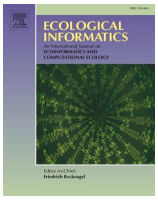
Ecological Informatics 63 (2021) 101291

Contents lists available at ScienceDirect

Ecological Informatics

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



An artificial intelligent framework for prediction of wildlife vehicle collision hotspots based on geographic information systems and multispectral imagery

Juan C. González-Vélez [a,*], Maria C. Torres-Madronero [b], J. Murillo-Escobar [c], Juan Carlos Jaramillo-Fayad [a]

[a] ALQUIMIA - Instituto Tecnológico Metropolitano - ITM, Medellín, Colombia
[b] MIRP Lab - Instituto Tecnológico Metropolitano - ITM, Medellín, Colombia
[c] Gi2b - Instituto Tecnológico Metropolitano - ITM, Medellín, Colombia

ARTICLE INFO

Keywords:
Spatial analysis
Machine learning
Pattern recognition
Wildlife-vehicle collision
Multispectral imagery

ABSTRACT

Wildlife-vehicle collision - WVC is a phenomenon that arises from the fragmentation of ecosystems by roads, limiting the mobility of individuals and putting at risk the stability of populations by increasing mortality. Colombia is not unaware of the problem of the WVC, evidenced in different scientific publications that describe the WVC in the roads of the country. Although the rise of artificial intelligence has significant advances in the prediction of spatial phenomena in recent years, it has not yet been sufficiently explored by Road Ecology. For this reason, this research aimed to develop a methodology to predict the sites of accumulation of WVC in eastern Antioquia, Colombia, based on artificial intelligence algorithms, geographic information systems - GIS, and multispectral image processing. During the development of this research, it was identified that the features most related to the WVC in the study area are: Distance to Forest, Distance to Biological Corridor, Ground Resistance to Movement, Cost of Movement, the bands of the Landsat 8 satellite: 9, 10, 11 and the normalized burning index (NBRI). Different machine learning algorithms were compared (k-nearest neighbours, support vector machines (SVM), random forests (RF), and artificial neural networks). SMOTE and ADASYN balancing techniques were applied. The results allowed to identify that the RF algorithm with ADASYN yielded the best performance when subjected to spatial-wise cross-validation (AUC-ROC 0.78 $\pm$ 0.12), surpassing the results of current state-of-the-art. Finally, the methodology was validated through a transfer learning experiment, training the RF-ADASYN algorithm with three zones of the eastern Antioquia region and validating on a different section (AUC-ROC = 0.87 $\pm$ 0.09), retraining the initial model with 5% of data from the validation database.

1. Introduction

Wildlife-vehicle collision - WVC is a phenomenon that is proportional to the roads growth and arises from the fragmentation of ecosystems, limiting the mobility of individuals and putting at risk the stability of populations by increasing mortality (Jaeger, 2015). WVC has severe consequences for ecosystems due to the loss of ecosystem services such as pest control, population control, seed dispersal, among others (Coffin, 2007). This phenomenon also affects road safety: WVC causes injuries, costs associated with vehicle repair, and human lives lost. It is estimated that approximately 2 million collisions between vehicles and large mammals occur in the United States each year, resulting in at least 29,000 people injured, 200 or more human deaths (van der Ree et al., 2015a, 2015b), as well as economic losses estimated at $4 billion each year (Cramer et al., 2015; van der Ree et al., 2015a, 2015b). An estimated 365 million vertebrates die each year on this country's roads (Davenport and Switalski, 2006). Due to this problem, some authors have considered WVC to be one of the main factors contributing to the loss of biodiversity (Laurance et al., 2014). For this reason, it is necessary to generate measures that mitigate the adverse effects of linear infrastructure on ecosystems and wildlife (Clevenger and Waltho, 2005).

One of the primary purposes of conservation researchers, road ecologists, and road infrastructure managers is identifying sites with the highest risk of WVC. These sites are identified through diagnostic

* Corresponding author.
E-mail address: juancgonzalez@itm.edu.co (J.C. González-Vélez).

https://doi.org/10.1016/j.ecoinf.2021.101291
Received 1 February 2021; Received in revised form 23 March 2021; Accepted 23 March 2021
Available online 16 April 2021

studies, which require plenty of resources, specialized human talent, and long periods to obtain systematic and significant information. Based on the information collected, spatial analyses are generated that allow mitigation and prevention measures for the WVC (Crawford et al., 2014; Cureton and Deaton, 2012; Danks and Porter, 2010; Forman et al., 2003; Girardet et al., 2015; Gunson et al., 2011; Madsen et al., 2002).

One of the most promising areas to be applied to this phenomenon is Artificial Intelligence: a set of learning techniques and algorithms that seek to give computer systems the ability to learn. Just as we humans learn from our experiences and the world, systems 'learn' from the data provided to them, allowing them to generate predictions, analysis, or even decision-making (Müller and Guido, 2016). In recent years there has been an exponential increase in the generation of research that makes use of learning algorithms as an input for the prediction of various spatial phenomena (Amiri et al., 2019; Bui et al., 2016; Bui et al., 2017; Bui et al., 2019; Durduran, 2010; Ghorbani et al., 2019; Harirforoush and Bellalite, 2019; Jaafari et al., 2019).

Despite this, few studies use machine learning algorithms to predict the segments with the most significant accumulation of WVC (Pagany, 2020). Most of these studies use generalized linear models, regression techniques, and some classification algorithms. For instance, in (Pagany et al., 2020), Gaussian Naive Bayes, stochastic gradient descent and random forest with random cross-validation are used to predict WVC data. (Nguyen et al., 2021) used a logistic generalized linear model and random forest to predict WVC in Southern Tasmania. Moreover, (Serrón et al., 2020) used a random forest to predict and analyze WVC data in Uruguay. However, the previous works are oriented to predict WVC data, not the significant hotspots. Additionally, the used cross-validation methodology in these previous studies did not address the problems presented by data's spatial correlation, resulting in high performances due to overfitting (Schratz et al., 2019).

This research aims to propose a methodology to predict the sites of accumulation of WVC in eastern Antioquia, Colombia. Our methodology uses artificial intelligence algorithms, geographic information systems - GIS, and multispectral image processing. We performed group-based cross-validation to consider the spatial bias of the data. This research differs from other works by its use of classification algorithms, which road ecology has not sufficiently explored despite proven effective in other spatial-based researches e.g. (Bui et al., 2019). It also applies a novel validation technique for WVC data prediction, considering the spatial bias introduced by the spatial correlation, which to the best of our knowledge, has not been explored before. Finally, this paper shows an easy, replicable, and scalable methodology that reduces the costs and time required to identify the most significant WVC Hotspots. We describe how to learn and then transfer a known pattern to an unknown area using spatial information obtained from official maps or satellite imagery, potentially reducing the cost of identifying WVC in a non-studied road up to 95%.

2. Methods

Fig. 1 shows a flowchart for the methodology proposed in this work, consisting of a recollection of WVC reports, a geostatistical analysis of the point pattern, a hotspot identification stage, a characterization of the most relevant spatial descriptors of the hotspots, followed by a machine learning algorithm phase, a cross-validation algorithm, and a transfer learning stage. Each of these stages are explained in the following section.

2.1. Study area

The study area is a 71 km road network, with a high flow of vehicles, connecting the municipalities of Envigado, La Ceja, El Carmen de Viboral, and Rionegro in the San Nicolás Valley (Colombia). This region presents temperatures between 9 and 24 °C. The area has secondary vegetation cover, agricultural mosaics, pastures, forest plantations, and open forests. Thus, it is an area with a high presence of fauna. The most common animal species in the area are the Red-tailed Squirrel (*Notosciurus granatensis*), the Central American Agouti (*Dasyprocta punctata*), the Common Possum (*Didelphis marsupialis*), the Paca (*Cuniculus paca*), and the Mountain Dog (*Potos flavus*) (García-Morera and Giraldo-Iral, 2018).

2.2. WVC and spatial information recollection

This paper used WVC reports from the Recosfa App dataset (RECOSFA, 2019). Each report includes the spatial coordinates (latitude, longitude), photographs, the animal class, and the species (when possible to identify). The data is obtained from car surveys carried out by ITM researchers and reports from several road administrators between 2016 and 2020 following the recommendations outlined by (Smith and van der Ree, 2015). For August 2020, the Recosfa App database included 6204 WVC reports in Colombia and 837 in the Study Area, consisting of 527 mammals, 178 birds, 82 amphibians, 47 reptiles, and 3 reports whose class could not be identified. Likewise, *Didelphidae* was the most reported with 335 records, followed by *Rhinella marina* and *Notosciurus granatensis*. Fig. 2 shows the study area and the division segments. Table 1 summarizes the number of WVC reports contained in each segment.

We used several environmental and geographic information data to characterize the WVC phenomenon. Table 2 summarizes the information layers, the year of construction, source, and derived spatial information. We used road and river layers collected in 2017 (DANE, 2017), a digital elevation map, tree cover loss data, and land cover maps. Using this data, we obtained the distance to roads and rivers, a watershed model, the distance to cover loss, and the nearest forest. Figs. 3 and 4 show some of the spatial information maps used in this study.

Additionally, a Landsat 8 image composite was derived from data collected between 2014 and 2018; the composition was obtained using

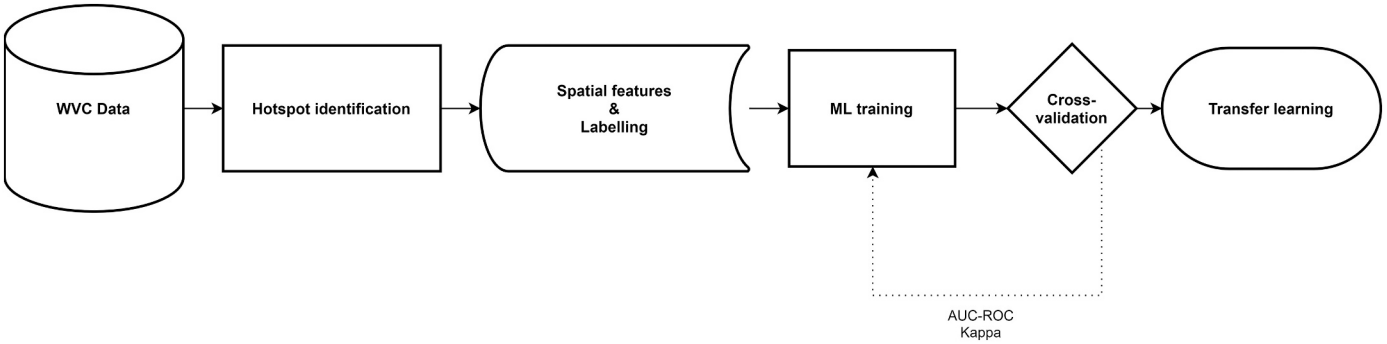


Fig. 1. Proposed methodology to predict WVC hotspots using machine learning algorithms.

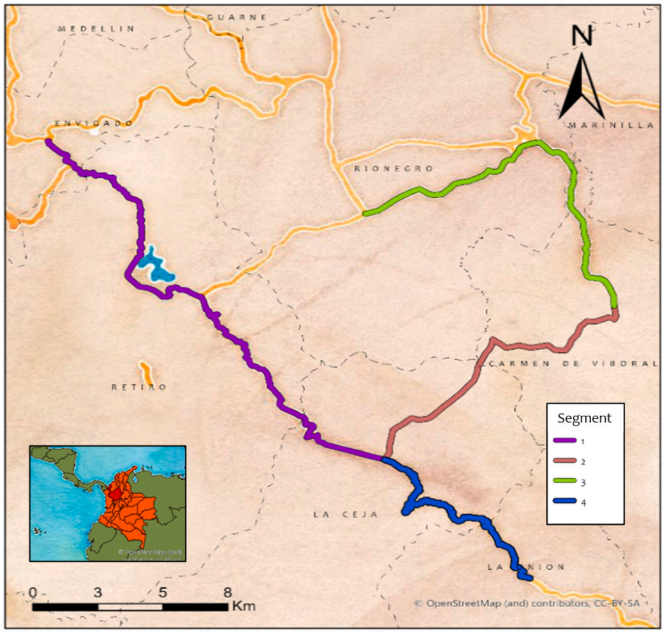


Fig. 2. Map of the study area: San Nicolás Valley (Colombia). The colored segments correspond to the division of the study area on 4 linear segments.

Table 1
Number of WVC reports by each segment.

Road segment	Number of reports
1	261
2	199
3	196
4	181

Table 2
Spatial Information recollected and derived spatial information.

Layer	Year	Source	Derived layers
Roads	2017	(DANE, 2017)	Euclidean distance
Rivers	2017	(DANE, 2017)	Euclidean distance
DEM	2000	(USGS, 2000)	Altitude, Watershed model
Tree cover	2017	(Hansen et al., 2013)	Distance to tree loss
Forest/no forest	2016	(IDEAM, 2016)	Distance to forest
Landsat 8 composite	2014–2018	(USGS, 2013)	Bands 1–11, Spectral indices
Land cover	2017	(IDEAM, 2017)	Land cover classification

Google Earth Engine Simple composite algorithm for Landsat raw imagery with a maximum of 5% of cloud score. This algorithm applies standard top of atmosphere (TOA) calibration and then assigns a cloud score to each pixel using the SimpleLandsatCloudScore algorithm. It selects the lowest possible range of cloud scores at each point and then computes per-band percentile values from the accepted pixels.

Several spectral indices were derived from the composite, such as Normalized Vegetation Index - NDVI (Jackson, 1983), Green Normalized Vegetation Index - GNDVI (Gitelson et al., 1996), Enhanced Vegetation Index - EVI, Advanced Vegetation Index - AVI, Soil-adjusted vegetation index - SAVI, Normalized Difference Moisture Index - NDMI, Moister Stress Index - MSI, Green Cover index - GCI, Bare Soil Index - BSI and Normalized Burn Rate Index - NBRI (Baynes, 2004).

To model animal movement, a least-cost of movement map was generated for the study area. This model was made using the Linkage Mapper toolbox for Arcmap 10.6 (McRae et al., 2008). A reclassification of the GIS layers according to the ethology of the target species: crab-eating fox (*Cerdocyon thous*) was made due to his constant presence in the study area (see Supplementary material) as suggested by (Beier et al., 2011), in which each of the map's pixels acts as a resistance network in which the path with the least voltage loss will be found (Dickson et al., 2018).

The shows some of the spatial information maps collected including vegetation indexes such as NDVI, a Landsat 8 cloudless composite image, a digital elevation map for the study area, WVC reports collected, distance to rivers, distance to forest, distance to tree loss, and distance to roads among others.

2.3. Geostatistical analysis and WVC hotspot identification

Based on the WVC reports, we performed a pattern analysis to identify areas with statistically significant WVC point accumulations. A K Ripley analysis (Clevenger et al., 2003) was implemented using the Siriema Software (Coelho et al., 2014), with an initial observation distance of 100 m and increments of 100 m until convergence. Clusters were generated on the spatial scale identified by Moran autocorrelation test (Eshel, 2011).

Finally, a 2D hotspot analysis was also performed using the Siriema Software. For this, the significant distance band identified by the spatial autocorrelation analysis and a division of 1000 equidistant segments in the study area were used. A point distribution was obtained with its corresponding segment accumulation values - HS, and the upper and lower confidence limits (UCL and LCL, respectively). We generated two classes: class 1 corresponds to road segments whose HS exceeds UCL, and class 0 includes the remaining segments, corresponding to the hotspot and not-hotspot segments.

2.4. Feature sampling and selection

We associated each hotspot and not-hotspot segments with the information layers and derived variables. This representation space contains only spatial information without WVC data, allowing a classification based on the area's features and not on the distribution of WVC points. The feature dataset contained each segment's mean feature value extracted at different circular buffer radios: 90 m, 150 m, and 300 m through focal statistics as shown by (Ha and Shilling, 2018). A matrix of 3000 track segments and 96 features were created, corresponding to the spatial descriptors of the segments proposed in Table 2 (see Supplementary material).

Then, feature selection techniques such as Mutual Information (MI) (Qian et al., 2020), Chi-square (Bahassine et al., 2020), and ANOVA F-score (Güneş et al., 2010) were applied to identify the most relevant information for hotspot identification. A random forest classifier was used to select the best subset of features, comparing the obtained area under the curve (AUC) of the receiver operating characteristic (ROC). The classifier was iteratively trained with a subset of features to select the optimum number of features.

A validation process was designed to reduce the effects of spatial bias caused by the proximity between segments. We evaluated the model's predictive capacity using group-wise cross-validation techniques (Pedregosa et al., 2011), using 4 training and validation folds. Each fold has 9 training and 3 testing segments. These segments are exchanged until each of them, at least once, is part of the testing set.

2.5. Classification comparison and selection

Since there is a higher number of non-hotspot segments, we applied balance techniques such as adaptive synthetic sampling - ADASYN (He et al., 2008), minority synthetic oversampling technique - SMOTE (Chawla et al., 2002), KMeans SMOTE (Douzas et al., 2018), Borderline SMOTE and SVM SMOTE (Nguyen et al., 2011). For the last one, it was necessary to estimate the SVM parameters embedded in the algorithm

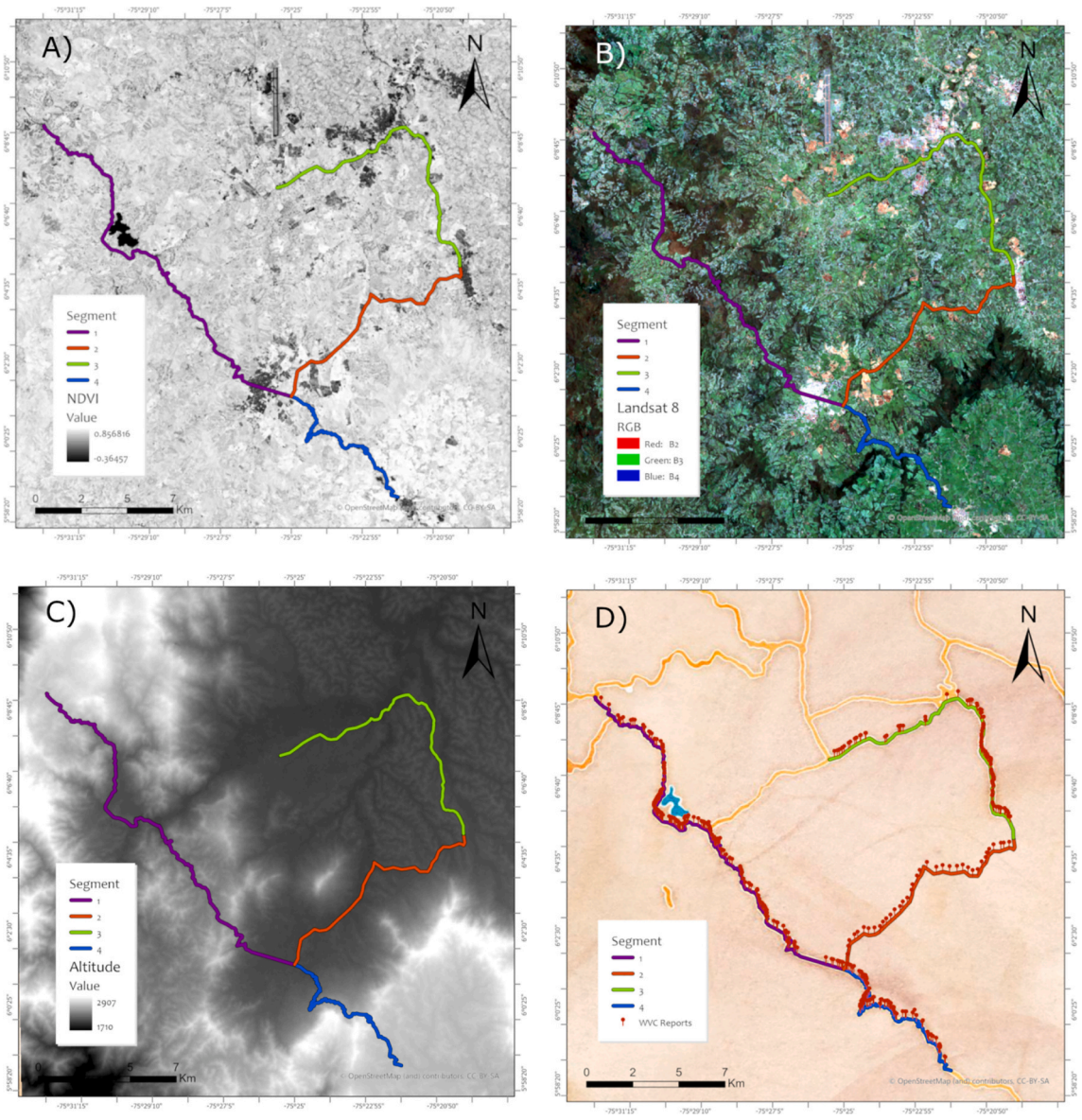


Fig. 3. Spatial information maps for Study Area. A) NDVI index, B) Landsat 8 composite image, C) Altitude, D) WVC reports.

using a GridSearchCV approach (Pedregosa et al., 2011).

We compared four supervised classifiers: K - Nearest Neighbours - KNN (Guo et al., 2003), Support Vector Machines - SVM (Cortes and Vapnik, 1995), Artificial Neural Networks - ANN (Haykin, 1998), and Random Forests - RF (Breiman, 2001). The algorithms were implemented in Python using scikit-learn classification tools (Pedregosa et al., 2011). These methods were selected for their use in similar applications to predict wildlife hotspots (Bui et al., 2019; Peng et al., 2014). To identify the best algorithm to predict WVC hotspots, a Friedman's non-parametric statistical test and an LSD multiple comparison test was made by computing the AUC-ROC, confusion matrices, and Kappa statistics yielded by each algorithm (Riffenburgh, 2006).

The parameters of each classifier were optimized using grid search for KNN, SVM, and ANN. The number of neighbours of KNN was established for values between 1 and 300. For SVM, the grid search determined the amplitude and standard deviation parameters, with values between 0.0625 and 8 and between 0.0005 and 100, respectively. For ANN, we searched for the best number of neurons per layer and activation function. In the RF algorithm, a genetic algorithm - GA was used for optimization (Olson et al., 2016).

2.6. Transfer learning

Although there is a methodology popularly known as transfer learning, which consists of transferring a previously trained neural network architecture with images of a different element to the one to be classified or detected, this is not the method used in this research (Morid et al., 2021). Here, transfer learning is understood as the use of a trained classification model readjusted with new data. The refinement is performed with few new samples. Then, the updated classification model is applied to a validation set.

To validate the selected algorithm, a K fold approach was used to partition the cross-validation segments into training and transfer blocks with a ratio of 3:1 respectively, which at the end of each evaluation will change until each segment has been at least one time part of the test set. To evaluate the impact of different percentages of data added to the training dataset, different percentages of aggregated data were tested to readjust the model, determining the ideal percentage needed to make predictions in areas with minor to none WVC data.

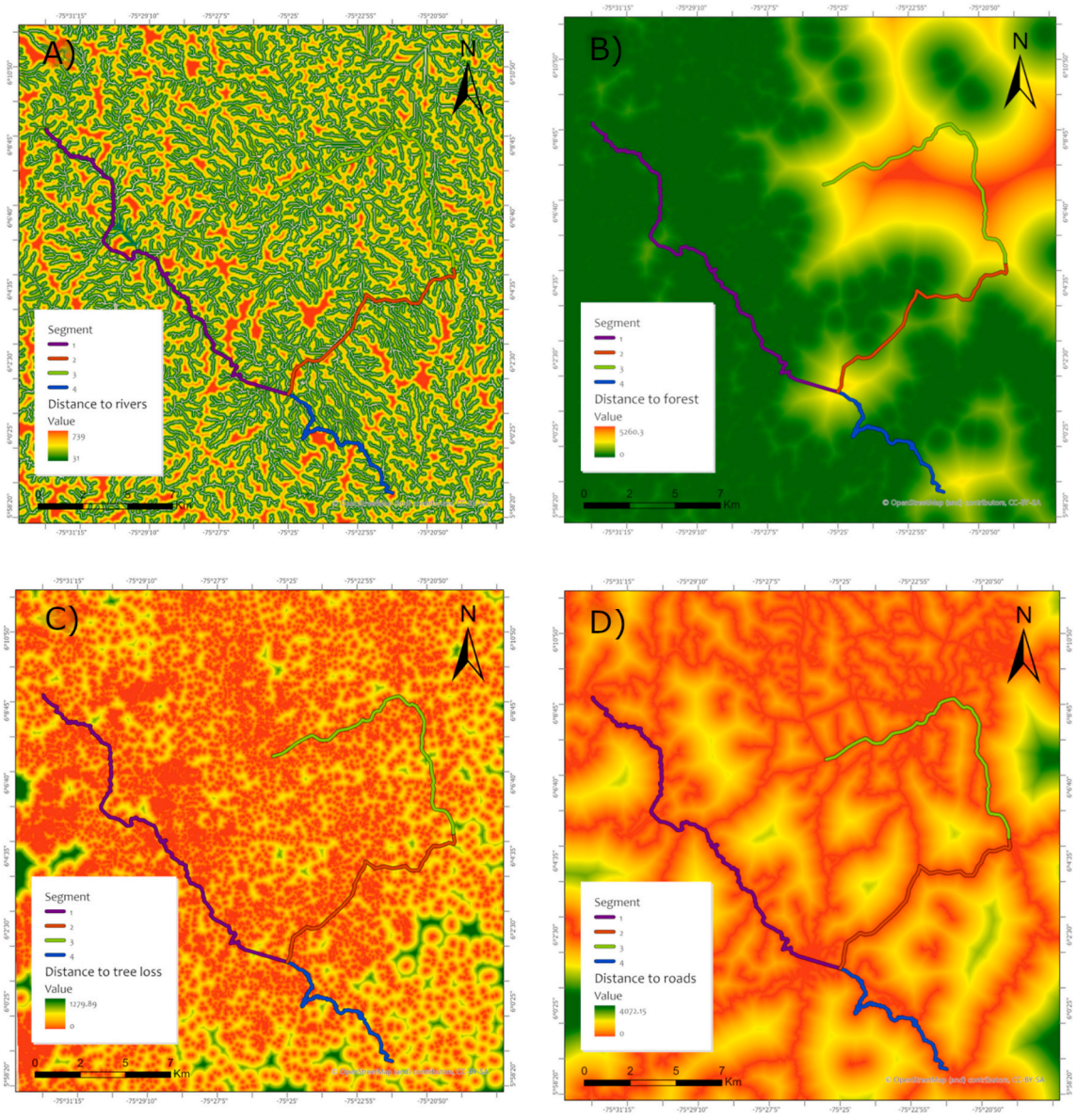


Fig. 4. Spatial information maps for Study Area. A) Distance to rivers, B) Distance to the forest, C) Distance to tree loss D) Distance to roads.

3. Results

3.1. Hotspot prediction

Fig. 5 shows K Ripley statistics for each segment of the study area. We identified significant clustering of points by comparing the observed pattern L(r) with the upper (ULC) and lower (LCL) confidence limits of a random distribution on the evaluated segment. Significant clusters were identified between 0.1 km(r) and 16 km(r) in segment 1, between 0.1 km (r) and 3.5 km(r) and between 6.8 km(r) and 10.4 km(r) in segment 2, between 0. 1 km(r) and 6.5 km(r) in segment 3, and between 0.1 km(r) and 3.7 km(r), between 5.1 km(r) and 5.3 km(r), and, between 5.5 km(r) and 6.5 km(r) in segment 4.

Cluster groups were created to allow autocorrelation tests considering the point aggregation intensities, each cluster was made using a search radius of 300 m as shown as a significant band in all evaluated segments by the K Ripley analysis. Spatial autocorrelation analysis of the cluster group distribution were performed, identifying the distance bands of 1.3 km and 269 m with a significant clustering pattern, with a positive spatial autocorrelation ($I = 0.18$, $I = 0.064$) for the segments 1 and 2, respectively. Likewise, it was evident that the 1.3 km distance band had significant clusters with a positive spatial autocorrelation ($I = 0.47$, $I = 0.21$) for segments 3 and 4, respectively.

Finally, Fig. 6 shows the intensity of the hotspots identified by the 2D hotspot analysis by a gradient color. In segment 1, significant clusters were identified between kilometers 7.5 and 12.65, 13.5, 14, and between kilometers 17.5 and 18.9. In the segment 2, significant grouping patterns were identified between kilometers 0.5 to 0.6, 2.1 to 3.6, and kilometers 5.6, 7.6, 8, and 11. For segment 3, kilometers 2 and 7 were identified as significant clusters. Finally, in segment 4, kilometers 0 and 0.5, and kilometers 1 to 6 were identified as significant clusters.

Because of the nature of the data, an imbalance was detected between the labels no-hotspot and hotspot with a 4:1 ratio, respectively, requiring the application of techniques that allow not only a synthetic balance of the feature matrix but also an adequate measurement of the performance of the machine learning algorithms (Hoens and Chawla, 2013; Japkowicz, 2013). The imbalance is caused by WVC data's nature, where WVC reports are expected to be grouped in the routes of connectivity severed by the construction of the road as suggested by (Mader, 1984). Due to the spatial dependence between the collected data, a

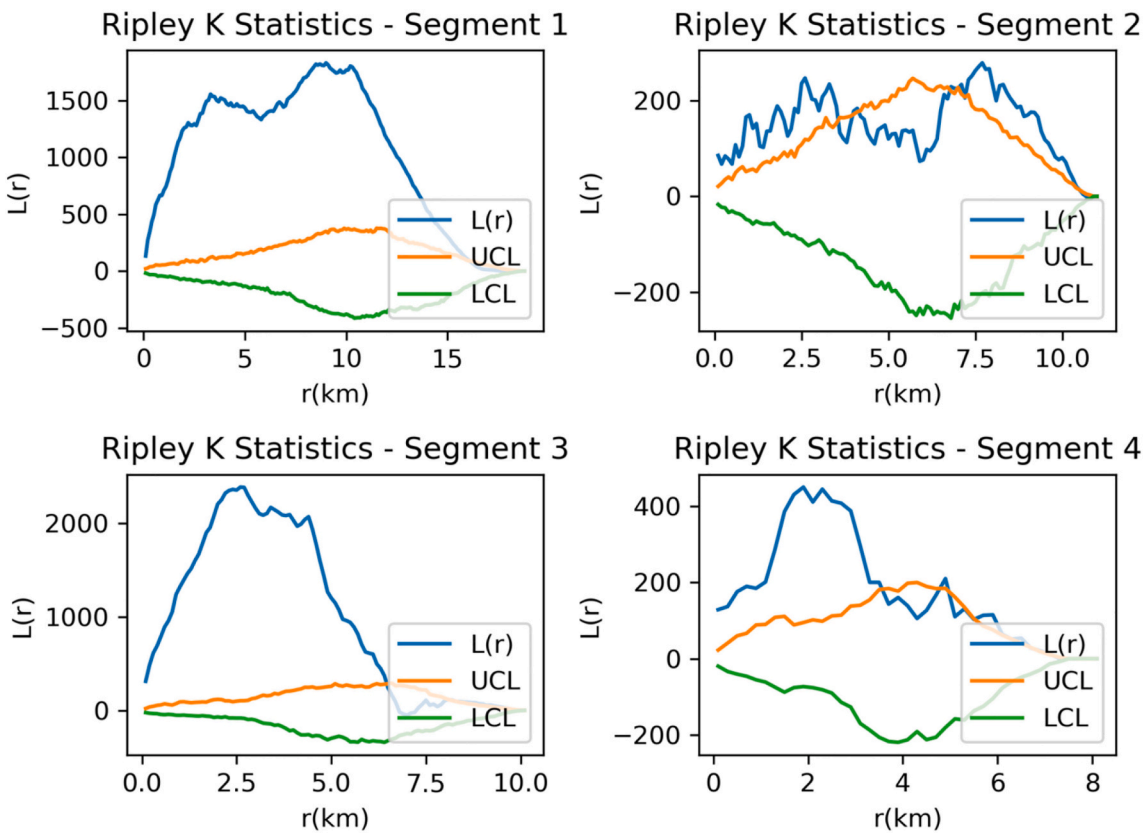


Fig. 5. K Ripley statistics for the study area, significant clusters are identified when the observed pattern L(r) surpass the upper confidence limit (UCL) of a random distribution for the same road segment.

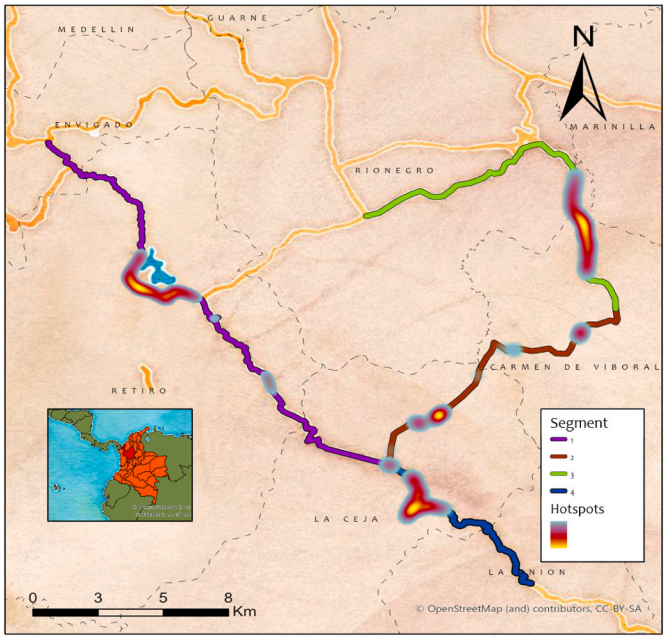


Fig. 6. Hotspots identified in the study area.

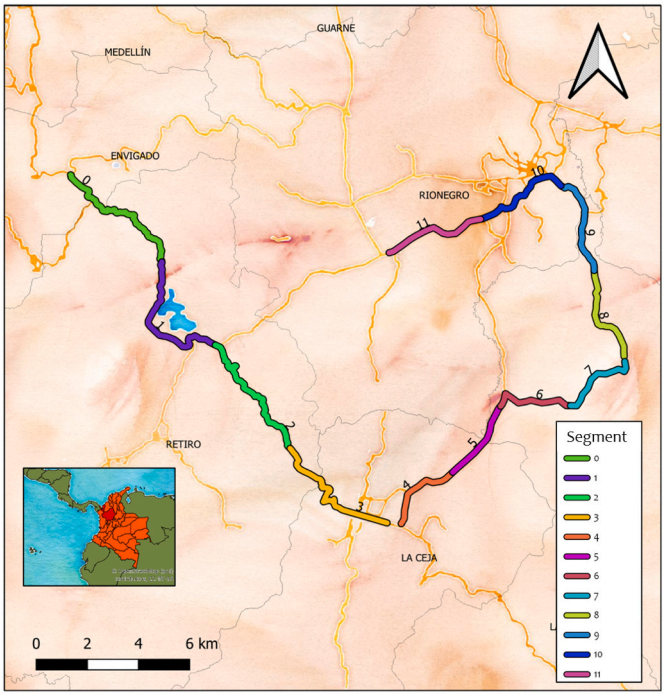


Fig. 7. Division of the training segments (1,2&3) in 12 subsegments.

validation algorithm was designed with 12 sub-segments, each one with 250 points distributed sequentially in the training area. Fig. 7 shows the sub-segment distribution in the study area.

The feature selection was carried out using the univariate selection method of the best K characteristics using Chi-square as the information criterion (Bahassine et al., 2020), Mutual Information (MI) (Qian et al., 2020), and the F value of ANOVA (F-ANOVA) (Güneş et al., 2010). The area under the ROC curve of a non-optimized Random Forest (RF) classifier was evaluated with a different number of features to select the number that would yield the best classification result. Fig. 8 shows the

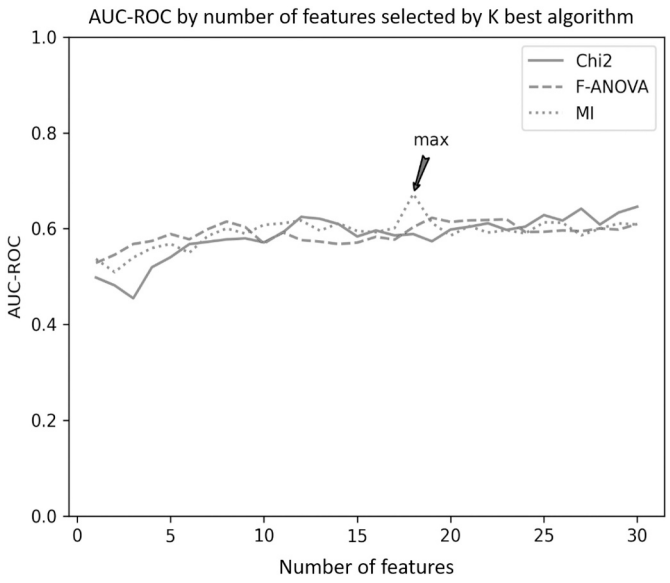


Fig. 8. Grid search of the ideal number of features selected by each of the univariate selectors.

results of each iteration of the algorithm with each of the univariate selection method (Chi2, F-ANOVA, and MI). We can note that the RF algorithm trained with a 17 subset of features identified by the MI selection was the optimal configuration, yielding the highest value (AUC-ROC = 0.67) among the tested features subsets.

Table 3 summarize the selected subset of features identified by the MI method, the buffer radius in which the feature was sampled, the amount of information provided by each feature in bits, and its contribution to the AUC-ROC. We can see that the same variable, e.g. Landsat Band 11 (TIRS 2), provides a different amount of information when the mean value of the raster is sampled at different buffer distances, additionally can be seen that the feature Distance to corridor sampled at 300 m yielded the most significant contribution to the RF algorithm.

With the identified features by the MI, custom-made machine learning algorithms: RF, SVM, KNN, and Neural Networks were tested using scikit-learn classification tools (Pedregosa et al., 2011). Table 4 shows the results of each of the optimized algorithms, the balance techniques employed, the F1-score, Accuracy, Kappa, and AUC-ROC performance metrics. In Table 4, we can observe the best results in terms of AUC and Kappa for each algorithm with its respective balance method: ANN BORDERLINE SMOTE (0.63 ± 0.11, 0.1586), RF ADASYN (0.78 ± 0.12, 0.3429), SVM KMEANS SMOTE (0.59 ± 0.17, 0.1372) and KNN BORDERLINE SMOTE (0.59 ± 0.08, 0.165).

Friedman non-parametric statistical test and the LSD multiple comparison test algorithm were performed using MATLAB software for all the classification algorithms. Table 5 shows the results obtained for the multiple comparison test between RF ADASYN and the other algorithms, showing significant statistical differences (90% C.I.) regarding the compared classifiers except with itself and KNN SMOTE, KNN KMEANS, and ANN BORDERLINE. Thus, the RF algorithm, which also has the highest performance of all the compared algorithms, was chosen together with the ADASYN balance method to be used to predict the WVC phenomenon with transfer learning.

Table 3
Optimal subset of features identified by the MI univariate feature selection method.

Feature	Mean sampling buffer	Amount of information (bits)	Feature contribution
Landsat Band 11 (TIRS 2)	300 m	0.228374	0.073298
Altitude	300 m	0.210617	0.066922
Distance to forest	300 m	0.197259	0.067143
Landsat Band 10 (TIRS 1)	300 m	0.195181	0.045025
Distance to corridor	300 m	0.174479	0.095792
Resistance	150 m	0.172832	0.030807
Resistance	300 m	0.169023	0.043452
Altitude	150 m	0.167544	0.074664
Distance to forest	150 m	0.167544	0.055238
Distance to corridor	150 m	0.159924	0.074783
Distance to corridor	90 m	0.152511	0.070427
Distance to forest	90 m	0.152351	0.059134
Landsat Band 9 (Cirrus)	300 m	0.150367	0.053969
Landsat Band 11 (TIRS 2)	150 m	0.146185	0.044474
Landsat NBRI index	300 m	0.144953	0.042359
Altitude	90 m	0.128334	0.039412
Movement cost	300 m	0.122733	0.063102

Table 4
Performance comparison of all the machine learning algorithms with balance methods.

Classification method	Resampling method	F1-Score	Precission	Kappa	AUC-ROC
ANN	SVM SMOTE	0.61394	0.7068	0.0368	0.54 ± 0.17
	SMOTE	0.6661	0.7136	0.0508	0.57 ± 0.16
	KMEANS SMOTE	0.6917	0.7392	0.1174	0.60 ± 0.16
	BORDERLINE SMOTE	**0.7006**	**0.7574**	**0.1586**	**0.63 ± 0.11**
	ADASYN	0.6755	0.7428	0.1235	0.63 ± 0.15
RF	SVM SMOTE	0.7332	0.8426	0.2355	0.71 ± 0.16
	SMOTE	0.7322	0.8442	0.2572	0.75 ± 0.12
	KMEANS SMOTE	0.6852	0.8378	0.2055	0.70 ± 0.11
	BORDERLINE SMOTE	0.7626	0.8628	0.3151	0.72 ± 0.14
	ADASYN	**0.7736**	**0.8624**	**0.3429**	**0.78 ± 0.12**
SVM	SVM SMOTE	0.7169	0.8448	0.0891	0.48 ± 0.18
	SMOTE	0.7169	0.8495	0.104	0.54 ± 0.14
	KMEANS SMOTE	**0.7556**	**0.8343**	**0.1372**	**0.59 ± 0.17**
	BORDERLINE SMOTE	0.7277	0.8032	0.1166	0.46 ± 0.17
	ADASYN	0.764	0.7844	0.1252	0.47 ± 0.17
KNN	SVM SMOTE	0.701	0.777	0.152	0.57 ± 0.1
	SMOTE	0.6997	0.7696	0.1679	0.59 ± 0.1
	KMEANS SMOTE	0.7096	0.8251	0.1709	0.61 ± 0.1
	BORDERLINE SMOTE	**0.688**	**0.7762**	**0.165**	**0.59 ± 0.08**
	ADASYN	0.6948	0.7763	0.1674	0.56 ± 0.09

Best result for each classification method and significative p-values ($p<0.1$) are indicated in bold

Table 5
Multiple comparison test between the RF ADASYN algorithm and the other tested algorithms.

Selected algorithm	Compared algorithm	*p-Value*
RF ADASYN	SVM_SVMSMOTE	**0.0027**
	ANN_SVMSMOTE	**0.0065**
	RF_SVMSMOTE	0.8576
	KNN_SVMSMOTE	**0.0363**
	SVM_SMOTE	**0.0092**
	ANN_SMOTE	**0.0120**
	RF_SMOTE	0.8109
	KNN_SMOTE	**0.1000**
	SVM_KMEANS	**0.0130**
	ANN_KMEANS	**0.0727**
	RF_KMEANS	0.3696
	KNN_KMEANS	0.1063
	SVM_BORDERLINE	**0.0142**
	ANN_BORDERLINE	0.1882
	RF_BORDERLINE	0.7649
	KNN_BORDERLINE	**0.0595**
	SVM ADASYN	**0.0313**
	ANN ADASYN	**0.0595**
	KNN ADASYN	**0.0828**

Best result for each classification method and significative p-values ($p<0.1$) are indicated in bold

3.2. Transfer learning

To validate the methodology for the prediction of WVC hotspots, a transfer learning methodology was implemented using the RF ADASYN algorithm selected in the previous section. The algorithm was fitted with the spatial information of the remaining segments, excluding the one in which the prediction was to be made. Additionally, to evaluate the incidence of retraining data on the model's performance, different percentages of the total length of the segment to be predicted were evaluated. Fig. 9 shows the ROC graphs with different percentages of the spatial information from the prediction segment used to retrain the algorithm, 0%, 1%, 5%, 10%, 15%, and 20%. We can note the presence of overfitting phenomenon starting at 10% of the total data used to retrain the algorithm, being clear due to the saturation of the model's performance.

Finally, Fig. 10 presents a prediction map for all the segments of the study area using the transfer learning approach. Each prediction was made using cross-validation methodology with 5% of retraining data. In it, we can see that the false positive predictions (in which the algorithm mistakenly classifies a segment as WVC hotspot) are grouped especially at the edges of the hotspot areas due to the similarities in this "buffer" zone, in which the changes of the spatial features are barely perceptible.

4. Analysis of results and discussion

Regarding the proposed features, the methodology used in this investigation was outlined by (Amiri et al., 2019; Ghorbani et al., 2019; Jaafari et al., 2019; Kantola et al., 2019; Thach et al., 2018; Wang et al., 2019) among others, who have used spatial variables as input for the prediction of different spatial phenomena. Besides, an attempt was made to extend the feature database by using the Landsat satellite's spectral bands to obtain as much information as possible about the vegetation coverage, which is similar to the use made by (Ascensão et al., 2019). The characteristics proposed in this project are based on the previous works presented by (Ascensão et al., 2019; Fabrizio et al., 2019; Gonçalves et al., 2018; Ha and Shilling, 2018; Kantola et al., 2019) among others.

In the hotspot identification, segment 1 was identified as the road segment with the most reports: 281. It is a consequence of the proximity of segment 1 to the Rio Nare Reserve, the San Miguel and Cerros de San Nicolás protected areas, the primary ecological nodes in the area in terms of area and integrity. On the other hand, segment 4 has a greater distance to protected areas and forests, reducing the roads' impact on these protected ecosystems (Forman et al., 2003; van der Ree et al., 2015a, 2015b).

K Ripley analysis and spatial autocorrelation analysis allow us to determine that the phenomenon of WVC in the study area was not randomly mediated, consistent with the work made by (Clevenger et al., 2003). Additionally, it was possible to determine that the hotspots of fauna roadkill are related to their neighbours in terms of point accumulation, being more similar to their neighbours than to distant points of aggregation, as described by (Getis and Ord, 2010; Griffith, 2015; Ord and Getis, 2010).

For the feature selection, the Mutual Information (MI) metric has proven to be a method with promising results allowing the selection of the features with the highest amount of relevant information for the ranking algorithm's output (Zhou et al., 2020). In this study, the selected features are related to the land use around the events, the distance to forest cover, the distance to water sources, among other features as shown by (Ascensão et al., 2019; Fabrizio et al., 2019; Gonçalves et al., 2018; Ha and Shilling, 2018; Kantola et al., 2019), related to ecosystem quality (Hansen et al., 2013). Also, the selection of features at different scales was necessary, as shown by (Ha and Shilling, 2018), obtaining different contribution results to the prediction with the same feature measured at different scales.

The features selected: Distance to forest 300 m, Distance to biological corridor 300 m, Resistance 150 m, Resistance 300 m, distance to forest 150 m, distance to biological corridor 150 m, distance to biological corridor 90 m, distance to forest 90 m, Movement cost 300 m, are directly related to the ecological connectivity, which is consistent with (Mader, 1984; Mansergh and Scotts, 1989). The features: Landsat band 11,300 m, Landsat band 10,300 m, Landsat band 9300 m, Landsat band 11,150 m, and NBRI_300 m capture the temperature of the ground, which decreases in the presence of vegetation, allowing the algorithm the quantitative observation of the thermal regulation provided by the forests and plant cover (Shen et al., 2019), providing more attractive areas for animal movement as described by (Maffei and Andrew, 2003).

This work used a group-sensitive cross-validation, thus ensuring that the results do not contain biases related to the high spatial correlation between neighboring segments, allowing the model to be evaluated in completely unknown scenarios. According to the current knowledge of the authors, group-sensitive cross-validation has not been used to validate predictive models of WVC in the past. However, group validation is a technique usually used to validate spatial predictive models, as reported by (Kajornrit and Wong, 2013).

As it is known, the success in the implementation of a machine learning algorithm depends on the choice of appropriate parameters when training a model (Smets et al., 2007). Although each method's optimization method was selected according to the complexity of the optimization problem, the methods described are of the meta-heuristic type (Kurniasih et al., 2019). In the nearest neighbour algorithm (KNN), there is only one parameter to optimize, so it was decided to use the GridSearch method. However, there are faster methods such as those proposed by (Fukunaga and Narendra, 1975), (Moreno-Seco et al., 2002), (Baek and Sung, 2000). Due to the low amount of data in the training phase and the fact that the algorithm execution time was less than 5 min, other algorithms that could reduce the training time were not considered necessary.

Likewise, in the case of the Vector Support Machine (SVM) algorithm, although there is a great variety of possible kernels to be used in the training stage, it was decided to use the radial-based kernel due to the positive results it has shown in different applications such as those presented by (Duan and Liu, 2012), (Ye and Li, 2012), among others. Besides, limiting the optimization problem to a single Kernel allowed to focus on finding the C and Gamma values employing specific search methods with cross-validation methods, guaranteeing a selection of parameters adjusted to different data sets, avoiding overtraining the model (Wang et al., 2012).

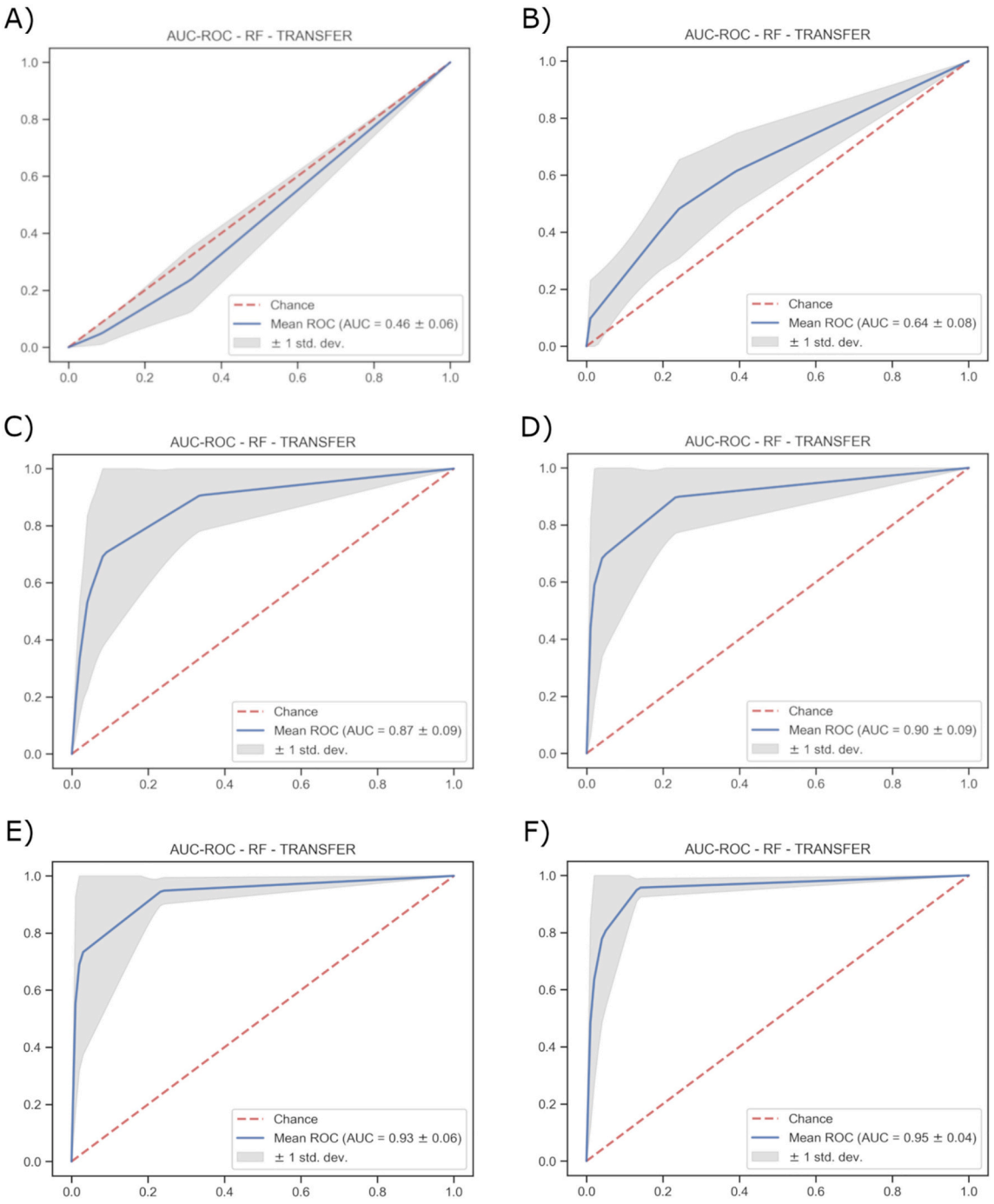


Fig. 9. AUC-ROC of the ADASYN RF classifier when subjected to different % of data aggregated to the training set. A) 0%, B) 1%, C) 5%, D) 10%, E), 15%, F) 20%.

About artificial neural networks (ANN), the optimization of the network structure is a problem which, according to the authors' current understanding, does not yet have an optimum solution. Therefore, in this project, it was decided to carry out a GridSearch optimization of the network structure parameters utilizing multiple optimization algorithms' repetitions. Finally, the random forest algorithm was optimized using genetic algorithms. This algorithm allows the optimization of multiple parameters efficiently and effectively (Kramer, 2017).

Regarding the comparison of the classifiers, the RF ADASYN algorithm provided the best result (0.78 ± 12), improving the result obtained by (Ascensão et al., 2019), being the precedent that most closely approximates the methodology applied in this research. However, comparing results of different methodologies applied in different study areas are particularly difficult.

Finally, positive results were observed for all the methods implemented. This type of methodologies has not been sufficiently explored for the WVC phenomenon, requiring a higher amount of research to improve the results present here, especially the neural networks, since they have proven to be especially effective in predicting different spatial phenomena with excellent results (Bui et al., 2016; Bui et al., 2017; Bui et al., 2019; Ghorbani et al., 2019; Jaafari et al., 2019).

To validate the methodology for predicting hotspots for fauna from multispectral imagery and geographic information systems, we proposed to adopt the Transfer Learning technique (He and Ma, 2013) used in neural networks to implement it in this particular classification problem. Different percentages of the track segments of the validation

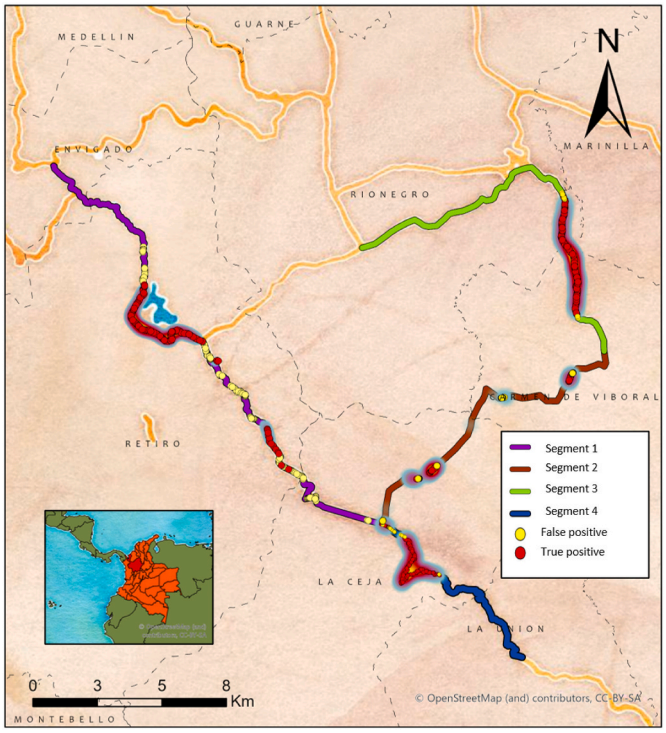


Fig. 10. Predicted Hotspots using 5% of data aggregated to the training set.

section were used to train the model. In Fig. 9 can be seen that the results of the classifier are improved with a higher proportion of data from the validation section used during the training, obtaining values of AUC ROC from 0.64 ± 0.08, corresponding to the retrained classifier with 1% of the validation data, to 0.95 ± 0.04, corresponding to the retrained classifier with 20% of the validation data.

We consider that the results using 5% retrain data show the most realistic approach to deploy this methodology in a real scenario, in which only 5% of the road would be needed to do a car survey to predict the 95% left with a good performance of the algorithm. However, this should be tested in other areas using a selected class or species data with a specific temporal window, tests we could not do due to the lack of enough information about a single species or class to identify the significant WVC hotspots.

5. Conclusions

This research has developed a novel methodology for predicting the most significant accumulation of WVC on roads in Eastern Antioquia, based on artificial intelligence algorithms, geographic information systems, and multispectral image processing. This includes a characterization of the WVC hotspots based on multispectral images, a feature selection using univariate selection methods, a selection of the machine learning model with the best fit, and a transfer learning experiment in areas unknown to the model. Tests shown that the RF ADASYN algorithm was the best performing algorithm (AUC-ROC = 0.78 ± 0.12) and (AUC-ROC = 0.87 ± 0.09) when retrained with 5% of the total length of the road to be predicted.

The methodology used in this work has the potential to reduce the response times of the academy, control bodies, and road operators to the phenomenon of WVC, allowing the estimation of areas with potential hotspots to be validation by diagnostic studies that will identify and propose mitigation measures. Also, the constructed methodology seeks to fill a partial void of information about the application of classification models for the prediction of WVC hotspots. It also seeks to set a standard on how should this type of algorithms be validated, considering the spatial bias introduced by the spatial autocorrelation.

This work shows a theoretical approach to the prediction of WVC hotspots in areas with few data collected. However, it is necessary to carry out field validation of the results obtained through this methodology, which is considered the future work of this project. Although this paper shows a better performance than the others present in the current state-of-the-art, it is essential to highlight that comparing methods in different areas and different data is demanding due to the different ecological settings and the differences between them of data recollection.

Regarding using a composite Landsat 8 image, we clarify that the authors acknowledge that temporal vegetation changes and other detailed spatial changes are not visible with this approach. However, since the aim to include the Landsat image in this work is to give information about the presence/absence of vegetation and other objects near the roads, the methodology should not be considerably affected by this, specially considering that Colombia has two seasons, wet and dry. Future works are encouraged to develop models considering seasonality, and spatial changes, allowing to narrow down the WVC Hotspot for some particular species or season.

Lastly, we note that this work has been done using all vertebrate animal class data. It was made to ensure a sufficient amount of data for the training stage. However, it would be ideal to recollect enough information on a single species to make a more precise WVC Hotspot identification because of the different behavior of even similar species.

Declaration of Competing Interest

The authors do not report any conflict of interests.

Acknowledgements

We would like to highlight the important contribution of Víctor Colino-Rabanal, Juan D. Martinez-Vargas and Tobías Leyva-Pinto who had the kindness of reviewing and make valuable contributions to this work. Also, we would like to thank the App Recosfa users who make this work possible by reporting WVC. This work was developed within the project: "Development of a methodology to predict the sites with the highest accumulation of wildlife roadkill (Hot Spots) on roads managed by the La Pintada Concession based on artificial intelligence algorithms and geographic information systems", carried out at the ITM and identified with the code P20249.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecoinf.2021.101291.

References

Amiri, M., Pourghasemi, H.R., Ghanbarian, G.A., Afzali, S.F., 2019. Assessment of the importance of gully erosion effective factors using Boruta algorithm and its spatial modeling and mapping using three machine learning algorithms. Geoderma 340, 55–69. https://doi.org/10.1016/j.geoderma.2018.12.042.

Ascensão, F., Yogui, D., Alves, M., Medici, E.P., Desbiez, A., 2019. Predicting spatiotemporal patterns of road mortality for medium-large mammals. J. Environ. Manag. 248, 109320. https://doi.org/10.1016/j.jenvman.2019.109320.

Baek, S.J., Sung, K.-M., 2000. Fast K-nearest-neighbour search algorithm for nonparametric classification. Electron. Lett. 36, 1821. https://doi.org/10.1049/el:20001249.

Bahassine, S., Madani, A., Al-Sarem, M., Kissi, M., 2020. Feature selection using an improved Chi-square for Arabic text classification. J. King Saud Univ. 32, 225–231. https://doi.org/10.1016/j.jksuci.2018.05.010.

Baynes, J., 2004. Assessing forest canopy density in a highly variable landscape using Landsat data and FCD mapper software. Aust. For. 67, 247–253. https://doi.org/10.1080/00049158.2004.10674942.

Beier, P., Spencer, W., Baldwin, R.F., McRae, B.H., 2011. Toward best practices for developing regional connectivity maps. Conserv. Biol. 25, 879–892.

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

Bui, D.T., Pradhan, B., Nampak, H., Bui, Q.-T., Tran, Q.-A., Nguyen, Q.-P., 2016. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibilitgy modeling in a high-frequency tropical cyclone area using GIS. J. Hydrol. 540, 317–330. https://doi.org/10.1016/j.jhydrol.2016.06.027.

Bui, D.T., Bui, Q.-T., Nguyen, Q.-P., Pradhan, B., Nampak, H., Trinh, P.T., 2017. A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area. Agric. For. Meteorol. 233, 32–44. https://doi.org/10.1016/j.agrformet.2016.11.002.

Bui, D.T., Ngo, P.-T.T., Pham, T.D., Jaafari, A., Minh, N.Q., Hoa, P.V., Samui, P., 2019. A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping. CATENA 179, 184–196. https://doi.org/10.1016/j.catena.2019.04.009.

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

Clevenger, A.P., Waltho, N., 2005. Performance indices to identify attributes of highway crossing structures facilitating movement of large mammals. Biol. Conserv. 121, 453–464. https://doi.org/10.1016/j.biocon.2004.04.025.

Clevenger, A.P., Chruszcz, B., Gunson, K.E., 2003. Spatial patterns and factors influencing small vertebrate fauna road-kill aggregations. Biol. Conserv. 109, 15–26. https://doi.org/10.1016/s0006-3207(02)00127-1.

Coelho, A.V., Coelho, I.P., Kindel, A., Teixeira, F.Z., 2014. Road mortality software Siriema: road mortality software. In: User's Manual V. 2.0. Universidade Federal do Rio Grande do Sul.

Coffin, A.W., 2007. From roadkill to road ecology: a review of the ecological effects of roads. J. Transp. Geogr. 15, 396–406. https://doi.org/10.1016/j.jtrangeo.2006.11.006.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20, 273–297. https://doi.org/10.1007/bf00994018.

Cramer, P., Olsson, M., Gadd, M.E., van der Ree, R., Sielecki, L.E., 2015. Transportation and Large Herbivores, in: Handbook of Road Ecology. John Wiley & Sons Ltd, pp. 344–352. https://doi.org/10.1002/9781118568170.ch42.

Crawford, B.A., Maerz, J.C., Nibbelink, N.P., Buhlmann, K.A., Norton, T.M., Albeke, S.E., 2014. Hot spots and hot moments of diamondback terrapin road-crossing activity. J. Appl. Ecol. 51, 367–375. https://doi.org/10.1111/1365-2664.12195.

Cureton, J.C., Deaton, R., 2012. Hot moments and hot spots: identifying factors explaining temporal and spatial variation in turtle road mortality. J. Wildl. Manag. 76, 1047–1052. https://doi.org/10.1002/jwmg.320.

DANE, 2017. Geoportal DANE - Descarga del Marco Geoestadistico Nacional (MGN).

Danks, Z.D., Porter, W.F., 2010. Temporal spatial, and landscape habitat characteristics of moose-vehicle collisions in Western Maine. J. Wildl. Manag. 74, 1229–1241. https://doi.org/10.2193/2008-358.

Davenport, J., Switalski, T.A., 2006. Environmental impacts of transport related to tourism and leisure activities. In: The Ecology of Transportation: Managing Mobility for the Environment. Springer Netherlands. https://doi.org/10.1007/1-4020-4504-2_14.

Dickson, B.G., Albano, C.M., Anantharaman, R., Beier, P., Fargione, J., Graves, T.A., Gray, M.E., Hall, K.R., Lawler, J.J., Leonard, P.B., Littlefield, C.E., McClure, M.L., Novembre, J., Schloss, C.A., Schumaker, N.H., Shah, V.B., Theobald, D.M., 2018. Circuit-theory applications to connectivity science and conservation. Conserv. Biol. 33, 239–249. https://doi.org/10.1111/cobi.13230.

Douzas, G., Bacao, F., Last, F., 2018. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf. Sci. 465, 1–20. https://doi.org/10.1016/j.ins.2018.06.056.

Duan, H., Liu, X., 2012. Lower C limits in support vector machines with radial basis function kernels. In: 2012 International Symposium on Information Technologies in Medicine and Education. IEEE. https://doi.org/10.1109/itime.2012.6291416.

Durduran, S.S., 2010. A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform. Expert Syst. Appl. 37, 7729–7736. https://doi.org/10.1016/j.eswa.2010.04.068.

Eshel, G., 2011. Autocorrelation. In: Spatiotemporal Data Analysis. Princeton University Press. https://doi.org/10.23943/princeton/9780691128917.003.0008.

Fabrizio, M., Febbraro, M.D., Loy, A., 2019. Where will it cross next? Optimal management of road collision risk for otters in Italy. J. Environ. Manag. 251, 109609. https://doi.org/10.1016/j.jenvman.2019.109609.

Forman, R.T.T., Sperling, D., Bissonette, J.A., Clevenger, A.P., Cutshall, C.D., Dale, V.H., Fahrig, L., Heanue, K., France, R.L., Goldman, C.R., et al., 2003. Road Ecology: Science and Solutions. Island Press.

Fukunaga, K., Narendra, P.M., 1975. A branch and bound algorithm for computing k-nearest neighbors. IEEE Trans. Comput. C-24, 750–753. https://doi.org/10.1109/t-c.1975.224297.

García-Morera, Y., Giraldo-Iral, L., 2018. Recopilación de Información de FAUNA en la Jurisdicción de CORNARE, hasta el año 2015 v1.1. https://doi.org/10.15472/lyfjyt.

Getis, A., Ord, J.K., 2010. The analysis of spatial association by use of distance statistics. Geogr. Anal. 24, 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

Ghorbani, M.A., Deo, R.C., Kashani, M.H., Shahabi, M., Ghorbani, S., 2019. Artificial intelligence-based fast and efficient hybrid approach for spatial modelling of soil electrical conductivity. Soil Tillage Res. 186, 152–164. https://doi.org/10.1016/j.still.2018.09.012.

Girardet, X., Conruyt-Rogeon, G., Foltête, J.-C., 2015. Does regional landscape connectivity influence the location of roe deer roadkill hotspots? Eur. J. Wildl. Res. 61, 731–742. https://doi.org/10.1007/s10344-015-0950-4.

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ. 58, 289–298. https://doi.org/10.1016/s0034-4257(96)00072-7.

Gonçalves, L.O., Alvares, D.J., Teixeira, F.Z., Schuck, G., Coelho, I.P., Esperandio, I.B., Anza, J., Beduschi, J., Bastazini, V.A.G., Kindel, A., 2018. Reptile road-kills in Southern Brazil: composition hot moments and hotspots. Sci. Total Environ. 615, 1438–1445. https://doi.org/10.1016/j.scitotenv.2017.09.053.

Griffith, D., 2015. Spatial statistics and geostatistics: basic concepts. In: Encyclopedia of GIS. Springer International Publishing, pp. 1–16. https://doi.org/10.1007/978-3-319-23519-6_1650–1.

Güneş, S., Polat, K., Yosunkaya, Şebnem, 2010. Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome. Expert Syst. Appl. 37, 998–1004. https://doi.org/10.1016/j.eswa.2009.05.075.

Gunson, K.E., Mountrakis, G., Quackenbush, L.J., 2011. Spatial wildlife-vehicle collision models: a review of current work and its application to transportation mitigation projects. J. Environ. Manag. 92, 1074–1082. https://doi.org/10.1016/j.jenvman.2010.11.027.

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2003. KNN model-based approach in classification. In: On The Move to Meaningful Internet Systems 2003: CoopIS DOA and ODBASE. Springer Berlin Heidelberg, pp. 986–996. https://doi.org/10.1007/978-3-540-39964-3_62.

Ha, H., Shilling, F., 2018. Modelling potential wildlife-vehicle collisions (WVC) locations using environmental factors and human population density: a case-study from 3 state highways in Central California. Ecol. Inform. 43, 212–221. https://doi.org/10.1016/j.ecoinf.2017.10.005.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853. https://doi.org/10.1126/science.1244693.

Harirforoush, H., Bellalite, L., 2019. A new integrated GIS-based analysis to detect hotspots: a case study of the city of Sherbrooke. Accid. Anal. Prev. 130, 62–74. https://doi.org/10.1016/j.aap.2016.08.015.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation, 2nd. ed. Prentice Hall PTR, USA.

He, H., Ma, Y., 2013. Imbalanced Learning. Wiley. https://doi.org/10.1002/9781118646106.

He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE. https://doi.org/10.1109/ijcnn.2008.4633969.

Hoens, T.R., Chawla, N.V., 2013. Imbalanced datasets: from sampling to classifiers. In: Imbalanced Learning. John Wiley & Sons Inc, pp. 43–59. https://doi.org/10.1002/9781118646106.ch3.

IDEAM, 2016. Mapa Bosque No Bosque.

IDEAM, 2017. MAPA ECOSISTEMAS CONTINENTALES, COSTEROS Y MARINOS - IDEAM.

Jaafari, A., Zenner, E.K., Panahi, M., Shahabi, H., 2019. Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability. Agric. For. Meteorol. 266-267, 198–207. https://doi.org/10.1016/j.agrformet.2018.12.015.

Jackson, R.D., 1983. Spectral indices in N-space. Remote Sens. Environ. 13, 409–421. https://doi.org/10.1016/0034-4257(83)90010-X.

Jaeger, J.A.G., 2015. Improving environmental impact assessment and road planning at the landscape scale. In: Handbook of Road Ecology. John Wiley & Sons Ltd, pp. 32–42. https://doi.org/10.1002/9781118568170.ch5.

Japkowicz, N., 2013. Assessment metrics for imbalanced learning. In: Imbalanced Learning. John Wiley & Sons Inc, pp. 187–206. https://doi.org/10.1002/9781118646106.ch8.

Kajornrit, J., Wong, K.W., 2013. Cluster validation methods for localization of spatial rainfall data in the northeast region of Thailand. In: 2013 International Conference on Machine Learning and Cybernetics. IEEE. https://doi.org/10.1109/icmlc.2013.6890861.

Kantola, T., Tracy, J.L., Baum, K.A., Quinn, M.A., Coulson, R.N., 2019. Spatial risk assessment of eastern monarch butterfly road mortality during autumn migration within the southern corridor. Biol. Conserv. 231, 150–160. https://doi.org/10.1016/j.biocon.2019.01.008.

Kramer, O., 2017. Genetic Algorithm Essentials. Springer International Publishing. https://doi.org/10.1007/978-3-319-52156-5.

Kurniasih, J., Utami, E., Raharjo, S., 2019. Heuristics and metaheuristics approach for query optimization using genetics and Memetics algorithm. In: 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS). IEEE. https://doi.org/10.1109/icoris.2019.8874909.

Laurance, W.F., Clements, G.R., Sloan, S., O'Connell, C.S., Mueller, N.D., Goosem, M., Venter, O., Edwards, D.P., Phalan, B., Balmford, A., Ree, R.V.D., Arrea, I.B., 2014. A global strategy for road building. Nature 513, 229–232. https://doi.org/10.1038/nature13717.

Mader, H.-J., 1984. Animal habitat isolation by roads and agricultural fields. Biol. Conserv. 29, 81–96. https://doi.org/10.1016/0006-3207(84)90015-6.

Madsen, A.B., Strandgaard, H., Prang, A., 2002. Factors causing traffic killings of roe deer *Capreolus capreolus* in Denmark. Wildl. Biol. 8, 55–61. https://doi.org/10.2981/wlb.2002.008.

Maffei, L., Andrew, B. Taber, 2003. Área de acción, actividad y uso de hábitat del zorro patas negras, *Cerdocyon thous*, en un bosque seco. Mastozoología Neotropical.

Mansergh, I.M., Scotts, D.J., 1989. Habitat continuity and social organization of the mountain pygmy-possum restored by tunnel. J. Wildl. Manag. 53, 701. https://doi.org/10.2307/3809200.

McRae, B.H., Dickson, B.G., Keitt, T.H., Shah, V.B., 2008. Using circuit theory to model connectivity in ecology evolution and conservation. Ecology 89, 2712–2724. https://doi.org/10.1890/07-1861.1.

Moreno-Seco, F., Micó, L., Oncina, J., 2002. Extending LAESA fast nearest neighbour algorithm to find the k nearest neighbours. In: Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 718–724. https://doi.org/10.1007/3-540-70659-3_75.

Morid, M.A., Borjali, A., Fiol, G.D., 2021. A scoping review of transfer learning research on medical image analysis using ImageNet. Comput. Biol. Med. 128, 104115. https://doi.org/10.1016/j.compbiomed.2020.104115.

Müller, A.C., Guido, S., 2016. Introduction to Machine Learning With Python: A Guide for Data Scientists, 1st ed. In: A Guide for Data Scientists. O'Reilly Media.

Nguyen, H.M., Cooper, E.W., Kamei, K., 2011. Borderline over-sampling for imbalanced data classification. Int. J. Knowledge Eng. Soft Data Parad. 3, 4. https://doi.org/10.1504/ijkesdp.2011.039875.

Nguyen, H.K.D., Buettel, J.C., Fielding, M.W., Brook, B.W., 2021. Predicting Spatial and Seasonal Patterns of Wildlife-vehicle Collisions in High-risk Areas. https://doi.org/10.1101/2021.01.17.427044.

Olson, R.S., Bartley, N., Urbanowicz, R.J., Moore, J.H., 2016. Evaluation of a tree-based pipeline optimization tool for automating data science. In: Proceedings of the 2016 on Genetic and Evolutionary Computation Conference - 16. ACM Press. https://doi.org/10.1145/2908812.2908918.

Ord, J.K., Getis, A., 2010. Local spatial autocorrelation statistics: distributional issues and an application. Geogr. Anal. 27, 286–306. https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

Pagany, R., 2020. Wildlife-vehicle collisions - influencing factors data collection and research methods. Biol. Conserv. 251, 108758. https://doi.org/10.1016/j.biocon.2020.108758.

Pagany, R., Valdes, J., Dorner, W., 2020. Risk prediction of wildlife-vehicle collisions comparing machine learning methods and data use. In: 2020 10th International Conference on Advanced Computer Information Technologies. IEEE. https://doi.org/10.1109/acit49673.2020.9208946.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, Édouard, 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Peng, L., Niu, R., Huang, B., Wu, X., Zhao, Y., Ye, R., 2014. Landslide susceptibility mapping based on rough set theory and support vector machines: a case of the Three Gorges area China. Geomorphology 204, 287–301. https://doi.org/10.1016/j.geomorph.2013.08.013.

Qian, W., Huang, J., Wang, Y., Shu, W., 2020. Mutual information-based label distribution feature selection for multi-label learning. Knowl.-Based Syst. 195, 105684. https://doi.org/10.1016/j.knosys.2020.105684.

RECOSFA, 2019. Red Colombiana de Seguimiento de Fauna Atropellada.

Riffenburgh, R.H., 2006. Chapter summaries. In: Riffenburgh, R.H. (Ed.), Statistics in Medicine, Second edition. Academic Press, Burlington, pp. 533–580. https://doi.org/10.1016/B978-012088770-5/50067-8.

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

Serrón, A., Coitiño, H., Segura, Á., 2020. Atropellos de mamíferos en la Región Este de Uruguay y su relación con los atributos del paisaje. INNOTEC, pp. 139–157.

Shen, W., Li, M., Huang, C., He, T., Tao, X., Wei, A., 2019. Local land surface temperature change induced by afforestation based on satellite observations in Guangdong plantation forests in China. Agric. For. Meteorol. 276-277, 107641. https://doi.org/10.1016/j.agrformet.2019.107641.

Smets, K., Verdonk, B., Jordaan, E.M., 2007. Evaluation of performance measures for SVR hyperparameter selection. In: 2007 International Joint Conference on Neural Networks. IEEE. https://doi.org/10.1109/ijcnn.2007.4371031.

Smith, D.J., van der Ree, R., 2015. Field methods to evaluate the impacts of roads on wildlife. In: Handbook of Road Ecology. John Wiley & Sons Ltd, pp. 82–95. https://doi.org/10.1002/9781118568170.ch11.

Thach, N.N., Ngo, D.B.-T., Xuan-Canh, P., Hong-Thi, N., Thi, B.H., Nhat-Duc, H., Dieu, T. B., 2018. Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: a comparative study. Ecol. Inform. 46, 74–85. https://doi.org/10.1016/j.ecoinf.2018.05.009.

USGS, 2000. USGS EROS Archive - Digital Elevation - Shuttle Radar Topography Mission (SRTM) 1 Arc-second Global.

USGS, 2013. Landsat 8.

van der Ree, R., Smith, D.J., Grilo, C., 2015a. The ecological effects of linear infrastructure and traffic. In: Handbook of Road Ecology. John Wiley & Sons Ltd, pp. 1–9. https://doi.org/10.1002/9781118568170.ch1.

van der Ree, R., Smith, D.J., Grilo, C., 2015b. Handbook of Road Ecology. John Wiley & Sons Ltd. https://doi.org/10.1002/9781118568170.

Wang, T., Ye, X., Wang, L., Li, H., 2012. Grid search optimized SVM method for dish-like underwater robot attitude prediction. In: 2012 Fifth International Joint Conference on Computational Sciences and Optimization. IEEE. https://doi.org/10.1109/cso.2012.189.

Wang, H., Qin, F., Zhang, X., 2019. A spatial exploring model for urban land ecological security based on a modified artificial bee colony algorithm. Ecol. Inform. 50, 51–61. https://doi.org/10.1016/j.ecoinf.2018.12.009.

Ye, Z., Li, H., 2012. Based on radial basis kernel function of support vector machines for speaker recognition. In: 2012 5th International Congress on Image and Signal Processing. IEEE. https://doi.org/10.1109/cisp.2012.6469807.

Zhou, H., Wang, X., Zhang, Y., 2020. Feature selection based on weighted conditional mutual information. Appl. Comput. Inform. https://doi.org/10.1016/j.aci.2019.12.003 ahead-of-print.