mortality hot spots. For Canadian moose, Laliberté and St-Laurent (2020) found interaction between collision risk and slope and elevation, for deer interaction with road sinuosity and the fraction of mature coniferous stands.

*Research gaps.* Few WVC studies could have been based solely on computed quantitative (ratio data) derived from spatially and semantically fine-grained geodata through specifically developed neighbourhood functions. The excellent data supply from several official Swiss data providers allows us to do just that. Furthermore, in our opinion further evidence is needed for selecting one segmentation approach over the other. For that reason we conducted a comparative study opposing different segmentation approaches with and without coldspots. No large scale WVC study has been done for the diverse Swiss landscape

**Table 1** List of geodata sources used characterizing the natural and built environment embedding collision hotspots

Landscape variable	Type	Data set	Layer	Year	Source
Road network	Line	swissTLM3D	Roads and paths	2018	Swisstopo
Waterways	Line	swissTLM3D	Waterways	2018	Swisstopo
Waterways	Line	swissTLM3D	Landuse type	2018	Swisstopo
Forest	Polygon	swissTLM3D	Landuse type	2018	Swisstopo
Hedges	Line	swissTLM3D	Tree and hedge rows	2018	Swisstopo
Primary areas	Polygon	Vector25	Built-up area	2008	Swisstopo
Built-up areas	Polygon	Vector25	Primary areas	2008	Swisstopo
Canton boundaries	Polygon	Boundries3D	Canton boundaries	2018	Swisstopo
DTM	Raster	Digital elevation models	swissALTI3D in 2 m resolution	2018	Swisstopo
DSM	Raster	Digital elevation models	Digital surface model, 2 m resolution	2018	Swisstopo
Biogeographic regions	Polygon	Biodiversity	Biogeographic regions	2004	FOEN
Wildlife corridors	Polygon	Biodiversity	Wildlife corridors transregional	2017	FOEN
Noise	Raster	Noise	Noise and traffic noise at night	2015	FOEN
Vegetation height	Raster	National forest inventory	Vegetation height index	2016	WSL
Traffic volume	Line	National traffic model	Mean daily traffic load and speed	2010	ARE

All data sets are of fine spatial and semantic granularity, originate official data providers, and are directly machine-readable without need for additional manual data capture

types. Our goal is to compare our results with the listed related studies. We furthermore compare two contrasting landscape types, the midlands vs. the alps, in order to achieve regional factors and factor rankings.

## Methods

### Data and preprocessing

Our study is based on a dataset counting in total 43,000 collision records from the Swiss Cantons of Zurich and Fribourg (midlands) as well as Grisons (alps), mainly collected by emergency services and hunting associations. Although the precise data capture procedures vary in between institutions and regions, the minimal attribute set per record includes coordinates, date, and species. The study is limited to larger species, that is roe deer (*C. capreolus*), red deer (*C. elaphus*), wild boar (*S. scrofa*), and chamois (*R. rupicapra*), the most responsible species for WVC in Switzerland. We used data from 2010 to 2015 offering the longest data overlap for all three Cantons. After restricting the collision data to our target

species and the before mentioned time span, a total of 12,431 records were used for our final analysis.

The key geodata set for the environmental annotation of the road segments was swissTLM3D, the large-scale topographic landscape model issued by the Swiss National Mapping Agency Swisstopo.<sup>1</sup> SwissTLM3D is the most extensive and accurate 3D vector dataset for Switzerland. This source provided crucial geodata for the road network, hydrology features, forests and further vegetation layers (Table 1). Swisstopo also provided accurate digital terrain and surface models (DTM, DSM, 2 m resolution). Geodata by the National Mapping Agency is complemented by additional environmental data from the Federal Office for the Environment FOEN<sup>2</sup> (biodiversity, traffic noise), the Swiss Federal Institute for Forest, Snow and Landscape Research WSL<sup>3</sup> (vegetation height index based on Lidar), and the Federal Office for Spatial Development ARE<sup>4</sup> (traffic volumes).

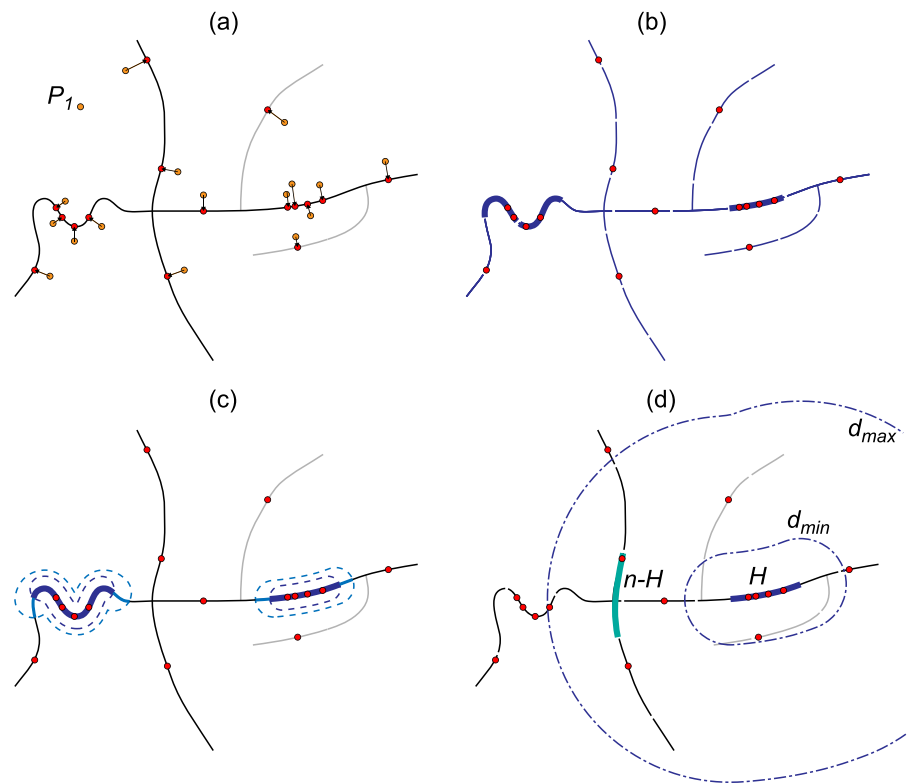
<sup>1</sup> <https://shop.swisstopo.admin.ch/en/products/landscape/tlm3D>.

<sup>2</sup> <https://www.bafu.admin.ch/bafu/en/home.html>.

<sup>3</sup> <https://www.wsl.ch/en/services-and-products>.

<sup>4</sup> <https://www.are.admin.ch/are/en/home.html>.

**Fig. 1** Map matching and segmentation. **a** Map matching to the closest road,  $P_1$  lies outside maximal matching threshold, **b** fixed-length segmentation, **c** KDE-based segmentation with two kernel sizes  $kde_n$  and  $kde_w$ , and **d** selection of coldspot control segments  $n-H$ , within a donut-shaped neighbourhood between  $d_{min}$  and  $d_{max}$  around  $H$



The road dataset was filtered excluding traffic infrastructures where wildlife collisions were impossible (driving bans, tunnels, bridges, ferries) or very rare (fenced motorways, narrow gravel roads with low traffic loads). The collision data revealed that accidents with our target species (see above) hardly ever occur within the built-up area, hence roads within the built-up area were also excluded. This filtering resulted in a road network of 18,100 km. The collision records were finally allocated to the closest road with a map matching procedure having a maximal matching threshold of 100 m (Fig. 1a).

### Segmenting the road network

This study used two approaches for segmenting the roads, a fixed-length segmentation and a data-driven segmentation based on a Kernel Density Estimation of the collision records.

**Fixed-length segmentation.** For the deterministic fixed-length segmentation *reg200*, the road network was cut into segments of a predefined length (Fig. 1b). The selection of 200 m as segment length is based on the literature (Elvik 2008; Garrah et al.

2015) and on the median of the KDE-based segmentation discussed below. Collision hotspots could thereafter be defined with a minimal threshold  $n$  of incidents per segment.

**KDE-based segmentations.** The second segmentation was based on the overlay of the road network with collision record density isolines computed using Kernel Density Estimation KDE (Bíl et al. 2013). This procedure requires two parameters (Fig. 1c): first radius  $r$  for the density kernel, and second,  $p$  the density percentile threshold defining the cluster boundary outline. The literature recommends  $r$ -values between 50 and 500 m (Bíl et al. 2013). For this study we defined and used two KDE segmentations, one based on a narrow kernel and one based on a wide kernel:

- $kde_n$ . Narrow kernel with  $r = 100$  m and  $p = 95\%$  percentile.
- $kde_w$ . Wide kernel with  $r = 200$  m and  $p = 90\%$  percentile.

The validation of the models in Sect. **Statistical analysis (predictions)** require control segments, that is coldspots of similar road segments with little to



no collisions. With the fixed-length segmentation all segments not identified as hotspots can serve as controls and no further coldspot selection is required. For the KDE-based segmentation, however, these control segments must be defined. Our procedure to construct coldspots is closely related to Seiler et al. (2016) (Fig. 1d). coldspot segments must have a similar length as the hotspot segment ( $\pm 10$  m), have the same road category, and must be located within a donut-shaped neighbourhood delineated by a minimal and a maximal distance ( $d_{\min}$ ,  $d_{\max}$ ).

The preliminary of requiring the same road category shall prevent the selection of control segments with entirely different characteristics (e.g. side roads) that could bias the results (Bíl et al. 2019). This in turn excludes road category from the list of possible predictors for the KDE-based segmentation models. By contrast, in the models based on the fixed-length segmentation, road category can be a predictor.

### Spatial modelling of local factors

Compared to related work, our study is entirely based on computed quantitative environmental factors derived from fine-grained official geodata (Table 2). This required the development and tailoring of a set of geospatial operations for the environmental annotation of the road segments. The study used five categories of factors with increasing computational complexity. All but the first category required tailored geospatial operationalizations for the factors based on Geographic Information Systems and Science (GIS) functions and spatial data science routines.

- *Attribute factors*. Primary feature attributes or combinations thereof (e.g. road speed limit),
- *Form factors*. Derived from geometry or shape (e.g. sinuosity of a road segment),
- *Distance factors*. Nearest neighbour distances to target features (e.g. distance to forest),
- *Areal neighbourhood factors*. Areal characteristics within a defined neighbourhood around the road segment (e.g. forest share within 200 m buffer),
- *Complex, modeled factors*. Derived based on a specifically developed model including one to many geodata layers (e.g. leading structures).

The used *attribute factors* include the road\_category from swissTLM3D (characterized in width

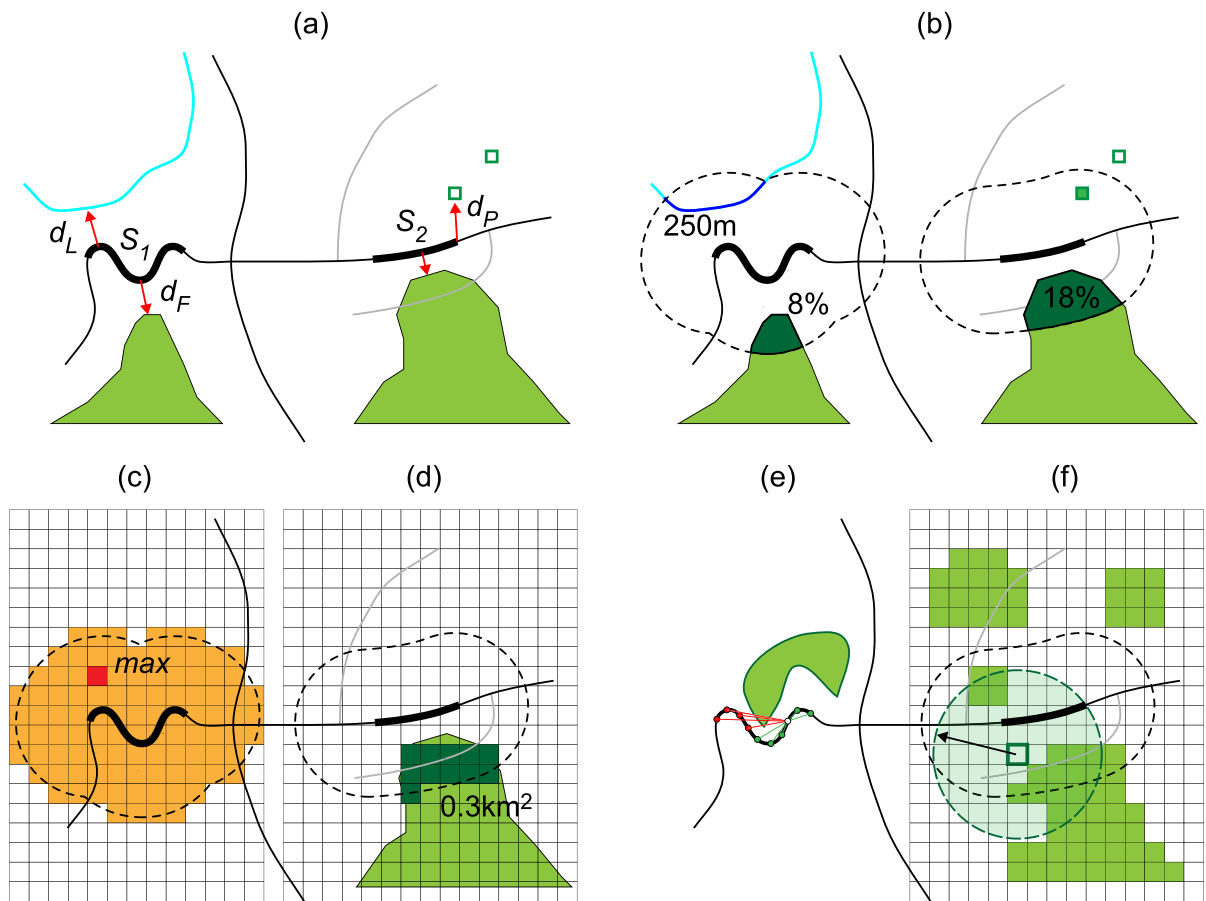
categories), a modelled daily traffic volume in vehicles per 24 h, and a modeled average speed per segment. The *form factor* sinuosity was used as a proxy for visibility. It was computed as the ratio between the Euclidean distance between start and endpoint of a segment and the actual segment length (Fig. 2a).

All *distance factors* assess per segment the nearest neighbour distance to point, line, or polygon features of the targeted landscape element (Fig. 2a). Such simple distance factors used in the study were `dist.forest`, `dist.builtup`, and `dist.corridor`, the last one giving the distance to wildlife corridors. `dist.water` was a compound distance factor assessing the shortest distance to several hydrology layers, even including point (well), line (creek, river), and polygon features (ponds, lakes).

*Areal neighbourhood factors* characterize the spatial composition of landscape elements within a buffer around the road segment, for both vector and raster data (Fig. 2b–d). This includes the areal share of a landscape element (% forest or % primary areas) and particularly for raster data indicating spatial variation within a buffer also zonal statistics (min, max, mean, standard deviation within buffer). Since every choice of a buffer width is somewhat arbitrary, we computed for some buffers various buffer widths and subsequently used them in the predictor ranking (see Sect. [Statistical analysis \(predictions\)](#)). In our study we finally used `noise_Xm` (X referring to the variable buffer widths), `primary_areas_200`, and `vegetation_height_Xm`.

Exploiting the fine-grained detail of our base data we furthermore developed three *complex factors*. In accordance with Seiler et al. (2016) we used leading structures, that is linear landscape elements leading the wildlife towards the roads and hence potentially predicting collisions. We propose, in contrast to aerial image interpretation, two analytical approaches calculating leading structures from geodata.

`leadstruct.DTM` derives leading structures from the digital terrain model using run-off hydrology tools. To this end, leading structures are modelled as ridges and trenches. We then propose the use of two buffer zones around a road segments,  $b_{inner}$  and  $b_{outer}$  (Fig. 3a). Leading structures are then identified and counted that “lead” from  $b_{outer}$  to  $b_{inner}$ . The same methodology is proposed for `leadstruct.TLM`, however, now based on the



**Fig. 2** Operationalizing factors. **a** Form and distance factors, Segment  $S_1$  is more sinuous than  $S_2$ ,  $d_p$  indicates the nearest neighbour distance to point features,  $d_L$  to line features, and  $d_P$  to polygon features. **b** Areal neighbourhood factors with buff-

ers for vector data and **c**, **d** raster data. **e** Line of sight visibility. **f** Complex factor feeding\_Xm modeled from white cells with low vegetation surrounded by a fraction of green high vegetation

swissTLM3D topographic landscape model vector data set considering linear features such as forest edges, water bodies, and hedges (Fig. 3b).

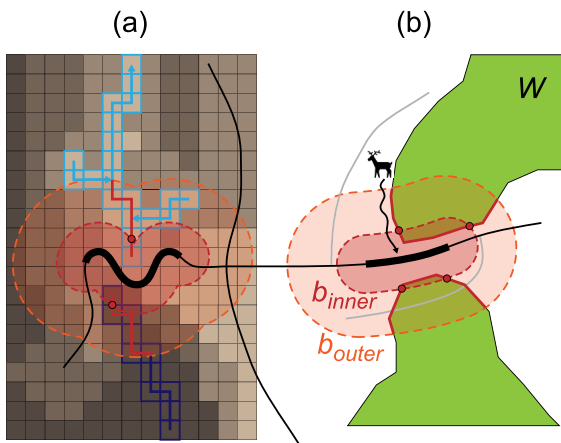
feeding\_Xm models the availability of feeding grounds within a neighbourhood buffer based on a fine-grained Lidar-based vegetation height index. Feeding grounds are modelled as low vegetation (potential feeding area) that is surrounded by high vegetation giving cover (Fig. 2f). Complementary to the simple sinuosity proxy for visibility, we furthermore modeled a complex visibility

using a line-of-sight approach and the digital surface model (DSF, see Fig. 2e).

#### Statistical analysis (predictions)

The goal of the statistical modelling was the identification of significant parameters for the prediction of hotspots and coldspots. The analysis was based on the data of wildlife accidents from three Swiss cantons: Zurich, Fribourg and Grisons, for which the data was available. The two types of segmentation used resulted in a different distribution of segments.





**Fig. 3** Operationalizing leading structures. **a** Ridges and valleys derived from a DTM, and **b** linear leading structures leading from an outer zone into an inner zone

In case of both kernel models ( $kde_n$  and  $kde_w$ ) the distribution was balanced with an equal number of hotspots and corresponding coldspots. The fixed-length segmentation resulted in a highly unbalanced distribution of segments, where hotspots accounted for only about 5% of total segments. Therefore, the statistical analysis below was done on the segments generated with kernel algorithms.

The first step of the statistical analysis was performed on the three cantons separately. We started with univariate tests to check for significant differences (significance level  $\alpha = 5\%$ ) between hotspots and coldspots. Since the investigated parameters were continuous and not normally distributed, non-parametric paired Wilcoxon tests were used. Furthermore, we tested correlations of the parameters with Spearman coefficients to reduce multicollinearity among the predictors. To choose the optimal radius for the

**Table 2** List of environmental factors used in our models, their operationalisation, and sources

Type	Factor	Description	Data source
Attribute	road_category	Road category 4 m, 6 m, 8 m, 10 m, fast road	swissTLM3D
	traffic	Mean daily traffic volume in (number of vehicles per pro 24 h)	National traffic model ARE
	speed	Modeled speed	National traffic model ARE
Form	sinuosity	Sinuosity as a proxy for visibility	swissTLM3D
Nearest neighbour	dist.water	Shortest distance to standing or flowing waters (m)	swissTLM3D
	dist.forest	Shortest distance to forest (m)	swissTLM3D
	dist.builtup	Shortest distance to built-up area (m)	swissTLM3D
	dist.corridor	Shortest distance to wildlife corridor (m)	Wildlife corridors FOEN
Areal neighbourhood	noise_Xm	Traffic noise at night (min, max, mean, std) for buffers 20 m, <b>50 m</b> , 100 m, 200 m, 300 m, 500 m	Traffic noise BAFU
	primary_areas_200m	Primary areas (%) within buffer of 200 m	Vector25
	vegetation_height_Xm	Vegetation height (min, max, mean, std) for buffers 20 m, 50 m, <b>100 m</b> , 200 m, 300 m, 500 m	Vegetation height WSL
Complex	leadstruct.TLM	Number of leading structures (forest edges, water bodies, hedges)	swissTLM3D
	leadstruct.DTM	Number of ridges and ditches	swissALTI3D
	leadstruct.all	Union of all leading structures	swissTLM3D, swissALTI3D
	feeding_Xm	Potential feeding grounds, low vegetation surrounded by cover (%) for buffers 100 m und 200 m	Vegetation height WSL
	visibility	Line of sight model with DSM	swissTLM3D, swissALTI3D und DSM

zonal parameters, we have built several logistic regression models and observed their performance.

This first analysis showed that there are differences mostly between Canton of Zurich and Grisons, therefore we have decided to perform further analysis and build models for Switzerland's mountain and midland regions separately. Zurich and the main part of Fribourg represent the Swiss midlands, while the Canton of Grisons and the remainder of the Canton of Fribourg are representatives of the Alpine region.

Due to differences between the alpine region and the Swiss midlands, separate models were developed. A total of four logistic regression models and two ensemble classifiers (FML for Midland and Alp region separately) were established.

- *KNR*. KDE-based segmentation, *N*arrow kernel *kde<sub>n</sub>*, Logistic Regression model, for two landscape types: KNR Midlands and KNR Alps.
- *KWR*. KDE-based segmentation, *W*ide kernel *kde<sub>w</sub>*, Logistic Regression model, for two landscape types: KWR Midlands and KWR Alps.
- *FML*. Fixed-length segmentation, *M*achine *L*earning tree based classifier, for two landscape types: FML Midlands and FML Alps.

Independent of region and segmentation type, all models were built from randomly selected 70% of the data, while the remaining 30% was used as test data for the validation of the model, analogous to Seiler et al. (2016). All models were evaluated with the following performance metrics: sensitivity (ability to identify hotspots), specificity (ability to identify coldspots) and misclassification error. The goal being to identify as much hotspots as possible, the models were optimized with respect to sensitivity.

Regression models were built for KDE-based road segments, where each hotspot was followed by a corresponding coldspot. The important parameters of the regression models were selected by bidirectional stepwise procedure based on the AIC. Furthermore we also investigated the pseudo  $R^2$  and Area Under the Curve (AUC) for the models.

In the case of regular segmentation, with highly unbalanced data sets building a regression model even with Bayesian approach was not successful. Therefore, for this segmentation tree based machine learning methods were applied. We investigated random forest and ensemble sklearn Extra Trees

classifier with Gini impurity, which resulted in best performance. For the hyperparameter tuning GridSearchCV was used. The models were evaluated with k-fold cross-validation.

#### Software and hardware

All GIS operations were executed with ArcGIS Pro (2.4) and Python 3.6 with the libraries pandas 0.25.1 and geopandas 0.6.0. We furthermore used PostgreSQL 9.4.5, with the extension PostGIS 2.2.1. Spatial data science and statistical modelling used R 3.6.0 with the libraries sf 2.8-1 and tidyverse. Machine learning-based modelling used python 3.6 with scikit-learn 0.23 package.

## Results

### Local factors

The first analysis, including univariate paired Wilcoxon tests and regression models for the three cantons separately, excluded average daily traffic volume and speed due to lack of significance and the amount of missing values (~ 10%). Average traffic volume was found to correlate positively with road noise. These two parameters are, however, indirectly represented in the model via the variable road noise (*noise\_Xm*). Similarly, *road\_category* was not a significant factor in the regression models for all of the cantons and hence not investigated further. The following radii were chosen for the zonal parameters: 50 m for road noise (mean, *noise\_50m*), 100 m for feeding grounds (*feeding\_100m*), 200 m for primary areas (*primary\_areas\_200m*) and 100 m for vegetation height (*vegetation\_height\_100m*). The complex visibility did not significantly improve the model and was only significant for one region in the univariate tests. For this reason, the visibility was excluded from further analysis. It was nevertheless incorporated into the models indirectly via the parameter *sinuosity*.

The univariate tests of individual variables indicated differences between coldspot and hotspot segments. However, the significance of the factors varies from region to region. Table 3 shows a summary of the results of the univariate paired Wilcoxon test for the final variables used for modeling with chosen

**Table 3** Results (p-values) of the univariate paired Wilcoxon tests between hotspots and corresponding coldspots

Variable	Midland KNR	Alps KNR	Midland KWR	Alps KWR
sinuosity	< 0.001*	< 0.001*	0.0013*	< 0.001*
dist.builtup	< 0.001*	0.0696	0.5975	0.8400
dist.forest	< 0.001*	0.0689	0.4469	0.0119*
dist.water	0.7750	0.0089*	0.7346	0.4801
dist.corridor	0.1857	0.0269*	0.5072	0.2111
feeding_100m	< 0.001*	0.0378*	< 0.001*	0.0297*
feeding_200m	< 0.001*	0.1095	< 0.001*	0.0639
vegetation_height_100m	< 0.001*	0.0020*	< 0.001*	0.0947
noise_50m	< 0.001*	< 0.001*	< 0.001*	< 0.001*
primary_areas_200m	< 0.001*	0.0442*	< 0.001*	0.8108
leadstruct.DTM	0.7707	0.8259	NA <sup>1</sup>	NA <sup>1</sup>
leadstruct.TLM	< 0.001*	0.1963	< 0.001*	0.4662
leadstruct.all	0.0023*	0.4124	< 0.001*	0.4662
visibility	< 0.001*	0.6828	NA <sup>2</sup>	NA <sup>2</sup>

\*Significant (significance level  $\alpha = 5\%$ ) <sup>1</sup>exact p-values could not be computed due to ties <sup>2</sup>no data available

**Table 4** Variables included in the regression models ranked in decreasing order of their explanatory power

Midland KNR	Alps KNR	Midland KWR	Alps KWR
dist.forest***	sinuosity***	feeding_100m***	sinuosity**
sinuosity**	dist.forest***	vegetation_height_100m***	feeding_100m
noise_50m**	noise_50m**	noise_50m***	noise_50m*
dist.builtup**	dist.water*	sinuosity***	leadstruct.TLM*
feeding_100m*	primary_areas_200m*	primary_areas_200m**	dist.forest
leadstruct.TLM	feeding_100m*	leadstruct.TLM	
vegetation_height_100m	vegetation_height_100m*		
dist.water	leadstruct.DTM		

Significance: \*\*\* $p < 0.001$ , \*\* $p < 0.01$ , \* $p < 0.05$

radii for the zonal parameters. The parameters that were significantly different independent of model and region are sinuosity, feeding\_100m and noise\_50m. The parameters dist.builtup, dist.forest, feeding\_200m, as well as vegetation\_height\_100m and primary\_areas\_200m are significant only in midland regions. Due to the fact that data in fixed-length segmented model (FML) were highly unbalanced and not paired, we decided not to carry out univariate tests in FML.

The performance of the two KDE-based models is comparable for both regions, Midlands and Alps. However, each of the two models contained a slightly different combination of variables. Regardless of the model, the following

variables are always significant: sinuosity, feeding\_100m and noise\_50m (Table 4). The following parameters (in decreasing order of significance): dist.forest, sinuosity, dist.builtup, noise\_50m, and feeding\_100m are significant in the KNR model for the Midland region. Other variables of the model are not significant ( $p\text{-value} > 0.05$ ), but were still included in the model because of the AIC and better performance (3.75% increase in sensitivity and 2.08% decrease in misclassification error). In the KNR model for the Alpine region, on the other hand, the parameters sinuosity, dist.forest, noise\_50m, dist.water, primary\_areas\_200m, feeding\_100m and finally vegetation\_height\_100m are significant, arranged in

**Table 5** Feature importances in FML model ranked in decreasing order

Midland FML	Alps FML
vegetation_height_100m	road_category_6m
noise_50m	road_category_4m
road_category_6m	sinuosity
feeding_100m	noise_50m
primary_areas_200m	primary_areas_200m
sinuosity	road_category_throughway
road_category_4m	vegetation_height_100m
road_category_throughway	feeding_100m
dist.water	road_category_8m

**Table 6** Included variables in models

Variable	Midland KNR	Alps KNR	Midland KWR	Alps KWR	Midland FML	Alps FML
sinuosity	X	X	X	X	X	X
dist.builtup	X					
dist.forest	X	X		X		
dist.water	X	X			X	
dist.corridor						
feeding_100m	X	X	X	X	X	X
vegetation_height_100m	X	X	X		X	X
noise_50m	X	X	X	X	X	X
primary_areas_200m		X	X		X	X
leadstruct.TLM	X		X	X		
leadstruct.DTM		X				
leadstruct.all						
visibility						
road_category					X	X
traffic						
speed						

decreasing order. Table 4 summarizes the parameters included in the regression models. The classification of the segments between the KNR and KWR models differs in the Midland region in 23% and in the Alpine region in 21% of the segments. In both regions, the KNR model identified twice as many hotspots as the KWR model.

In the classification of the FML model the most significant factors on which the decisions were made were road\_category\_4m and road\_category\_6m, noise\_50m, sinuosity and vegetation\_height\_100m, with slight differences between regions (Table 5). The most important features in the alpine region

were road\_category\_6m and road\_category\_4m, followed by sinuosity. Whereas, in the Midland region the most important features were road\_category\_6m together with noise\_50m, and vegetation\_height\_100m. Feature importance for the two FML models are summarized in Table 5.

Table 6 summarizes all investigated variables with indication of their presence and absence in the models. Table 7 shows the selected best model variants with standardized estimates.

**Table 7** Comparison of models

Model	Sensitivity (%)	Specificity (%)	Missclass. error (%)	Pseudo R <sup>2</sup> (%)	AUC
KNR Midlands	82.50	76.56	20.14	25.95	0.83
KNR Alps	88.64	50.00	27.03	29.75	0.84
KWR Midlands	79.46	57.94	31.05	23.24	0.82
KWR Alps	92.86	34.38	25.49	14.09	0.74
FML Midlands	77.19	76.80	23.18	NA	0.83
FML Alps	76.00	70.56	29.17	NA	0.76

### Model evaluation (predictions)

All models were able to significantly distinguish between hotspots and coldspots. However, all of them were more effective in the identification of hotspots than coldspots (Table 7).

In the Midland region, both regression models (KNR and KWR) can identify the hotspot segments relatively well (sensitivity 82.5% and 79.5%, respectively). The pseudo R<sup>2</sup> and the AUC of these two models are comparable. Higher specificity accounts for less false positives. Together with lower misclassification error, the KNR model is preferable. The performance of the FML models was slightly worse. In the case of the Alpine region, the performance of model KNR is also the best. The ability to identify hotspot segments is similar for both KDE based models (sensitivity 88.6% and 92.7%, respectively). The same applies to the parameters misclassification error and AUC. Specificity and pseudo R<sup>2</sup> are significantly lower in the KWR model (by around 15% each), which means that the corresponding model identifies non-hotspot segments worse and explains significantly less data. Therefore, for the Alpine region the KNR model is also preferable. The models based on tree algorithms are a little less effective in both Midland and Alpine region (sensitivity 77.2% and 76.0%, respectively). With ~ 71% specificity in the Alpine region the FML model results in a lower number of falsely identified hotspots, especially important for such unbalanced dataset. At the same time, sensitivity lower by ~ 14% on average than in the kernel based logistic regression models, results in more hotspots missed.

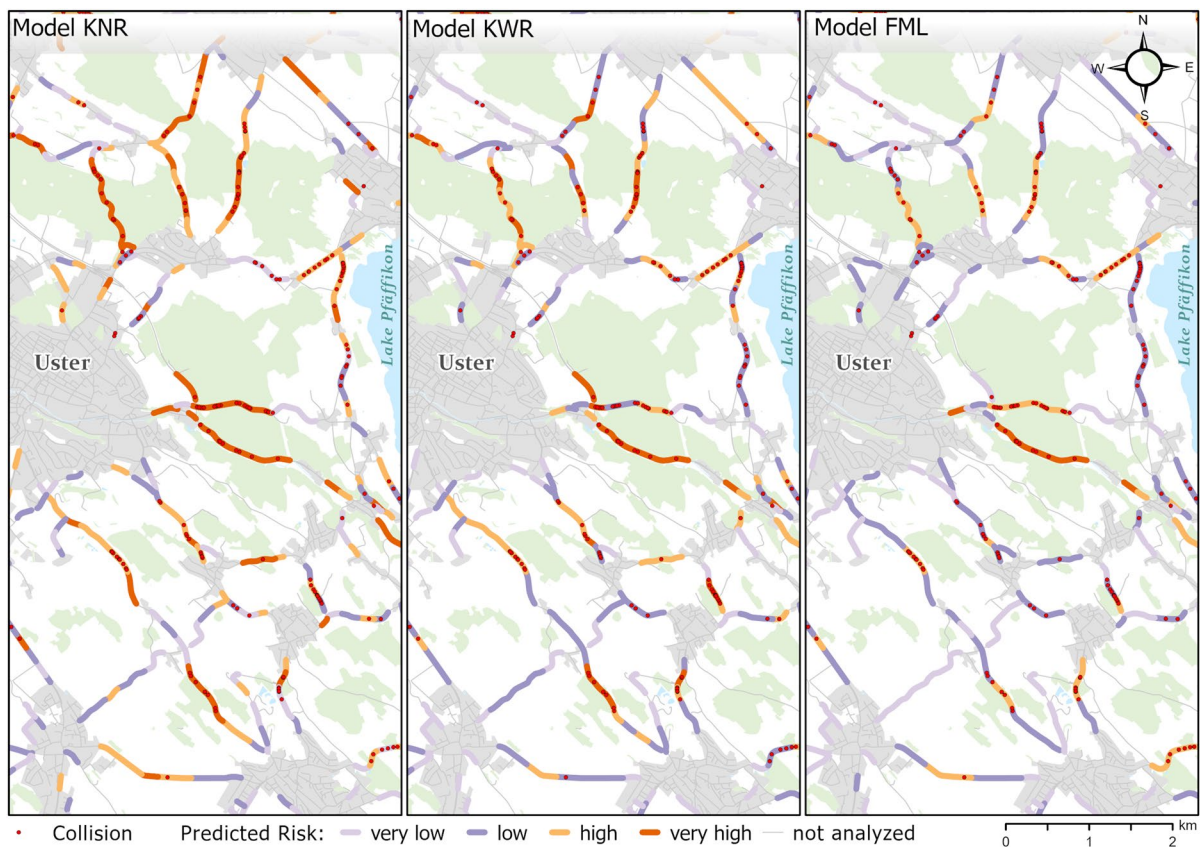
### Exemplary focus areas for Midlands and Alps models

The results section concludes with a closer look at the model results embedded in the geography and landscape ecology of two exemplary focus areas representative for the Midlands and Alps models. The featured maps reveal large-scale details about the strengths and weaknesses of the KNR, KWR, and FML models, allowing a comparison between the Midlands and Alps models, respectively. The Midland focus area is located near Uster, Canton Zurich, and represents the forested colline rural areas of the Swiss Midlands (Fig. 4). The Alps focus area lies near the lower Engadin capital of Scuol, Canton Grisons. It represents the subalpine zone with its elongated valleys stretching several elevation zones, typically traversed with a single major road and some branching minor roads connecting the rural outposts on elevated terraces (Fig. 5).

In order to improve readability of the maps, probabilities are shown as color gradients over discrete probability classes instead of continuous probabilities per segments. The road segments were identified binary as hotspots and coldspots based on the optimal probability cutoff value. The probability ranges of each class were then divided equally into two halves by a higher hotspot probability and a lower hotspot probability. Although this method does not reflect the distribution of the hotspot/coldspot segments, it does improve the visualization of the classified segments. Very high (dark red), high (orange), low (dark purple) and very low (light purple) accident probability can now be distinguished on the map. For orientation the maps show the settlement areas, a generalized road network and the forests.

All Midland models (Fig. 4) show good results for the collision clusters on the main traffic-heavy roads cutting across larger woodlands (variables *dist.*





**Fig. 4** Focus area Midlands. Network of several-traffic heavy main roads connecting provincial towns and cutting through expanded colline forests. All models capture the main hotspots

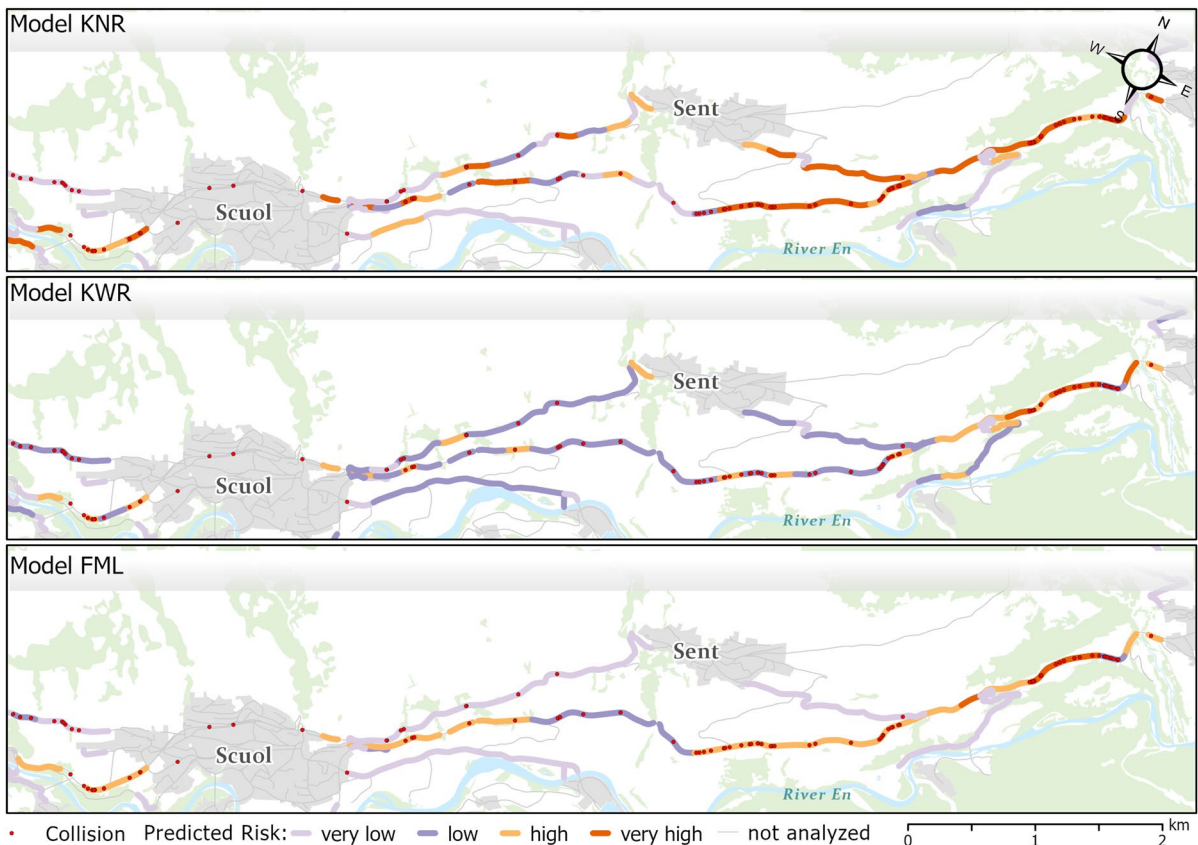
along the main roads, whilst FML misses some collision clusters in the northern forest belt. *Data source* Federal Office of Topography Swisstopo

forest, `noise_50m` as a proxy for traffic load and speed), partially even emphasising curvy sections (`sinuosity`). The prominent missed collision hotspots along Lake Pfäffikon (some success in KNR, less in KWR and even worse in FML) are indicative of noticeable deficiencies of the models in open agricultural sections. Here the lack of up-to-date geodata on annual crop changes limits the models' power. FML picks up on most hotspots, but mainly predicts the *high* risk category, hardly ever *very high*. The five most important features of the Midland FML model are related to vegetation and traffic infrastructure (`vegetation_height_100m`, `noise_50m`, `road_category_6m`, `feeding_100m`, as well as `primary_areas_200m`), whereas `sinuosity` and `dist.water` are less important. This may explain FML's solid results with the main hotspots along the fast main roads cutting through the forest

patches but also its failure on the curvy road along the Lake.

The Alps focus area is dominated by the main road spanning the lower Engadin Valley from the SW to the NE (the continuous central road with the majority of the collisions, Fig. 5). The much less busy roads reaching out to the smaller villages on the elevated terraces (e.g., Sent) feature much less collisions. All models pick up on this general pattern, again with KNR achieving the best results, here followed by FML. The better performance of the FML model in the Alps may be explained by its advantage of including explicit road categories as features (see Table 5, features `road_category_6m`, `road_category_4m`). The maps also illustrate the absence of forest and the much less dense settlement pattern in the mountainous areas explaining the reduced importance of `dist.forest` and `dist.builtup` in all Alps models. `Sinuosity`, the most important





**Fig. 5** Focus area Alps. Elongated lower Engadin Valley with main road trough larger villages in valley floor (Scuol) and smaller villages on the elevated terraces (Sent). KNR picks up

on most hotspots along the main road but also predicts some false positives, FML performs rather well. *Data source* Federal Office of Topography Swisstopo

variable of the KNR Alps model, may also explain the false positive hotspots along the curvy roads to sent in KNR.

## Discussion

In comparison to many related previous studies on WVC that used up to 50% categorical factors (e.g., Bíl et al. 2019; Seiler et al. 2016) our study used almost no categorical data with the only exception of road type. Instead, it could rely entirely on areal geospatial variables with excellent spatial and semantic granularity due to very good geodata availability. For example, instead of relying on qualitative information on the vegetation cover next to road interpreted from GoogleMaps or other aerial imagery we were able to compute a similar indicator from the

vegetation structure based on a nationwide LiDAR dataset (*feeding\_Xm*). This enhanced spatial and semantic granularity forced us to invest much more in the development of spatial neighbourhood functions relating the environmental factors to the hot-spot/coldspot segments, hence producing methodological progress regarding semantic road annotation neighbourhood functions tailored for WVC analysis. In that regard, a further methodological contribution comes in the form of the operationalization of leading structures, based on linear landscape features oriented towards the road segments.

The selection of variables simplified the predictive process. Using simple road sinuosity over the computationally expensive visibility approach significantly reduced the computation load. Similarly, *noise\_Xm* served as an excellent proxy for traffic volume and speed. Comparing the Midlands and Alp

models reveals several differences in their landscape ecology impacting on collision risk. Distance to settlement (`dist.builtup`) features only in Midland KNR, but in no Alps model where settlements are much sparser and have hence little predicting power. Similarly, distance to water (`dist.water`) is more important in mountainous areas whereas in the lowlands water features are ubiquitous and hence less predictive.

Since leading structures were of special interests to the experts, the influence of these parameters on model performance was investigated in detail. Including the leading structure was decreasing the AIC in all models, resulting in slightly lower misclassification error and slightly higher pseudo  $R^2$ . On average, either sensitivity or specificity was increased by 1.5%, except for KWR Midland. In this model specificity was increased by 5% with a decrease in sensitivity of 6%.

When comparing between a KDE data-driven segmentation approach with a fixed-length segmentation the performance of the models is comparable, with a slight favor of KNR model. This is mostly due to the KNR model's higher explanatory power compared to other models. This is not surprising as the way the data was prepared for both models is different. Whereas with the KDE models the segmentation is fitted to the hotspots, with the fixed-length segmentation the actual segment cutting points are somewhat arbitrarily positioned, resulting, for example, in splitting a hotspot into two neighboring segments, both however with lower collision counts. This issue can be considered as a one-dimensional case of the modifiable areal unit problem (MAUP) known in spatial data science, namely the effect of unstable patterns in choropleth maps when point data is aggregated to varying administrative units (Cressie 1996). Furthermore, logistic regression may outperform decision tree algorithms on smaller data sets and low signal to noise ratios. However, when increasing the size of a more imbalanced dataset the forest algorithms outperform it, which was the case in the regular segmentation approach.

For the current paper our validation of the models is limited to the three Cantons where we have access to actual collision data with the required spatial precision. Within Zurich, Fribourg and Grisons data-splitting was applied for all models, referred to as historical data validation in Rykiel (1996), resulting

in the performance numbers in Table 7. However, no ground truth data is available for the rest of Switzerland covered by our models.

Wildlife accidents often occur in situations where roads cross favorable wildlife habitats (Malo et al. 2004; Garrah et al. 2015). The animal species studied here are all bound to the forest, at least at lower altitudes. Our models performed good in forest areas distant from built-up areas and in the presence of feeding grounds. We had less good results with segments embedded in agricultural areas, due to a lack of up-to-date data on annual crop changes. Some crops can have a double impact on collision occurrence, with this impact even showing a positive feedback. A corn field next to a road for example represents a severe collision risk. On one hand there are more animals attracted because of the food and shelter that a field may offer. This is relevant for roe deer but also and increasingly for wild boar and red deer that are expanding their range in the lowlands of Switzerland (Graf et al. 2021). On the other hand the visibility is significantly reduced. Due to crop rotation the presence of different crops changes from year to year and also within one year the growth status of the crops and thus their attractiveness for wildlife changes. Not having annual crop data we are missing out important collision causes. As an alternative to crop data, species abundance could have been included as an even more direct predictor. However, this information is not available in a homogeneous form throughout the country and many cantons lack data with the required granularity for our purpose. We included abundance at least in an indirect form by using variables like forest, hedges, distance to corridors and vegetation height.

When comparing our ecological implications with related work, we find consent with related studies. Our overall most important factors are road sinuosity, food availability and traffic noise, in accordance with the related studies (Seiler et al. 2016; Bíl et al. 2019). In our study road sinuosity turned out to be an important factor increasing the risk of WVC. The more curvy a road is, the greater the risk of wildlife accidents. This can be explained by the fact that when the road is curvy, both drivers and wildlife recognise a potential hazard later compared to the situation on a straight road. Early hazard recognition is crucial for the prevention of accidents. It has already been shown that unfavourable visibility conditions caused

by sinuosity, vegetation or weather conditions can strongly influence the accident risk (van Langevelde and Jaarsma 2005; Barrientos and Bolonio 2009; Laliberté and St-Laurent 2020).

Low vegetation below 1 m surrounded by high vegetation, such as we find along forest edges, leading structures (hedges, groups of trees, streamside vegetation) or clearings are attractive areas for ungulates as they find both food and cover. If such grazing areas are located near transport infrastructures, wild animals stay close to the danger zone or even cross it. In comparable studies in Sweden and the Czech Republic, grazing areas also seem to be an important influencing factor in connection with wildlife accidents (Seiler et al. 2016; Bíl et al. 2019). In a study in Spain, grazing does not seem to be an important influencing factor, which can be explained by the main species studied there, the wild boar (Seiler et al. 2016). Wild boar also search for food in pastures or meadows, but are not interested in the actual grazing, but in invertebrates found in the soil.

Road noise is related to speed and traffic volume, which in turn are two important factors influencing accident risk (van Langevelde and Jaarsma 2005; Elvik 2008; Barrientos and Bolonio 2009; Huijser et al. 2015; Seiler et al. 2016). If speed increases, the time between the detection of a hazard and the potential collision between vehicle and wildlife decreases. The braking distance and impact energy also increase with speed. Average daily traffic volume (DTV) and speed were excluded from the initial analyses due to lack of significance. The lack of significance can be explained by the fact that both the hotspots and non-hotspots are located on main roads and therefore have similar speeds, and in addition the resolution of the data set is not sufficient for our analyses due to the punctual measurements. Also in terms of traffic volume, the measurement network may not be dense enough to prove the influence of this factor. Alternatively, the so-called *deadly trap hypothesis*, which predicts an increased number of falling deer at medium traffic volumes, could explain why no correlation between traffic volume and risk of wildlife accidents was found (Iuell 2003).

However, it is not the absolute number of vehicles on a road section (vehicles per day) that is decisive for the accident risk, but the time when the vehicles are on the road, because wildlife is mainly on the road at dusk and during the night (Bíl et al. 2020).

Studies show that the frequency of wildlife accidents varies depending on the time of year and time of day (Garrah et al. 2015) and that this seasonality is strongly species-dependent (Laliberté and St-Laurent 2020). Depending on the species, different roadside vegetation or other characteristics such as road salt as a measure for winter maintenance also influence the occurrence of wildlife accidents (Grosman et al. 2009; Gunson et al. 2011; Bíl et al. 2020).

In our models, wildlife corridors were not a significant factor in causing more accidents compared to other studies (Seiler et al. 2016). This could be explained by the fact that many roads also cross optimal wildlife habitats in places where there are no wildlife corridors. It is also possible that measures to reduce wildlife accidents have preferably been taken and implemented in wildlife corridors already.

## Conclusions and implications

In this study we present an extensive WVC study making use of very rich regional collision data sets, excellent fine-grained landscape geodata, using novel spatial neighbourhood functions for annotating road segments and machine learning for up-scaling the regional data to a national model. In accordance with most related studies, but based on geodata with much improved spatial and attributal granularities, we identified road sinuosity, browsing/forage availability, and traffic flow as key factors for WVCs. Our best models achieved sensitivities of 82.5% to 88.6%, with misclassifications of 20.14% and 27.03%, respectively. Our results also highlighted intrinsic limitations of modelling WVCs from land-cover data, especially in areas with transient vegetation cover (annual crop rotation). Further limitations arose from inhomogeneously collected collision data, adding uncertainties about the spatial and temporal precision and accuracy of collision incidents.

Our paper makes both methodological and ecological contributions to the theory of WVC. From a methodological perspective, we illustrate the added value of using fine-grained land-cover and ecological data. We also show how such detailed information can be annotated to road segments using spatial neighbourhood functions. Such functions can be implemented as straightforward buffer operations or more complex

models, as illustrated with the leading structures. The experimental section of our paper furthermore compares two different approaches of road segmentation, collision data-driven KDE and fixed-length segmentation, both for two different landscape types (midlands vs. alps) and combined with multiple aspects of sensitivity analysis (variable segmentation kernels, variable neighbourhood thresholds). This comparative experimental section illustrates and quantifies the importance and implications of modelling choices, key aspects of WVC analysis often overlooked.

As major ecological contributions we extrapolated solid national WVC models from three rich but heterogeneous regional data sets, with sensitivities beyond 82%. Our models and the therefore selected variables are in accordance to related work, acknowledging regional characteristics. We identified the most important collision factors for the studied Swiss landscape types (road sinuosity, browsing/forage availability and traffic noise), with few but interesting differences between midland and alpine landscapes (e.g. distance to built-up area is less important in sparsely populated mountain areas).

Even though our study benefited from very good geodata availability, it also highlighted key aspects that could further improve WVC modelling. First, annual crop data would be of outmost interest, we plan on using Sentinel data for that purpose (see e.g. Sigrist et al. 2022). Such data could also serve for modelling the abundance of target species throughout the country and for the different seasons. This spatially and temporally explicit estimation of target species abundance could improve the prediction of hotspots of WVC's, especially in regions dominated by agriculture. Other data sets available for some test regions but not for others were excluded from the beginning. For example, road illumination is so far only available for some regions, but would be a great asset when becoming available nationwide. Harmonization of collision data capture protocols will further extend the range of analytical options. Most importantly here is the harmonization of the time stamps (time of day) to include diurnal collision patterns. In future work we intend making use of additional ground truth data that will be made available for areas beyond our three Cantons. We plan to closely collaborate with collectors of the collision data nationwide, making sure the insights from our study help homogenizing data capture protocols whilst widening

the coverage of data collision data collection for validation purposes.

Since both drivers and wildlife tend getting accustomed to permanent warning measures, such as road signs or fix installed reflectors (Huijser et al. 2015; Benten et al. 2018), our study helps positioning more effective but expensive interactive prevention measures: warning systems alerting wildlife to approaching cars, and vice versa alerting drivers to present and active wildlife. Hence, our prediction maps will be used for pre-selecting collision hotspots, that will then be further investigated via in situ analysis and local decision makers.

**Acknowledgements** The authors thank all funding bodies listed below and an anonymous foundation for making this work possible. The authors also thank the Institute of Natural Resource Sciences, the Zurich University of Applied Sciences ZHAW for supporting this work. Finally, the authors also thank the anonymous reviewers for their constructive feedback.

**Author contributions** All authors contributed to the study conception and design. Material preparation and data preprocessing, and integration was done by NR. The spatial data science methods were conceptualized and implemented by PL and NR. The statistical modelling was conceptualized and performed by IK and AW. The first draft of the manuscript was written by PL, with significant contributions by SS, AW, and IK, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** Open access funding provided by ZHAW Zurich University of Applied Sciences. This work was supported by the research project *Prevention of animal vehicle collisions on traffic infrastructures* (VSS 2015/212), funded by the Swiss Federal Roads Office FEDRO, the Swiss Federal Office of the Environment FOEN, and the Swiss Federal Railway SBB.

**Data availability** The collision data used in this study and all datasets generated during the current study can not be made publicly available due to non-disclosure agreements with the original data providers, but can be requested from the corresponding author on reasonable request. All geodata used in this article (see Table 2) is available via the respective official data providers, that includes the Swiss National Mapping Agency Swisstopo (<https://www.swisstopo.admin.ch/>), the Swiss Federal Office for Spatial Development ARE (<https://www.are.admin.ch/>), the Swiss Federal Office for the Environment FOEN (<https://www.bafu.admin.ch/>), and the Swiss Federal Institute for Forest, Snow and Landscape Research WSL (<https://www.wsl.ch/>).

**Declarations**

**Competing interests** The authors have no relevant financial or non-financial interests to disclose. Not applicable.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Ascensão F, Yogui D, Alves M, Medici EP, Desbiez A (2019) Predicting spatiotemporal patterns of road mortality for medium-large mammals. *J Environ Manag* 248:109320
- Barrientos R, Bolonio L (2009) The presence of rabbits adjacent to roads increases polecat road mortality. *Biodivers Conserv* 18(2):405–418
- Benten A, Annighöfer P, Vor T (2018) Wildlife warning reflectors' potential to mitigate wildlife–vehicle collisions—a review on the evaluation methods. *Front Ecol Evol* 6:37
- Bíl M, Andrášik R, Janoška Z (2013) Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. *Accid Anal Prev* 55:265–273
- Bíl M, Andrášik R, Svoboda T, Sedoník J (2016) The KDE+ software: a tool for effective identification and ranking of animal–vehicle collision hotspots along networks. *Landsc Ecol* 31(2):231–237
- Bíl M, Andrášik R, Dul'a M, Sedoník J (2019) On reliable identification of factors influencing wildlife–vehicle collisions along roads. *J Environ Manag* 237:297–304
- Bíl M, Kubeček J, Andrášik R (2020) Ungulate–vehicle collision risk and traffic volume on roads. *Eur J Wildl Res* 66(4):1–10
- Castro M, De Santos-Berbel C, Iglesias L (2017) A comprehensive methodology for the analysis of highway sight distance. In: *Transport infrastructure and systems: proceedings of the AIIT international congress on transport infrastructure and systems, Rome, Italy, 10–12 April 2017*. CRC Press, p 193
- Cressie NA (1996) Change of support and the modifiable areal unit problem. *Geogr Syst* 3(2–3):159–180
- Elvik R (2008) A survey of operational definitions of hazardous road locations in some European countries. *Accid Anal Prev* 40(6):1830–1835
- Garrah E, Danby RK, Eberhardt E, Cunningham GM, Mitchell S (2015) Hot spots and hot times: wildlife road mortality in a regional conservation corridor. *Environ Manag* 56(4):874–889
- Graf R, Fischer C (eds) (2021) *Atlas der Säugetiere - Schweiz und Liechtenstein* [Atlas of mammals. Switzerland and Liechtenstein]. Swiss Society of Wildlife Biology, Haupt, Berne
- Grosman PD, Jaeger JA, Biron PM, Dussault C, Ouellet J-P (2009) Reducing moose–vehicle collisions through salt pool removal and displacement: an agent-based modeling approach. *Ecol Soc*. <https://doi.org/10.5751/ES-02941-140217>
- Gülci S, Emin AA (2015) Assessment of ecological passages along road networks within the Mediterranean forest using GIS-based multi criteria evaluation approach. *Environ Monit Assess* 187(779):1–23
- Gunson KE, Mountrakis G, Quackenbush LJ (2011) Spatial wildlife–vehicle collision models: a review of current work and its application to transportation mitigation projects. *J Environ Manag* 92(4):1074–1082
- Huijser M P, Mosler-Berger C, Olsson M, Strein M (2015) Wildlife warning signs and animal detection systems aimed at reducing wildlife–vehicle collisions. In: *Handbook of road ecology*. Wiley, Hoboken, p 198–212
- Iuell B (2003) Wildlife and traffic—a European handbook for identifying conflicts and designing solutions. In: *The XXII PIARC world road congress, 2003*. World Road Association PIARC, p 1–11
- Jung J, Olsen MJ, Hurwitz DS, Kashani AG, Buker K (2018) 3D virtual intersection sight distance analysis using LiDAR data. *Transp Res C* 86:563–579
- Kubicka M, Cela A, Mounier H, Niculescu S-I (2018) Comparative study and application-oriented classification of vehicular map-matching methods. *IEEE Intell Transp Syst Mag* 10(2):150–166
- Laliberté J, St-Laurent M-H (2020) In the wrong place at the wrong time: moose and deer movement patterns influence wildlife–vehicle collision risk. *Accid Anal Prev* 135:105365
- Malo JE, Suárez F, Díez A (2004) Can we mitigate animal–vehicle accidents using predictive models? *J Appl Ecol* 41(4):701–710
- Neumann W, Ericsson G, Dettki H, Bunnefeld N, Keuler NS, Helmers DP, Radeloff VC (2012) Difference in spatiotemporal patterns of wildlife road-crossings and wildlife–vehicle collisions. *Biol Conserv* 145(1):70–78
- Roth H (2012) *Präventionsprojekt weniger wildtierunfälle*. Technical report. Swiss Private Insurers, Zurich
- Rykiel EJ Jr (1996) Testing ecological models: the meaning of validation. *Ecol Model* 90(3):229–244
- Seiler A, Rosell C, Torellas M, Sjölund M, Rolandsen CM, Solberg E, Van Moorter B, Lindström I (2016) Case studies on the effect of local road and verge features on ungulate–vehicle collisions. Technical report
- Sigrist B, Signer C, Wellig SD, Ozgul A, Filli F, Jenny H, Thiel D, Wirthner S, Graf RF (2022) Green-up selection by red deer in heterogeneous, human-dominated landscapes of Central Europe. *Ecol Evol* 12(7):e9048
- van Langevelde F, Jaarsma CF (2005) Using traffic flow theory to model traffic mortality in mammals. *Landsc Ecol* 19(8):895–907

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.