Juan Carlos González Vélez

Predicción de puntos calientes de atropellamiento de fauna con base en algoritmos de inteligencia artificial, sistemas de información geográfica y procesamiento de imágenes multiespectrales

Predicción de puntos calientes de atropellamiento de fauna con base en algoritmos de inteligencia artificial, sistemas de información geográfica y procesamiento de imágenes multiespectrales

Juan Carlos González Vélez

Instituto Tecnológico Metropolitano
Facultad de Ingeniería
Medellín, Colombia
2020

Predicción de puntos calientes de atropellamiento de fauna con base en algoritmos de inteligencia artificial, sistemas de información geográfica y procesamiento de imágenes multiespectrales

Realizó:
Juan Carlos González Vélez

Tesis presentada como requisito para optar al título de:
Magíster en Automatización y Control Industrial

Directores:
Juan Pablo Murillo Escobar, MSc.
Juan Carlos Jaramillo Fayad, PhD.
Maria Constanza Torres Madroñero, PhD.

Grupo de investigación:
Química Básica, Aplicada y Ambiente

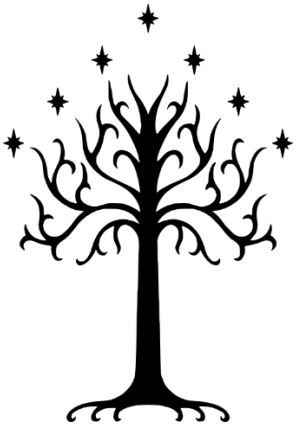
Instituto Tecnológico Metropolitano
Facultad de Ingeniería
Medellín, Colombia
2020

Agradecimientos

Margarita, por ser el faro que mantuvo este barco a flote.
Ignacio, por apoyarme siempre.
Andrés, que con su ejemplo siempre mostró que el trabajo duro da grandes frutos.
Jenny, por estar siempre a mi lado.
Carolina por recordarme lo que es la curiosidad e inocencia pura.

A Juan Carlos Jaramillo, por darme todas las oportunidades que me han definido.
A Juan Pablo Murillo, por apoyarme desde el primer momento que escuchó este proyecto.
A Maria Constanza, por brindarme la oportunidad de aprender a su lado.
A José Luis, por siempre brindar consejos sabios y certeros.

Al dios de la vida, que mantuvo a flote a mi familia, aun cuando las adversidades parecían ser más de lo que cualquiera de nosotros podía soportar.



What we do in life, echoes in eternity

Resumen

El atropellamiento de fauna es un fenómeno que surge a partir de la fragmentación de los ecosistemas por las vías, limitando la movilidad de los individuos y poniendo en riesgo la estabilidad de las poblaciones por el incremento de la mortalidad. Este fenómeno es estudiado por diferentes disciplinas que se integran en el campo de conocimiento denominado: ecología de carreteras. Colombia no es ajeno a la problemática del atropellamiento de fauna silvestre, evidenciado en diferentes publicaciones científicas que reportan este fenómeno en las vías del país. Aunque el auge de la inteligencia artificial ha tenido grandes avances en la predicción de fenómenos espaciales como incendios forestales, crecimientos súbitos de corrientes de agua, entre otros. Sin embargo, estos algoritmos aún no han sido suficientemente explorados por la ecología de carreteras. Por esta razón, esta investigación tuvo como objetivo desarrollar una metodología para predecir los sitios de mayor acumulación de atropellamiento de fauna en vías del Oriente Antioqueño con base en algoritmos de inteligencia artificial, sistemas de información geográfica y procesamiento de imágenes multiespectrales. Durante el desarrollo de esta investigación se identificó que las características más relacionadas con el atropellamiento de fauna en la zona de estudio son: distancia a Bosque, distancia a corredor biológico, resistencia del terreno al movimiento, costo de movimiento, las bandas 9, 10 y 11 del satélite Landsat 8 y el índice de quema normalizado (NBRI). A partir de este conjunto de características, se compararon diferentes algoritmos de aprendizaje de máquina (k-vecinos más cercanos, máquinas de soporte vectorial, bosques aleatorios y redes neuronales) balanceados por medio de las técnicas SMOTE y ADASYN. Los resultados obtenidos permiten identificar que el algoritmo de Bosques aleatorios (RF) con la técnica de balanceo ADASYN es el método de clasificación con mejor desempeño al ser sometido a validación cruzada por grupos (AUC-ROC 0.78 ± 0.12), superando los resultados alcanzados por investigaciones previas. Por último, se validó la metodología a través de un ejercicio de transferencia de aprendizaje, entrenando el algoritmo RF-ADASYN con 3 zonas del oriente antioqueño y validando sobre un tramo diferente (AUC-ROC = 0.87±0.09) reentrenando el modelo inicial con el 5% de la base de datos de validación.

Palabras Clave: Análisis espacial, Aprendizaje de Máquina, Reconocimiento de Patrones, Atropellamiento de fauna, Imágenes Multiespectrales

Abstract

The roadkill of fauna 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. This is studied by different disciplines integrated into the field of knowledge called Road Ecology. Colombia is not unaware of the problem of the running over of wild fauna, evidenced in different scientific publications that describe the phenomenon of the running over of wild fauna 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 had the objective of developing a methodology to predict the sites of more significant accumulation of fauna run-over in roads of the Antioquia East based on artificial intelligence algorithms, geographic information systems, and multispectral image processing. During the development of this research, it was identified that the characteristics most related to the roadkill of fauna 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: band 9, band 10, band 11 and the normalized burning index. From this set of characteristics, different machine learning algorithms were compared (nearest k-neighbors, vector support machines, random forests, and neural networks). SMOTE and ADASYN balancing techniques were applied. The results obtained allowed us to identify that the randomized forest (RF) algorithm with ADASYN balancing is the method with the best performance when subjected to cross-validation (AUC-ROC 0.78 ± 0.12), surpassing the results achieved by previous research. Finally, the methodology was validated through a transference exercise, training the RF-ADASYN algorithm with three zones of the eastern Antioquia region and validating on a different section (AUC-ROC = 0.87±0.09), retraining the initial model with 5% of data from the validation database.

Keywords: Spatial Analysis, Machine Learning, Pattern Recognition, Fauna Roadkill, Multispectral Imaging

Tabla de contenido

Pág.

Índice de figuras

Pág.

Índice de tablas

Pág.

1. Introducción

1.1. Justificación

El atropellamiento de fauna es una problemática que tiene un comportamiento directamente proporcional al crecimiento de las vías y se ha atribuido principalmente a la fragmentación de los ecosistemas causada por las infraestructuras lineales que limitan la movilidad de los individuos y ponen en riesgo la estabilidad de las poblaciones por el incremento de la mortalidad [1]. Esto tiene consecuencias graves para los ecosistemas debido a la pérdida de servicios ecosistémicos como: control de plagas, control de poblaciones, dispersión de semillas, entre otros, que benefician al ecosistema y en general a todos los que lo habitamos [2]. Sin embargo, este fenómeno también afecta al ser humano: las colisiones con la fauna provocan lesiones, pérdida de vidas y costos asociados a la reparación de los vehículos. Se estima que en Estados Unidos ocurren aproximadamente 2 millones de colisiones entre vehículos y mamíferos grandes cada año, resultando en al menos 29,000 personas lesionadas, 200 o más muertes humanas [3], así como pérdidas económicas estimadas en $4,000 millones de dólares cada año [4, 5]. Adicionalmente, se estima que 365 millones de vertebrados mueren cada año en las carreteras de este país [6]. Por otro lado, en Europa se estima que ocurren al menos 500,000 colisiones de vehículos con ungulados cada año [5]. Debido a la magnitud de esta problemática, algunos autores han considerado el atropellamiento de fauna como uno de los factores que más contribuyen a la pérdida de biodiversidad [7]. Por lo cual, es necesario generar medidas que permitan mitigar los efectos negativos de las infraestructuras viales sobre los ecosistemas y la fauna silvestre [8].

Uno de los principales propósitos de los investigadores en conservación, de los ecólogos de las carreteras, así como de los gerentes de las infraestructuras viales, es la identificación de los lugares con mayor riesgo de mortalidad de fauna silvestre. Estos lugares son identificados por medio de estudios diagnósticos, los cuales requieren de una gran cantidad de recursos, talento humano especializado y amplios periodos de tiempo para obtener información sistemática y significativa. A partir de la información recolectada, se generan análisis espaciales que permiten proponer medidas de mitigación y prevención al atropellamiento de fauna [9-16].

Colombia no es ajeno a la problemática del atropellamiento de fauna silvestre, y aunque hay pocos estudios, estos evidencian el fenómeno del atropellamiento de fauna

silvestre en las vías del país por medio de la cuantificación de la mortalidad en las carreteras, la identificación de las especies más afectadas y los lugares más frecuentes donde se presentan atropellamientos [17-28]. Por esta razón, el Instituto Tecnológico Metropolitano (ITM) de Medellín, ha buscado generar un espacio de investigación de alto impacto con un enfoque multidisciplinario, en el cual, la biología, la ecología, la geo estadística, las ciencias computacionales, entre otras disciplinas, pueden unir esfuerzos para prevenir, mitigar y compensar el atropellamiento de fauna silvestre por medio de técnicas y algoritmos que permitan ser más efectivos en la atención de esta problemática.

Quizás una de las áreas más prometedoras a ser aplicadas a este fenómeno es la Inteligencia Artificial; un conjunto de técnicas y algoritmos que buscan brindarle la capacidad a los sistemas computacionales de aprender. Del mismo modo que los humanos aprendemos de nuestras experiencias y el mundo, los sistemas "aprenden" de los datos que se les brinda, permitiendo generar predicciones, análisis o incluso la toma de decisiones [29]. En los últimos años se ha dado un aumento exponencial en la generación de investigaciones que hacen uso de algoritmos de aprendizaje como insumo para la predicción de diversos fenómenos espaciales [29-42], por esta razón, este proyecto pretende hacer uso de estos algoritmos como una herramienta adicional para la prevención y mitigación del atropellamiento de fauna silvestre. Sin embargo, al realizar la revisión bibliográfica, se evidencia una baja cantidad de investigaciones que hagan uso de algoritmos de aprendizaje de máquina para predecir los segmentos de mayor acumulación de atropellamientos.

Por lo tanto, esta tesis busca aportar a la construcción de medidas de mitigación y prevención cada vez más eficientes, disminuyendo los costos y tiempos para la identificación de los segmentos más probables de las carreteras a ser cruzados por la fauna, permitiendo continuar el aumento de la infraestructura vial de manera responsable y respetuosa con los ecosistemas.

1.2 Pregunta de investigación

¿Se podrán predecir puntos calientes de atropellamiento de fauna en vías del oriente antioqueño, que no cuenten con datos suficientes para realizar análisis espaciales convencionales, a través de métodos basados en inteligencia artificial y sistemas de información geográfica?

1.3 Objetivos de investigación

1.3.1 Objetivo general

Desarrollar una metodología para predecir los sitios de mayor acumulación de atropellamiento de fauna, en vías del Oriente Antioqueño con base en algoritmos de inteligencia artificial, sistemas de información geográfica y procesamiento de imágenes multiespectrales.

1.3.2 Objetivos específicos

Proponer un conjunto de características geográficas y ambientales de las vías del oriente antioqueño asociadas a zonas con acumulación de atropellamientos de fauna con base en el procesamiento de imágenes satelitales multiespectrales y mapas oficiales.

Determinar el algoritmo de inteligencia artificial que mejor se ajuste a la predicción de puntos calientes de atropellamiento de fauna en carreteras del Oriente antioqueño teniendo en cuenta el desbalance de clases.

Validar la metodología propuesta para la predicción de puntos calientes de atropellamiento de fauna por medio de transferencia de aprendizaje en las vías del oriente antioqueño

1.4 Organización del documento

Este documento consta de siete capítulos, los cuales están organizados de la siguiente manera: una introducción que tiene como propósito la descripción del objeto de estudio, los objetivos e hipótesis de este trabajo. El capítulo de estado del arte y marco conceptual recopila el devenir científico de la temática del atropellamiento de fauna silvestre, cómo ha evolucionado y cuáles son las metodologías más usadas para su estudio y predicción. Así mismo, se realiza un compendio de investigaciones que hacen uso de la inteligencia artificial para la predicción de fenómenos espaciales, además de un marco conceptual en donde se introducen los conceptos teóricos que fueron utilizados en este estudio. El capítulo 3 está dedicado a la descripción del marco metodológico que fue utilizado durante esta investigación para las fases de caracterización del atropellamiento de fauna, el desarrollo de modelos de predicción de

atropellamiento y la validación por medio de transferencia de aprendizaje. El capítulo 4 presenta los resultados logrados, las métricas de desempeño de los modelos desarrollados de sus predicciones y el resultado de la transferencia de aprendizaje. Finalmente, un capítulo de conclusiones y recomendaciones, un capítulo de bibliografía y un capítulo de anexos

2. Estado del arte y marco conceptual

2.1. Estado del arte

Teniendo en cuenta que este trabajo tiene como problema gestor el atropellamiento de fauna, se estableció el estado actual de esta problemática, desde la década de 1950 hasta el momento. Así mismo se realizó un estudio de los métodos que se han usado para analizar la información obtenida en esta temática y como los modelos de aprendizaje de máquina pueden ser útiles para la predicción de este fenómeno.

Con el objetivo de garantizar la repetitividad de esta revisión bibliográfica se presentan las ecuaciones de búsqueda usadas: para el tema de ecología de carreteras *((animal AND vehicle AND collision) OR (roadkill) OR (wildlife AND vehicle AND collision) OR (animal AND roadkill)) AND (road)*; para el tema de análisis espacial visual y sistemas de información geográfica; *((animal AND vehicle AND collision) OR (roadkill) OR (wildlife AND vehicle AND collision) OR (animal AND roadkill)) AND (Satellite imagery OR multi spectral)* Por último, para el tema de modelos y algorítmicos de inteligencia artificial: *(predict OR Model OR probability) AND (roadkill OR Animal Vehicle Collision OR Wildlife Vehicle Collision)*. En vista de los pocos resultados que se encuentran de esta búsqueda, específicamente que usen algoritmos de aprendizaje de máquina, se complementa con la siguiente ecuación: *((ArcGIS) OR (GIS)) AND ((artificial intelligence) OR (machine learning))*. Estas ecuaciones fueron utilizadas el 12 de febrero del 2020 en bases de datos indexadas como ScienceDirect, Scopus, Web of Science y Google Scholar. Esta revisión fue realizada en estas bases de datos debido a la presencia de artículos que describen la problemática en múltiples lugares del mundo, incluyendo artículos de Brasil y otros países latinoamericanos con similitudes a Colombia. Es importante mencionar que esta búsqueda fue complementada con artículos similares que fueron citados por artículos ubicados en esta revisión.

2.1.1 Ecología de carreteras

La Ecología de Carreteras es una disciplina que estudia los efectos negativos de las vías y las autopistas sobre los ecosistemas [5]. Entre los efectos negativos más investigados se encuentran: la contaminación química, sonora y lumínica provocada por los vehículos que transitan por las carreteras; la pérdida de hábitat adyacente a la vía; degradación del hábitat y del suelo; el efecto barrera y el efecto borde que impiden el movimiento natural de las especies; el atropellamiento de fauna; dispersión de especies invasoras o especies generalistas a lo largo del trazado de la vía, entre otros [3]. De los efectos mencionados, el atropellamiento de fauna ha sido ampliamente estudiado, encontrándose en publicaciones científicas desde los años 1950, especialmente en Estados Unidos y Europa, donde se reportan constantes colisiones con Ungulados y otros animales de tamaño considerable [43-45]. A pesar de esto, no fue sino hasta 1974 que se comienza a hablar de los efectos de las carreteras de manera formal, entre ellos el atropellamiento de fauna silvestre, como un fenómeno asociado a las carreteras [46, 47].

Las preocupaciones por este fenómeno surgen inicialmente como respuesta a los problemas de seguridad vial. Desde la década de 1950 se han implementado diferentes señalizaciones viales para advertir a los conductores acerca de la posible presencia de animales sobre la vía, incluso, modificándolas con luces intermitentes con el objetivo de aumentar su visibilidad y efectividad [48, 49]. A pesar de estas medidas de prevención y mitigación, diversas investigaciones reportan grandes cantidades de atropellamientos, por lo cual, el fenómeno comienza a ser considerado como uno de los factores que más contribuyen a la pérdida de poblaciones de algunas especies, así como un factor que genera cambios en la genética de las especies [50-53].

A partir de la década de 1980 se comienzan a generar las primeras estrategias a gran escala para la mitigación de este fenómeno, así como la continuación de investigaciones que reportan el fenómeno del atropellamiento y los efectos negativos de otras infraestructuras lineales como las líneas eléctricas y de transporte de petróleo [54-58]. Así mismo, la conectividad ecológica y el efecto barrera producido por la transformación del paisaje comenzó a ser identificada como un factor determinante del atropellamiento de fauna [59, 60].

A partir de la década de 1990 surgen estudios acerca de la efectividad de las medidas de mitigación, especialmente enfocándose en el paso de animales a través de

alcantarillas de cajón, pasos inferiores, superiores y otras estructuras [61-65]. Así mismo, comienzan a surgir metodologías de análisis de información de atropellamiento de fauna, los cuales serán descritos con mayor detalle en la sección 2.1.2. De igual manera, comienza a desarrollarse un mayor interés en explicar el fenómeno, avanzando en la identificación de las variables que lo propician y los costos económicos asociados a esta problemática [62, 66-74]

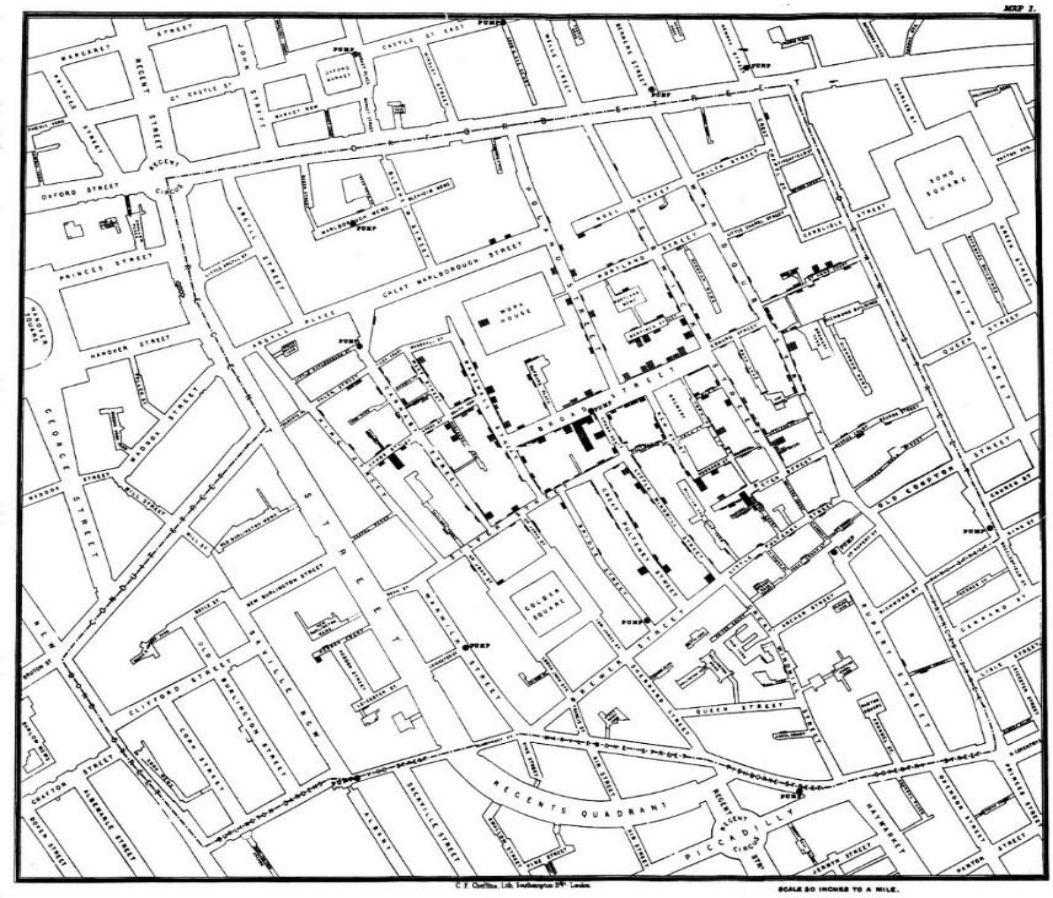
La década del 2000-2010 se caracterizó por la publicación de documentos que sintetizan el devenir científico previo, generando textos de referencia y actualización que se convirtieron en los pilares de esta disciplina científica [2, 12, 75]. Del mismo modo, durante esta época surgen propuestas de modelos estadísticos que buscan predecir el atropellamiento de fauna, los cuales serán descritos posteriormente. Por último, la década del 2010-2020 se ha caracterizado por una continuación de los esfuerzos a nivel mundial por determinar los factores que propician el atropellamiento de fauna, los modelos y técnicas que mejor se ajustan al fenómeno, así como el incremento de la cantidad de estudios en zonas en vía de desarrollo como Asia y América Latina. Cabe aclarar que durante esta década surgen documentos clave en el desarrollo de la ecología de carreteras, así como documentos emitidos por entidades estatales en diferentes países con recomendaciones para la construcción adecuada de las carreteras [5, 76-81].

2.1.2. Metodologías de análisis del atropellamiento de fauna

Análisis espacial visual

La humanidad es cada vez más dependiente de los datos [82], la mayoría de los cuales son recolectados por dispositivos móviles que nos acompañan todo el tiempo. Estos datos son acompañados por coordenadas geográficas, fecha y hora, ubicando la información recolectada en el espacio y en el tiempo, los cuales al ser representados en un mapa pueden brindar aún más información. Uno de los ejemplos más conocidos de los usos de la información espacial es el mapa de John Snow (Figura 1) durante la epidemia de cólera de 1854 en Londres, Inglaterra. John Snow, un médico tratante de la enfermedad, ubicó las víctimas de esta enfermedad en un mapa y según el patrón observado, pudo identificar que la bomba de agua de Broad Street estaba contaminada con la bacteria causante de la enfermedad [83].

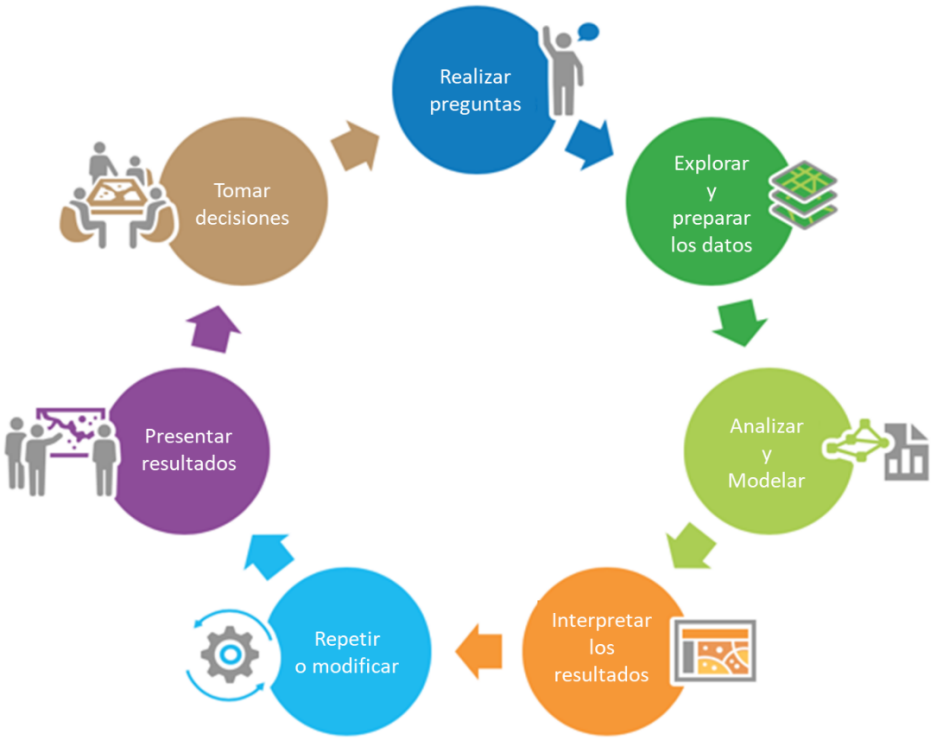
Figura 1. Mapa de la Cólera, John Snow, Soho ciudad de Londres 1854. Fuente: [83]



El análisis espacial se entiende como cualquier método o técnica analítica empleada en el estudio de diversos fenómenos naturales, económicos, socioculturales, entre otros, que utilizan datos representados en una o en varias escalas espaciales [84]. Estos análisis nos ayudan a entender donde se encuentra ubicado un fenómeno, como se relaciona con su entorno y que acciones se deben tomar en caso de que dicho fenómeno presente consecuencias negativas para la sociedad. Estos análisis pueden ser visuales o pueden ser cuantificables al aplicar técnicas estadísticas. En la Figura 2 se observa el flujo de trabajo más común para realizar cualquier tipo de análisis espacial.

Con el paso de los años y con la llegada de la carrera espacial, las diferentes disciplinas que realizaban análisis espacial contaron con una herramienta impensable hasta ese momento: las imágenes satelitales. En 1957 el programa Sputnik logró poner el primer satélite artificial en órbita [85], a partir de este hito, diversos gobiernos lanzaron múltiples programas de satélites de reconocimiento terrestre para uso militar, así como satélites equipados con cámaras de alta resolución sensibles a diferentes

Figura 2. Flujo de trabajo del análisis espacial. Fuente: curso "Getting Started with Spatial Analysis", Esri (2019).



bandas espectrales para aplicaciones civiles y científicas. Entre los satélites más destacados se encuentran: Landsat (1-8), Sentinel (1-6), VRSS-1, Rapid Eye, entre otros. A partir de las imágenes captadas por estos satélites y con la ayuda de Sistemas de información Geográficas (SIG) ha sido posible obtener y analizar gran cantidad de información espacial, permitiendo a científicos de diferentes áreas conocer más acerca de nuestro planeta y realizar análisis espaciales visuales a partir de estas [86].

Los satélites captan imágenes multiespectrales con múltiples dimensiones de información. En el caso del satélite Sentinel 2, cada imagen consta de 12 bandas del espectro electromagnético (entre los 443 nm y 2190 nm); por su parte, las imágenes captadas por el satélite Landsat están formadas por 11 bandas del espectro electromagnético (entre los 433 nm y 1250 nm), haciendo posible la identificación de diferentes elementos en la superficie a partir de su reflectancia. De estas imágenes se pueden calcular índices que permiten la identificación de elementos de interés, por ejemplo, la vegetación puede ser observada con mayor detalle a partir del índice de vegetación de diferencia normalizada (NDVI), el cual se calcula a partir de las bandas entre 845 nm y 885 nm, y las bandas entre 630 nm y 650 nm [87].

Debido a los altos índices de atropellamiento de fauna silvestre, es usual que los investigadores en Ecología de Carreteras realicen análisis espaciales de la información recolectada por monitoreos diagnósticos en terreno, complementados con información espacial proveniente de imágenes satelitales. Es importante mencionar que los monitoreos diagnósticos requieren metodologías estandarizadas para el levantamiento de información del fenómeno del atropellamiento de fauna, siendo considerados como una herramienta fundamental en la ecología de carreteras [5].

Sistemas de información geográfica (SIG) para la identificación de puntos calientes de atropellamiento

Los SIG son un conjunto de herramientas computacionales para el procesamiento, análisis y almacenamiento de información georreferenciada, la cual puede ser utilizada en múltiples áreas del conocimiento entre las que se destacan: la Geografía, la Cartografía, la Geología, entre otras disciplinas científicas, convirtiéndose en la principal herramienta para el análisis de fenómenos espaciales a partir de herramientas computacionales y geo estadísticas [31].

Una de las aproximaciones más comunes al fenómeno del atropellamiento de fauna es evaluarlo como un fenómeno espacial, dependiente de múltiples variables ecológicas y técnicas de la vía [10, 11, 14, 88-92], por lo cual, es común que las investigaciones hagan uso de los SIG y las imágenes multiespectrales para su análisis, esto debido a la disponibilidad de imágenes en cortos periodos de tiempo y una diversidad de herramientas integradas en la plataforma SIG para su análisis. A partir de las variables ecológicas y técnicas identificadas, es usual que los análisis busquen localizar las zonas con mayor probabilidad de cruce o colisión de los animales con vehículos que transitan por las vías [13, 93-96]. A continuación, se hará una descripción del estado del arte de los métodos de análisis espacial que han sido usados en el fenómeno del atropellamiento de fauna.

El atropellamiento de fauna es un fenómeno ecológico que tiene relación con el espacio y el tiempo, por lo cual, es necesario incorporar en los análisis del atropellamiento de fauna el componente espacio-temporal [97-103]. Tradicionalmente la identificación de los puntos calientes se ha efectuado por técnicas de acumulación, en las cuales, un Kernel, usualmente circular, genera un mapa de calor según la agrupación de puntos, asumiendo los segmentos de vía con mayor intensidad como

puntos calientes [9, 10, 104, 105]. A partir de esta técnica, han surgido variaciones como las descritas por Coelho, et al. [106] en su software Siriema. En este software se presenta el análisis de puntos calientes 2D Hotspot para datos de atropellamiento de fauna, el cual busca identificar las agrupaciones de puntos estadísticamente significativas a lo largo de la vía, realizando agrupaciones de ellas en un radio r y comparando sus resultados con una distribución normal [106].

Teniendo en cuenta que el atropellamiento de fauna es un fenómeno espacial, es importante determinar las escalas espaciales a las cuales se presenta mayor agrupación, por lo cual, técnicas geo estadísticas como el K de Ripley han sido aplicadas [95, 104, 106, 107]. Otras técnicas han buscado tener una mayor significancia desde el punto de vista estadístico haciendo uso de técnicas geo estadísticas, como por ejemplo el análisis de puntos calientes optimizado, el cual identifica agrupaciones significativas por medio del índice de autocorrelación espacial [108, 109].

En una gran cantidad de los análisis espaciales para la identificación de puntos calientes de atropellamiento de fauna, así como en los estudios que buscan explicar este fenómeno, es usual que se haga uso de las imágenes satelitales multiespectrales como insumo para producir mapas que contengan las variables espaciales de interés para cada estudio. Algunos de los mapas más usados que usan imágenes satelitales multiespectrales como insumo son: mapas de coberturas vegetales [110, 111], mapas de Índices de vegetación normalizada (NDVI) [112, 113], imágenes aéreas para ser analizadas visualmente [114], así como imágenes para clasificación supervisada según la variable que se desee mapear [113, 115].

2.1.3 Modelos predictivos para el atropellamiento de fauna

Los estudios del fenómeno del atropellamiento tuvieron su auge en la década de los años 70, donde se presentaban análisis estadísticos, generalmente descriptivos, cuyo objetivo principal fue el entendimiento de la problemática y las variables asociadas a este. Sin embargo, a partir de finales de 1990 comenzaron a publicarse modelos estadísticos y matemáticos que estiman donde se presentarían atropellamientos de fauna silvestre en las vías a partir de datos recolectados por censos diagnósticos u otras técnicas [115, 116].

Inicialmente estos modelos buscaron identificar las horas del día y temporada climática en las cuales se presenta una mayor probabilidad de colisiones con animales en vías férreas y carreteras. Estos modelos de tipo regresivos se basaron en el comportamiento de ungulados y variables que influyen en sus patrones de movimiento como comportamientos migratorios, fases lunares, temporada climática y hora del día, entre otras variables [89, 116]. A pesar de tomar variables espaciales en los modelos, estos se enfocaron principalmente en los patrones temporales, dejando de lado algunas variables ecológicas asociadas a segmentos de mayor acumulación de atropellamientos.

Posteriormente, se desarrollaron modelos lineales generalizados, que buscan predecir los lugares de mayor riesgo de colisión con base a variables ecológicas. Estos modelos tienen en cuenta diversas variables como: variables topográficas: pendiente y elevación; variables de paisaje: tamaños de parches, tamaños de núcleos, tipos de coberturas vegetales, proporción de cobertura boscosa o urbana; índices espectrales como el NDVI; variables climáticas como el promedio de precipitaciones al mes, temperatura y humedad promedio, entre otras; variables asociadas a la vía como tipo de tráfico vehicular, densidad de tráfico, visibilidad del segmento, tipo de vía, velocidad máxima permitida, velocidad efectiva en el sector; entre otras variables, [117-120]. En los últimos años se han publicado modelos bayesianos que usan variables similares para ajustarse al fenómeno[111, 121].

Usualmente los modelos han sido validados principalmente por 2 técnicas, el criterio de información de Akaike (AIC) y el error cuadrático medio (RMSE), los cuales son utilizados para seleccionar el modelo con el menor error posible en la predicción del fenómeno del atropellamiento, esto comparando la predicción realizada por el modelo en una zona conocida [16, 115, 122, 123].

Algoritmos de aprendizaje aplicados a fenómenos espaciales

Los algoritmos de aprendizaje de máquina tienen como propósito emular el concepto de aprendizaje humano. Este campo de estudio es una intersección entre la estadística y las ciencias computacionales, también conocido como analítica predictiva, aprendizaje estadístico, entre otros [29]. Gracias a los avances logrados en los algoritmos de aprendizaje de máquina, en los últimos años se han realizado una gran cantidad de trabajos orientados al análisis espacial por medio de estos algoritmos [30,

32-34, 36-39, 41]. Dichos análisis usualmente tienen como objetivo predecir la ubicación y magnitud de diversos fenómenos espaciales.

Algunos de estos algoritmos nacen de modificaciones de técnicas estadísticas existentes, las cuales pueden ser utilizadas para predecir. Entre los más usados se destacan: el algoritmo de vecinos más cercanos (KNN) y sus derivados como el índice de vecinos más cercanos (NNI), las máquinas de soporte vectorial (SVM), las redes neuronales artificiales (RNA), el algoritmo de Bosques Aleatorios (RF), entre otros. En el caso del algoritmo KNN, se busca predecir a partir de los vecinos más cercanos la clase de un dato desconocido por medio de consenso de clase [124]. Este algoritmo se caracteriza por tener una baja dificultad de implementación, buen desempeño y bajo costo computacional.

Del mismo modo, existen algoritmos que se basan en técnicas de regresión estadística para realizar predicciones, un ejemplo común son las SVM, estas buscan separar un grupo de puntos a través de una función de ajuste conocida como hiperplano. Para hacerlo, se ubican las observaciones más alejadas del promedio y a partir de estas se genera una función de ajuste usando mínimos cuadrados, siendo la función resultante la frontera de decisión o hiperplano, la cual permite realizar clasificaciones de puntos desconocidos según las características que lo describan [125]. Una de las aplicaciones más conocidas de las SVM es la clasificación supervisada de imágenes satelitales, logrado a partir del análisis de los espectros de reflectancia de las imágenes [126].

En los últimos años se ha popularizado el uso de métodos de clasificación en cascada como las redes neuronales; estas pertenecen a un nuevo grupo de algoritmos de inteligencia artificial denominados algoritmos de aprendizaje profundo. Estos algoritmos de aprendizaje hacen uso de múltiples ecuaciones lineales llamadas neuronas, que al estar unidas forman una red interconectada de funciones ponderadas con una salida común. Usualmente, estas ecuaciones son ajustadas a los datos a través de métodos de optimización heurística como la retro propagación o algoritmos de enjambre de partículas, los cuales permiten ajustar los parámetros de cada neurona para minimizar el error de predicción [29].

La literatura permite identificar que la inteligencia artificial es una herramienta útil para el análisis espacial, siendo evidente al encontrar aplicaciones a problemas tan complejos como: predicción de accidentes de tránsito [34], incendios forestales [31, 33, 37], crecientes súbitas [32], erosión de los suelos [39], conductividad en suelos [35],

polución en suelos [40] y seguridad ecológica [42]. Estos fenómenos se caracterizan por tener diversas componentes estocásticas, lo cual hace más compleja su predicción. Entre los algoritmos más usados para realizar análisis espacial, se encuentran: Bosques aleatorios [30], SVM [39], Redes neuronales artificiales [32] y modificaciones de estas como las Redes neuronales difusas [31], las cuales han generado predicciones con una precisión mayor al 90%.

Para el caso del fenómeno del atropellamiento de fauna, no se encuentran artículos científicos que apliquen técnicas de clasificación como las descritas anteriormente. Sin embargo, existen precedentes que usan técnicas de aprendizaje como el MaxEntropy y la regresión logística. Estas técnicas generan modelos probabilísticos según un conjunto de datos de entrenamiento, sin embargo, debido a su aproximación probabilística, estos algoritmos poseen una incertidumbre en su predicción, aunque adecuada para estimar el riesgo de atropellamiento, es inadecuada cuando se desea determinar si un segmento de vía corresponde o no a un punto caliente [127]. Las investigaciones mencionadas anteriormente tienen desempeños reportados a través del cálculo del área bajo la curva del receptor característico (AUC-ROC) del orden del 70%, los cuales son resultados prometedores y sirven como precedentes para esta investigación [90, 128, 129].

2.2. Marco conceptual

2.2.1 Caracterización del fenómeno del atropellamiento de fauna

Recolección de datos de atropellamiento de fauna

La recolección de datos de atropellamiento es esencial para prevenir, mitigar y compensar los efectos ambientales negativos de las carreteras sobre la fauna. En todos los casos, el éxito de la recolección de datos y la calidad de estos aumenta cuando se sigue un proceso estructurado [5]. Usualmente, este proceso consta de diferentes fases en donde se formula un objetivo claro para el estudio, se recolectan datos existentes, se construye un diseño de estudio eficaz para identificar las especies objetivo, se seleccionan los mejores métodos para la recopilación de datos y se diseña una metodología clara para el análisis de los datos utilizando técnicas apropiadas, lo cual debe llevar a publicar los resultados con el objetivo de comparar sus resultados con otros estudios similares [5].

Aunque existen numerosas publicaciones que detallan la aplicación de estos métodos de recolección de información [130, 131], cada método tiene sus limitaciones y sesgos, aunque el uso de múltiples técnicas aumentará la precisión y la calidad de la información recogida. Los estudios de atropellamientos de fauna en carretera registran el número, la ubicación y las especies de fauna que mueren en la carretera debido a colisiones entre vehículos y fauna silvestre (WVC) y se utilizan con frecuencia para identificar los lugares en los cuales se presenta más constantemente los atropellamientos (puntos calientes). Si adicionalmente, se recogen datos sobre el paisaje, la carretera y el tráfico en cada WVC, la influencia de estos puede ser un insumo para construir modelos predictivos de puntos críticos de colisión [93, 111-113, 132]. Entre los factores que deben considerarse es la modalidad de recorrido del área de estudio, el momento y la frecuencia del recorrido, la duración del recorrido y la seguridad de los observadores [104]. Aunque el uso de un vehículo permite a los investigadores cubrir mayor longitud, la velocidad más alta reduce la detectabilidad de los pequeños animales [107]. Así mismo, muestreos a pie permiten una detección más completa, pero reduce la longitud de la carretera que puede ser inspeccionada de forma fiable. Adicionalmente es recomendable un muestreo a lo largo de varias estaciones o temporadas climáticas. Si solo se quiere realizar el estudio en un período de tiempo limitado, los recorridos deben realizarse cuando las especies objetivo son más activas y es probable que se encuentren en la carretera, por ejemplo, durante la migración, la reproducción o la dispersión, etc. [5].

Análisis de distribución de puntos

A continuación, se describen algunas técnicas geo estadísticas aplicables al atropellamiento de fauna y que fueron usadas en esta investigación

K Ripley

El análisis K-Ripley fue introducido por Brian D. Ripley [133] en 1981, tiene como objetivo describir la agrupación o dispersión de los datos en un rango de escalas espaciales $K(r)$ [134, 135], para luego ser comparada contra una distribución normal $Ks(r)$. A través de este análisis, se identifican las distancias de agrupación estadísticamente significativas la cual deberá ser la distancia de segmentación de la vía para realizar análisis de agrupación de puntos [136]. $K(r)$ se define por la ecuación (1) [104, 106]:

$$K(r) = \frac{D}{n(n-1)} \sum_{i=1}^{n} 2r/Ci(r) \sum_{j \neq i} f_{ij} \qquad (1)$$

Donde: K(r) es el valor del estadístico K para la escala de observación r; *D* es la longitud de la vía en km; n corresponde al número de eventos de atropellamiento; r al radio de observación en metros; Ci(r) es la longitud de la vía al interior del círculo con radio r centrado en el atropellamiento i; f_{ij} es un índice de corrección cuyo valor es igual a 0 si j está por fuera del círculo con radio r y centrado en i, o igual a 1 si j está al interior de esta área.

$$L(r) = K(r) - Ks(r) \qquad (2)$$

Donde L(r) es la diferencia entre el valor de K observado para un radio r y el valor de K simulado para una escala r; Ks(r) es la media del valor K obtenido mediante múltiples permutaciones.

Autocorrelación espacial

Por su parte, el Índice de autocorrelación espacial I tiene por objetivo medir los cambios en los valores de agrupación de los segmentos generados en una serie de distancias, encontrando las distancias en las que los datos se agrupan de forma significativa (p valor < 0.05) comparada con sus vecindades y el universo de muestreo [137]. Para esto se utiliza la ecuación (3)

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \qquad (3)$$

Donde z_i es la desviación de un atributo con respecto a su media $(x_i - \bar{X})$, $w_{i,j}$ es el peso espacial entre el atributo i y el atributo j, n es igual a la cantidad total de atributos y S_0 es la suma de todos los pesos espaciales, tal como se describe en (4).

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \qquad (4)$$

La estadística z_I se calcula a partir de la ecuación (5):

$$Z_I = \frac{I - E[I]}{\sqrt{V[I]}} \quad (5)$$

donde:

$$E[I] = \frac{-1}{n-1} \quad (6)$$

$$V[I] = E[I^2] - E[I]^2 \quad (7)$$

2D Puntos Calientes

Otra técnica geo estadística aplicable a atropellamiento de fauna es el análisis de 2D Puntos Calientes, el cual identifica las agrupaciones estadísticamente significativas de eventos a lo largo de la vía, segmentando la vía en porciones iguales y sumando los valores en un radio de búsqueda r. El valor de r corresponde a la banda de distancia con un *Z* más alto y menor valor de *p* resultantes de la evaluación del índice de autocorrelación espacial en múltiples distancias [136]. El radio de observación r se genera partiendo desde el centro de cada segmento, para luego realizar una comparación del resultado con el mismo proceso realizado a una distribución al azar. La ecuación usada se describe a continuación:

$$H_i(r) = 2r/Ci(r) \sum_{i=1}^{n} f_{ij} \quad (8)$$

Donde: $H_i(r)$ corresponde al valor de agregación evaluado para un segmento i considerando una escala r; n es el número de eventos de atropellamiento, *r* es el radio de observación; *Ci(r)* la longitud de la vía al interior del círculo con radio r centrado en el atropellamiento *i*; f_{ij} es un índice de corrección igual a 0 si *j* está por fuera del círculo con radio *r* y centrado en *i*, o igual a 1 si *j* está al interior de esta área. Para evaluar la significancia estadística de las posibles agregaciones se usa la siguiente ecuación:

$$HS = H_i(r) - Hs(r) \quad (9)$$

Donde HS es la diferencia entre el valor de agregación observado para un radio r y ubicado sobre un segmento i y el valor medio de agregaciones simuladas para una escala r ubicada sobre el segmento i [136]. A partir de esto se genera un valor de *HS* por

cada segmento de vía, el cual, al ser visualizado en sistemas de información geográfica, muestra los puntos calientes de mayor intensidad y su ubicación espacial, descartando el azar en el evento de atropellamiento y tomando en cuenta la correlación entre los puntos y su distancia.

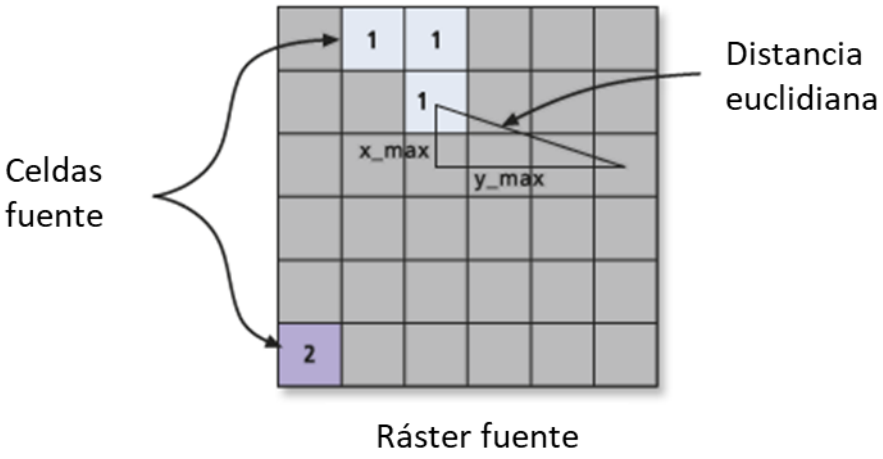
Extracción de características espaciales

Distancia euclidiana

La distancia euclidiana es la distancia ordinaria calculada entre dos puntos en un espacio euclídeo (Figura 3).

$$D = \sqrt{(x-y)^2} \quad (10)$$

Figura 3. Distancia euclidiana calculada en un raster. Fuente: [138]



Modelo digital de elevación

Los modelos de elevación digital es una representación gráfica y cuantitativa de la superficie de la tierra creada a partir de medición de elevación de terreno. Los modelos de elevación digital representan la elevación del terreno (Modelo digital de terreno) o la elevación de la superficie terrestre (Modelo digital de superficie), según se requiera. Para crear estos modelos, usualmente se utilizan mediciones de elevación en terreno y una interpolación matemática entre ellos. Sin embargo, en las últimas décadas se han desarrollado diferentes técnicas para estimar la altitud por medio de radar de

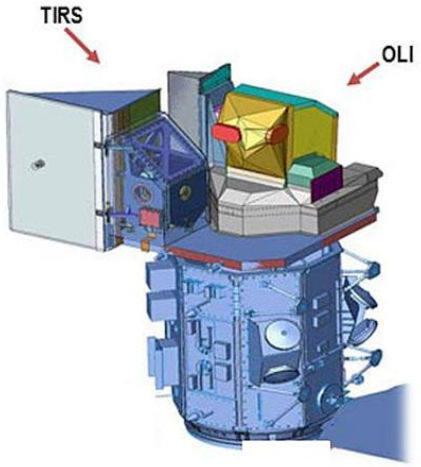
apertura sintética (SAR), esta es una forma de radar que se utiliza para crear imágenes bidimensionales o reconstrucciones tridimensionales de objetos, para lograrlo se utiliza el movimiento de una antena de radar sobre una región objetivo para proporcionar una resolución espacial más fina que la de los radares de exploración de haz convencional. Adicionalmente, la adquisición de imágenes SAR es independiente de la iluminación natural, por lo cual, las imágenes pueden ser tomadas de noche. El radar utiliza radiación electromagnética a frecuencias de microondas, lo que representa un beneficio significativo en comparación con otras técnicas de radar, lo anterior debido a que este tipo de onda no es absorbida por la atmósfera. Actualmente, existen múltiples satélites SAR en órbita, entre los cuales se encuentran los satélites: Sentinel-1A y el Sentinel-1B, los cuales proporcionan cobertura InSAR a escala global con un ciclo de repetición de 6 días. Así mismo, existen otros satélites en órbita como el RADARSAT-1, el TerraSAR-X y el Cosmo SkyMed [139, 140].

A partir de esto, la topografía de radar por transbordador (SRTM) fue utilizada en la misión del transbordador espacial en el año 2000, en donde un SAR, a bordo del Transbordador Espacial Endeavour fue utilizado para adquirir datos topográficos de la superficie terrestre. Para esto fue necesario el uso de dos antenas de radar ubicadas en la bahía de carga útil del transbordador y en un mástil de 60 metros extendido desde la bahía de la carga útil [141]. Esta misión realizó una medición por InSAR y fue realizada por Intermap Technologies quien fue el principal contratista para el procesamiento de los datos recolectados por la misión, a partir de los cuales se generaron un modelo de elevación digital para toda la superficie terrestre.

Imágenes del Satélite Landsat 8

El satélite de observación terrestre Landsat 8 (L8) es un sistema de exploración con propósitos civiles y científicos, está diseñado para una órbita sincrónica al Sol de 705 km de altitud, con un ciclo de repetición de 16 días, orbitando completamente la Tierra cada 98,9 minutos. El L8 lleva una carga útil de dos sensores: el Operational Land Instrument (OLI), y el Thermal Infrared Sensor (TIRS), formando de manera conjunta lo que se conoce como el Observatorio L8 (Figura 4) [142].

Figura 4. Componentes del Satélite Landsat 8 – Observatorio L8. Fuente: Data Users Handbook para el Landsat 8 – USGS [142]



El OLI y el TIRS recogen los datos conjuntamente para proporcionar imágenes coincidentes de las mismas áreas de superficie. El objetivo de la programación y la reunión de datos es proporcionar imágenes con baja nubosidad del planeta para cada estación del año. Cada imagen es corregida de forma radiométrica y registrada en una proyección cartográfica con corrección del movimiento terrestre, dando como resultado una imagen orto rectificada [142].

El sensor del OLI recoge datos de imágenes para 9 bandas espectrales de onda corta en una franja de 190 km con una resolución espacial de 30 metros por píxel (m/px), excepto la banda pancromática, cuya resolución es de 15 m por píxel [142]. Como el OLI, el TIRS utiliza fotodetectores de infrarrojos de pozo cuántico (QWIP) para medir la energía infrarroja térmica de onda larga (TIR) emitida por la superficie de la tierra, cuya intensidad es una función de la temperatura de la superficie.

En la Figura 5 se pueden observar las bandas de luz captadas por el L8 y sus respectivas resoluciones espaciales [142].

Figura 5. Bandas de distancia electromagnética captadas por los sensores OLI y TIRS del satélite Landsat 8. Fuente: Data Users Handbook para el Landsat 8 – USGS [142]

	LANDSAT 8	
	Longitud de onda (µm)	Resolución (m)
Banda 1 - Coastal Aerosol	0,435 – 0,451	30
Banda 2 - Blue	0,452 - 0,512	30
Banda 3 - Green	0,533 - 0,590	30
Banda 4 - Red	0,636 - 0,673	30
Banda 5 - Near Infrared (NIR)	0,851 - 0,879	30
Banda 6 - Short-wave Infrared (SWIR) 1	1,566 - 1,651	30
Banda 7 - Short-wave Infrared (SWIR) 2	2,107 - 2,294	30
Banda 8 - Panchromatic	0,503 - 0,676	15
Banda 9 - Cirrus	1,363 - 1,384	30
Banda 10 - TIR 1	10,60 - 11,19	100
Banda 11 - TIR 2	11,50 - 12,51	100

A partir de estas bandas espectrales, es posible realizar el cálculo de diferentes índices multiespectrales [143], algunas de las más utilizadas se presentan en la Tabla 1.

Tabla 1. Índices multiespectrales. Fuente: compilación propia

índice	Ecuación para el L8	Uso	Referencia
Índice normalizado de diferencia de vegetación (NDVI)	$\frac{NIR - RED}{NIR + RED}$	Resalta la cantidad de verde captada por un sensor multiespectral también conocido como biomasa relativa	[144]
Índice de vegetación de diferencia normalizada verde (GNDVI)	$\frac{NIR - GREEN}{NIR + GREEN}$	Determina la captación de agua y nitrógeno en el dosel vegetal	[145]
Índice de vegetación mejorada (EVI)	$2.5 * \frac{NIR - RED}{NIR + 6 * RED - 7.5 * BLUE + 1}$	Sensible a las zonas con vegetación densa	[146]
Índice de vegetación avanzada (AVI)	$\sqrt[3]{(RED) * (GREEN) * (RED - GREEN)}$	Destaca diferencias sutiles en la densidad del dosel	
Índice de suelo desnudo (BI)	$\frac{RED + BLUE - GREEN}{RED + BLUE + GREEN}$	Mejora la identificación de las zonas de suelo desnudo	[147-149]
Índice de sombra o Índice de sombra a escala (SI, SSI)	$\sqrt{(BLUE) * (GREEN)}$	Aumenta a medida que aumenta la densidad del bosque	
Índice de vegetación ajustada al suelo (SAVI)	$\frac{NIR - RED}{NIR + RED + 0.5} * 1.5$	Modificación del NDVI con un factor de corrección del brillo del suelo	[150]
Índice de diferencia de humedad normalizada índice de quema (NDMI o NDWI o NBRI)	$\frac{NIR - SWIR}{NIR + SWIR}$	Detecta cambios en el contenido de agua de la vegetación	[151, 152]
Índice de estrés hídrico (MSI)	$\frac{SWIR1}{NIR}$	Predicción de la productividad agrícola y la modelización biofísica de la vegetación	[153]
Índice de clorofila verde (GCI)	$\frac{NIR}{GREEN}$	Estimación el contenido de clorofila en las hojas de diversas especies de plantas	[154]

Selección de características

Con el objetivo de mejorar los resultados del proceso de clasificación de un algoritmo de aprendizaje de máquina, es importante realizar procesos de normalización y selección de características, a continuación, se describen algunas técnicas usadas por esta investigación.

Normalización

La normalización de los conjuntos de datos es un requisito común de muchos estimadores de aprendizaje de máquina, los cuales podrían comportarse de modo insatisfactorio si las características individuales no se asemejan a una distribución normal con media cero y varianza unitaria. Esto es importante debido a que, muchos elementos utilizados en la función de costo de la mayoría de los algoritmos de aprendizaje suponen que todas las características están centradas en torno a cero y tienen una varianza en el mismo orden. Por lo cual, si una característica tiene una varianza de órdenes de magnitud mayor que otras, podría dominar la función objetivo y hacer que el estimador no pueda aprender de otras características correctamente como se espera [155].

Por esta razón, es recomendable realizar una normalización de la base de datos, para esto, es usual utilizar métodos de escalamiento como la función MinMaxScaler propia de la librería Scikit-learn [156], los cuales utilizan los valores máximos y mínimos como referencia de la función de transformación. En (23) puede observarse la ecuación de normalización.

$$Caracteristica\ escalada = \sigma * (X_{max} - X_{\min}) + \mathrm{X}_{\min} \qquad (23)$$

Donde

$$\sigma = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (24)$$

Selección de características por K Best

La selección de características por medio del resultado de técnicas estadísticas univariadas es un proceso usualmente llamado reducción de dimensión, este proceso es fundamental previo al entrenamiento de un modelo de aprendizaje de máquina.

Información Mutua (MI) es una prueba estadística que mide que tanta información brinda una variable aleatoria de otra. Su resultado es una cantidad adimensional usualmente descrita con la unidad bits. Esta medida de Información Mutua puede ser útil como una forma de reducir la incertidumbre de un resultado al identificar variables que brinden información de otra variable igualmente aleatoria. Altos valores de información mutua indica una mayor reducción de la incertidumbre, mientras que una información mutua igual a cero indica que ambas variables son independientes [157]. En (25) se muestra la definición teórica para variables discretas de la medida de información mutua $I\ (X,\ Y)$, tal como fue descrita por [158].

$$I(X,Y) = \sum_{x,y} P_{xy}(x,y) log \frac{P_{xy}(x,y)}{P_x(x), P_y(y)} = E_{p_{xy}} \log \frac{P_{xy}}{P_x P_y} \qquad (25)$$

Donde

$$P_x(x) = \sum_{y} P_{xy}(x,y) \quad P_y(y) = \sum_{x} P_{xy}(x,y) \quad (26)$$

En términos prácticos, $I\ (X,\ Y)$ es una representación de la entropía o incertidumbre de una variable aleatoria X en la cual, al encontrar información de esta en otra variable aleatoria Y, se reduce la incertidumbre del valor que tomará la variable X dado la variable Y [157]. La selección de características por medio de métodos de estimación de la información mutua fue descrita originalmente por Ross [159] y se encuentra implementado en la librería Scikit-learn en la función mutual_info_classif [156] para Python.

Adicionalmente, la prueba de Chi cuadrado es una prueba estadística usada como prueba de independencia entre variables de una tabla de contingencia. En otras palabras, permite probar si la distribución de una variable difiere significativamente de las demás. La prueba de Chi cuadrado está definida en la ecuación (27).

$$\boldsymbol{\chi_c^2} = \sum_{i=0}^{n} \sum_{j=0}^{n} \frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}} \qquad (27)$$

Donde *c* son los grados de libertad dados por la ecuación $c = (r-1)(col-1)$, donde *r* es la cantidad de filas de la tabla de contingencia y *col* es la cantidad de columnas, $O_{i,j}$ es el valor observado de la muestra *i, j* y $E_{i,j}$es el valor esperado para la observación *i, j.* La selección de características por medio del Test de Chi cuadrado se encuentra implementado en la librería Scikit-learn en la función chi2 [156]

De igual manera se tiene la prueba F del análisis de la varianza (ANOVA), la cual es una prueba estadística que busca estimar los factores de varianza asociados en un conjunto de datos. El ANOVA se basa en la ley de la varianza total, en la cual la varianza observada en una variable particular se divide en componentes atribuibles a diferentes fuentes de variación. Esta prueba también se conoce como prueba de comparaciones múltiples ANOVA y se encuentra definida en la ecuación (28).

$$F = \frac{\frac{\sum_{i=1}^{K} n_i(\overline{Y}_i - \overline{Y})^2}{K-1}}{\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_i\right)^2}{N-K}} \qquad (28)$$

Donde $\overline{Y}_i$ denota la media de la muestra en el *i* -esimo grupo, n_i corresponde al número de observaciones del *i* -esimo grupo, $\overline{Y}$ corresponde a la media global de los datos y *K* el número de grupos. Adicionalmente, Y_{ij} corresponde a la observación j-esima del *i*-esimo grupo de *K* grupos posibles y *N* es el tamaño de la muestra general [160].

2.2.2 Entrenamiento y validación en la zona de entrenamiento

Técnicas de sobre muestreo

En caso de presentarse desbalance de clases en la base de datos de entrenamiento de un algoritmo de aprendizaje de máquina, es deseable utilizar técnicas de sobre muestreo para balancear la base de datos. A continuación, se muestran algunas técnicas utilizadas por este proyecto:

Muestreo sintético adaptativo (ADASYN)

El re-muestreo sintético adaptativo (ADASYN) se basa en la idea de generar de forma adaptativa muestras de datos minoritarios según sus distribuciones, para esto se

generan más datos sintéticos de las muestras de clases minoritarias que son más difíciles de aprender en comparación con las muestras minoritarias que son más fáciles de aprender. El método ADASYN no solo puede reducir el sesgo de aprendizaje introducido por la distribución original de los datos en desequilibrio, sino que también puede cambiar de forma adaptativa el límite de decisión para centrarse en las muestras difíciles de aprender [161].

El algoritmo ADASYN consiste en generar muestras sintéticas de la clase minoritaria a partir de muestras elegidas por medio del algoritmo de K Vecinos más cercanos – KNN. En la ecuación (29) se presenta la función matemática a partir de la cual se genera cada muestra sintética.

$$S_i = X_i + (X_{zi} - X_i) * \lambda \qquad (29)$$

Donde $(X_{zi} - X_i)$ es el vector de diferencia en *n* espacios dimensionales entre una muestra de la clase minoritaria al azar X_{zi} y el dato elegido por medio del algoritmo de KNN X_i y λ es un número aleatorio $\lambda \in [0,1]$ [161].

Técnica de sobre muestreo de minorías sintéticas (SMOTE)

La técnica de sobre muestreo de minorías sintéticas (SMOTE) permite balancear una base de datos creando ejemplos sintéticos por medio de reemplazo directo. Este enfoque se inspira en la técnica propuesta por Ha y Bunke [162] los cuales crearon datos de entrenamiento extra realizando operaciones como la rotación de las imágenes de entrenamiento. De igual manera, SMOTE genera ejemplos sintéticos a partir de las características. La clase minoritaria, por lo tanto, es sobre muestreada tomando muestras de la clase minoritaria y datos sintéticos a lo largo de los segmentos de la línea que unen los vecinos más cercanos de la clase minoritaria[163].

De esta técnica existen múltiples variaciones según la forma en la que elige las muestras, Borderline SMOTE realiza una selección de las muestras a re-muestrear a partir de un algoritmo de clasificación KNN [164], KMeans SMOTE lo realiza a partir del algoritmo de clasificación no supervisada KMeans [165] y el algoritmo SVM SMOTE decide cuál muestra será re muestreada a partir de la función de decisión creada por una máquina de Soporte Vectorial [166].

Algoritmos de Clasificación

Con el auge de la inteligencia artificial y el aprendizaje de máquina, han surgido incontables algoritmos de clasificación supervisada y no supervisada. En esta sección se presentan los algoritmos utilizados durante esta investigación y su fundamentación teórica.

Algoritmo de clasificación: K vecinos más cercanos (KNN)

El clasificador KNN es un modelo de aprendizaje de máquina no paramétrico, este algoritmo se caracteriza por memorizar el conjunto de datos de entrenamiento, por lo cual, aunque técnicamente no aprende de los datos, puede realizar predicciones con bastante precisión siempre y cuando no cambie el patrón observado en los datos. El algoritmo KNN en sí mismo es bastante sencillo y se puede resumir en los siguientes pasos: 1. Elegir el valor de k y una métrica de distancia. 2. Encontrar los vecinos más cercanos a la muestra que queremos clasificar. 3. Asignar la etiqueta de la clase por mayoría de votos.

Basado en la métrica de distancia elegida, el algoritmo KNN encuentra la cantidad k de muestras en el conjunto de datos de entrenamiento que están más cerca del punto desconocido, la etiqueta de la clase del nuevo punto de datos se determina entonces por un consenso entre las clases de sus vecinos más cercanos. La principal ventaja de este enfoque basado en la memoria es que el clasificador se adapta inmediatamente a medida que recogemos nuevos datos de entrenamiento. Sin embargo, el inconveniente es que la complejidad computacional para clasificar nuevas muestras crece linealmente con el número de muestras del conjunto de datos de entrenamiento [167]. En (30) se observa la ecuación más común para el cálculo de la distancia euclidiana.

$$D = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad (30)$$

Algoritmo de clasificación: Máquina de Soporte Vectorial (SVM)

Las máquinas de soporte vectorial (SVM) son un algoritmo de aprendizaje supervisado que puede ser utilizado en problemas como la clasificación por vectores de soporte (SVC) y la regresión por vectores de soporte (SVR). La implementación de este

algoritmo se basa en encontrar una función optimizada que permita separar los conjuntos de clasificación de la forma más ideal posible, permitiendo así clasificar los valores desconocidos y_n, para esto, se traza un hiperplano por medio de una función hallada por medio de mínimos cuadrados (31) estimando una función de costo de minimización del error (32).

$$L(w,b) = y_n(w^T x_n + b) > 1 - \zeta \;\; si\; y_n = 1 \mid < 1 + \zeta \;\; si\; y_n = -1 \quad (31)$$

$$\frac{1}{2}\|w\|^2 + C\sum_i \zeta^i \qquad (32)$$

Donde w es un vector de pesos, b es el sesgo, C es la variable de penalización y ζ es el factor de relajación de frontera tal como fue propuesto por Cortes y Vapnik [168].

Si la base de datos x_n no es linealmente separable por medio de una función lineal (31), es recomendable proyectar la base de datos x_n a un espacio dimensional superior usando funciones Kernel [167], entre los Kernel más usados se encuentran la función de base radial (rbf), Kernel polinomial, Kernel sigmoide, Kernel lineal, entre otros. La formulación matemática de estos puede ser observada en Hofmann, et al. [169].

Algoritmo de clasificación: Bosques Aleatorios (RF)

Los árboles de decisión son modelos de clasificación basados en un algoritmo de descarte por medio de preguntas, el cual, en el caso de los algoritmos de aprendizaje de máquina, se aprende una serie de preguntas que permiten inferir la clase a la que corresponde una muestra desconocida. Estas preguntas o comparaciones por medio de umbrales permiten separar los datos hasta llegar a la clase perteneciente. El concepto implementado por este algoritmo genera un árbol de decisión en el cual se separan los datos en ramas de decisión usando la característica que posea mayor ganancia de información (IG) (33) y repitiéndolo hasta llegar a un punto en que todas las hojas de cada rama no puedan dividirse más. Uno de los problemas con esta aproximación es que si la profundidad del árbol es muy grande, puede llevar fácilmente a un sobre ajuste del modelo [167], lo que implica que el algoritmo memoriza de una forma óptima los datos, pero que no genera un ajuste al fenómeno descrito por ellos.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \qquad (33)$$

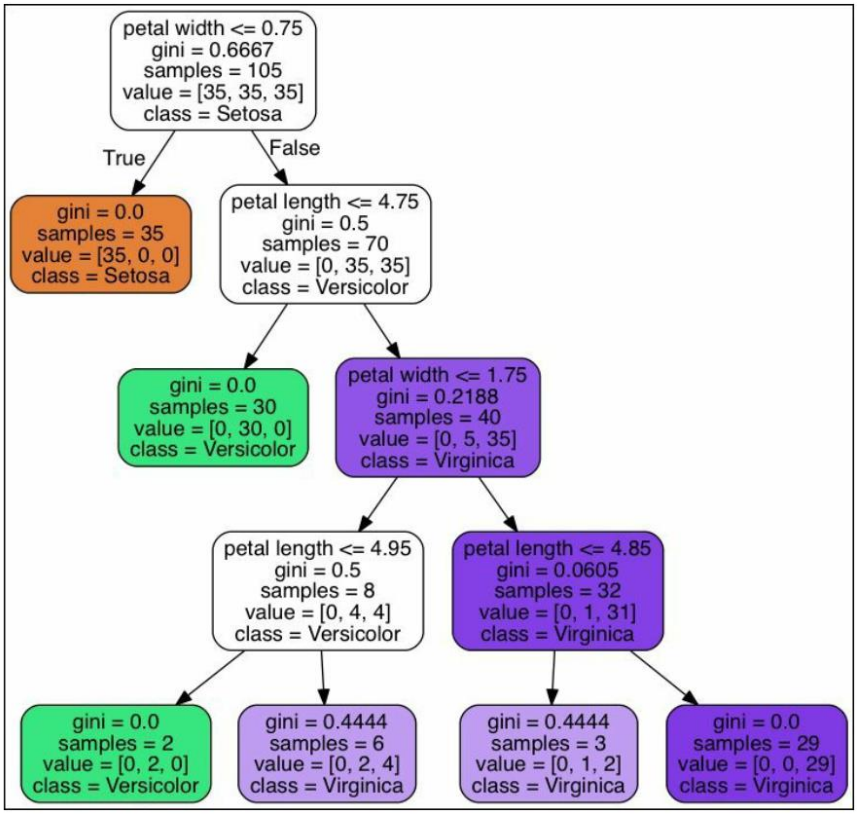
Donde, f es la característica sobre la que se va a realizar la división, D_p y D_j son los nodos padre del hijo *j*-esimo, I es la medida de impureza, N_p corresponde al número total de muestras en el nodo padre y N_j es el número de muestras en el nodo hijo j-esimo.

Como puede observarse, la ganancia de información corresponde a la diferencia de impureza entre el nodo padre y la suma de impurezas de los nodos hijos, entre menor sea la impureza entre los nodos hijo, mayor será la ganancia de información. Con el objetivo de disminuir el costo computacional, el algoritmo implementado en esta investigación divide cada nodo padre en dos nodos hijo D_{left} y D_{right} por lo cual la IG puede expresarse como:

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right}) \quad (34)$$

Siendo I la medida de impureza calculada a través del criterio de división, comúnmente este puede ser Impureza Gini, entropía y error de clasificación. Por lo tanto, el árbol de decisión, será la construcción de n divisiones hasta llegar a tener hojas puras para cada característica, de modo tal que permita la clasificación de cada muestra [167]. En la Figura 6 se muestra un árbol de decisión creado para la base de datos Iris de clasificación taxonómica de especies vegetales [170].

Figura 6. Árbol de decisión creado a través de la librería Scikit-learn [156] para la base de datos de prueba Iris. Fuente: [167]



Así mismo, los bosques aleatorios han ganado una enorme popularidad en las aplicaciones de la máquina aprendizaje durante la última década debido a su buen desempeño en la clasificación, escalabilidad y facilidad de uso. Intuitivamente, un bosque aleatorio puede ser considerado como un conjunto de árboles de decisión. La idea detrás de un bosque aleatorio es sumar las fronteras de decisión de árboles, lo cual permite construir un modelo más robusto que tenga un mejor rendimiento de generalización y sea menos susceptible al sobre entrenamiento. El algoritmo de bosque aleatorio puede ser resumido en cuatro simples pasos: se inicializa con una muestra aleatoria de arranque de tamaño n, a partir de este se realiza un árbol de decisión maximizando la ganancia de información, por último, se agrega la predicción de cada árbol para asignar la etiqueta de clase por mayoría de votos [167].

Algoritmo de clasificación: Redes Neuronales Artificiales (RNA)

Las redes neuronales artificiales (RNA) son capas de neuronas de procesamiento altamente interconectados que realizan una serie de transformaciones en los datos para aprender de los patrones detectados entre ellos. Modeladas a partir del cerebro humano, las RNA tienen como objetivo que las máquinas imiten el funcionamiento del

cerebro [171]. Las redes neuronales están basadas en el perceptrón propuesto por Rosenblatt [172] considerado el método más simple de una red neuronal artificial para problemas de clasificación linealmente separables por un hiperplano (35), este consiste por una única neurona con una variable de peso y un parámetro de sesgo.

$$v = \sum_{i=1}^{m} w_i x_i + b \quad (35)$$

Donde w_i serán los pesos calculados para las características x_i y un valor de sesgo o ajuste vertical b.

A partir de esto, se propuso la articulación de múltiples neuronas a través de funciones no lineales y su posterior entrenamiento por bloques de datos. Sin embargo, al observar un aumento en la complejidad del problema de optimización, se propuso realizar una propagación de los datos hacia delante de la red hasta llegar a la salida, para luego realizar una optimización hacia atrás, calculando el error general en la predicción de la red para luego calcular el error en cada una de las neuronas. Este algoritmo llamado retro propagación fue propuesto por David y James [173] y se considera como uno de los mayores logros en el desarrollo de las redes neuronales [171], en (35) y (36) puede observarse su formulación matemática.

$$\frac{dC}{dw^L} = \frac{dC}{da^L} * \frac{da^L}{dz^L} * \frac{dz^L}{dw^L} * \frac{dz^L}{db^L} \quad (35)$$

Donde C equivale a nuestra función de costo, a^L a la función de activación, z^L a la suma ponderada de la neurona, w^L al parámetro de pesos y b^L el parámetro de sesgo, cada uno de estos para la última capa L

Posteriormente, se recorre cada capa de la red hacia atrás hasta llegar a las entradas, permitiendo así ajustar el parámetro w de cada neurona, permitiendo de esta manera una función optimizada tal como se observa en (36) y (37)

$$\frac{dC}{dw^{L-1}} = \frac{dC}{da^L} * \frac{da^L}{dz^L} * \frac{dz^L}{da^{L-1}} \frac{da^{L-1}}{dz^{L-1}} * \frac{dz^{L-1}}{dw^{L-1}} \quad (36)$$

$$\frac{dC}{db^{L-1}} = \frac{dC}{da^L} * \frac{da^L}{dz^L} * \frac{dz^L}{da^{L-1}} \frac{da^{L-1}}{dz^{L-1}} * \frac{dz^{L-1}}{db^{L-1}} \quad (37)$$

Por último, se tendrá un parámetro llamado tasa de aprendizaje, la cual, indica la tasa a la cual se modificarán los pesos de las neuronas a cada paso del algoritmo de retro propagación, generando así una tasa de descenso de gradiente en la función de costo. Si este valor es muy grande, el algoritmo de retro propagación podría no encontrar los valores de mínimos de la función, mientras que, si es un valor muy bajo, el algoritmo de optimización podría quedarse en un mínimo local de la función, por lo cual, la elección de este parámetro influye directamente en el desempeño de la optimización de la red neuronal [171].

Métricas de validación

Una de las fases más importantes durante el proceso de implementación de un algoritmo de aprendizaje de máquina, será la medición o cuantificación de desempeño del algoritmo. Para esto, usualmente se utiliza la matriz de confusión del algoritmo como base para el cálculo de diferentes métricas de desempeño [174]. En la Figura 7 se muestra la matriz de confusión para un problema de clasificación binario.

Figura 7. Matriz de confusión para un problema de clasificación binaria. Fuente: imagen modificada de Raschka y Mirjalili [167]

		Predicción	
		Positivos	**Negativos**
Observación	**Positivos**	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (VN)

Las métricas usadas en esta investigación se presentan a continuación.

F1 Score

También conocida como el promedio ponderado de la precisión y la sensibilidad, la métrica F1 Score permite estimar la precisión de un algoritmo ponderado por la clase positiva, permitiendo de esta manera medir la capacidad de un algoritmo de predecirla de manera adecuada. Esta métrica es recomendada para problemas desbalanceados, debido a que mide la exactitud ponderada por la clase positiva, usualmente minoritaria en este tipo de problemas [167]. La ecuación para su cálculo puede observarse en (38)

$$F1 = 2 * \frac{Precisión * Sensibilidad}{Precisión + Sensibilidad} \quad (38)$$

Donde

$$Precisión = \frac{VP}{VP + FP} \quad (39)$$

$$Sensibilidad = \frac{VP}{FN + VP} \quad (40)$$

Kappa

Principalmente utilizada en problemas desbalanceados la métrica Kappa de Cohen permite reducir el sesgo inducido por el desbalance de clases sobre las métricas de desempeño tal como fue demostrado por Ben-David [175]. La idea principal de la métrica Kappa es compensar el sesgo producido por la probabilidad mayor de encontrar una muestra de la clase mayoritaria, tal como se muestra en (41)

$$Precisión\ estimada = \frac{E(VP) + E(VN)}{N} \quad (41)$$

Donde $E(VP)$ y $E(VN)$ son el producto de la distribución marginal de los verdaderos positivos y verdaderos negativos, tal como se muestra en (42) y (43)

$$E(VP) = \frac{VP * (VP + FP)}{N} \quad (42)$$

$$E(VN) = \frac{VN * (FN + VN)}{N} \quad (43)$$

La precisión estimada es luego normalizada para calcularse el valor Kappa tal como se muestra en (44)

$$K = \frac{precisión\ observada - precisión\ estimada}{1 - precisión\ estimada} \quad (44)$$

A través de esta métrica es posible diferenciar desempeños sobreestimados debido a desbalance en las clases de clasificación tal como fue descrito por Fernández, et al. [176].

Área bajo la curva de la respuesta característica del operador (AUC-ROC)

La curva de respuesta operativa característica del receptor (ROC) es una técnica estándar para evaluar clasificadores en conjuntos de datos que exhiben un desequilibrio de clases. Las curvas ROC logran tener una insensibilidad sesgada resumiendo el rendimiento de un clasificador en un rango de tasas positivas verdaderas (TVP), estimado como se muestra en la ecuación (45) y tasas positivas falsas (TFP) (46) [177]. Evaluando los modelos con varias tasas de error, las curvas ROC son capaces de determinar qué proporción de instancias será correctamente clasificadas para una determinada TFP.

$$TFP = \frac{FP}{VN + FP} \quad (45)$$

$$TVP = \frac{VP}{VP + FN} \quad (46)$$

Para generar la curva ROC, cada punto es generado al mover la frontera de decisión del clasificador, de este modo se cuantifica la capacidad real de un clasificador, el área bajo la curva (AUC) -ROC se ha convertido en la métrica estándar de facto para evaluar clasificadores para aplicaciones desbalanceadas [178]. Esto se debe al hecho de que es independiente tanto del umbral seleccionado como de las probabilidades previas, además de ofrecer un único número para comparar clasificadores. Uno de los principales beneficios que presenta el AUC-ROC es que puede considerarse que mide la frecuencia con que una instancia de clase positiva aleatoria se clasifica por encima de una instancia de clase negativa aleatoria, cuando se clasifica según su probabilidad [176].

Métodos de optimización

Con el objetivo de maximizar el desempeño de los algoritmos de aprendizaje de máquina, es usual que se utilicen métodos de optimización de hiper parámetros. A continuación, se presentan los métodos usados en esta investigación.

Algoritmos genéticos (GA)

El algoritmo genético fue concebido inicialmente por Holland como un medio para estudiar el comportamiento adaptativo [179]. Sin embargo, se han considerado en gran medida como métodos de optimización de algoritmos inspirado en la recombinación genética como medio para generar nuevas soluciones candidatas.

Los algoritmos genéticos son enfoques de búsqueda heurística aplicables a una amplia gama de problemas de optimización. La evolución es la base de los algoritmos genéticos, la actual variedad y éxito de las especies es una buena razón para creer en el poder de la evolución. Las especies son capaces de adaptarse a su entorno. Se han convertido en estructuras complejas que permiten la supervivencia en diferentes tipos de entornos, el apareamiento y la obtención de descendencia para evolucionar pertenecen a los principios fundamentales del éxito de la evolución. Estas son buenas razones para adaptar los principios evolutivos a la solución de los problemas de optimización [180].

Un algoritmo genético básico de optimización de hiper parámetros puede ser reducido a una serie de pasos para seleccionar los mejores candidatos: una etapa de inicialización, usualmente aleatoria, de los hiper parámetros de la población, una etapa de combinación aleatoria de los hiper parámetros de dos miembros de la población [181], generación de mutaciones aleatorias en los parámetros a optimizar [182], una etapa de evaluación del desempeño de los algoritmos descendientes y por último, una etapa de selección de los mejores descendientes [180].

3. Marco metodológico

Este proyecto se divide en 3 fases: una fase de caracterización del atropellamiento de fauna a partir de mapas e imágenes multiespectrales, seguida por una fase de selección de características mediante técnicas univariantes, por último, una fase de entrenamiento y validación en la cual se busca encontrar el algoritmo que mejor se ajuste al fenómeno del atropellamiento, además de realizar una transferencia de aprendizaje generando predicciones en zonas que no posean información suficiente para realizar medidas de mitigación.

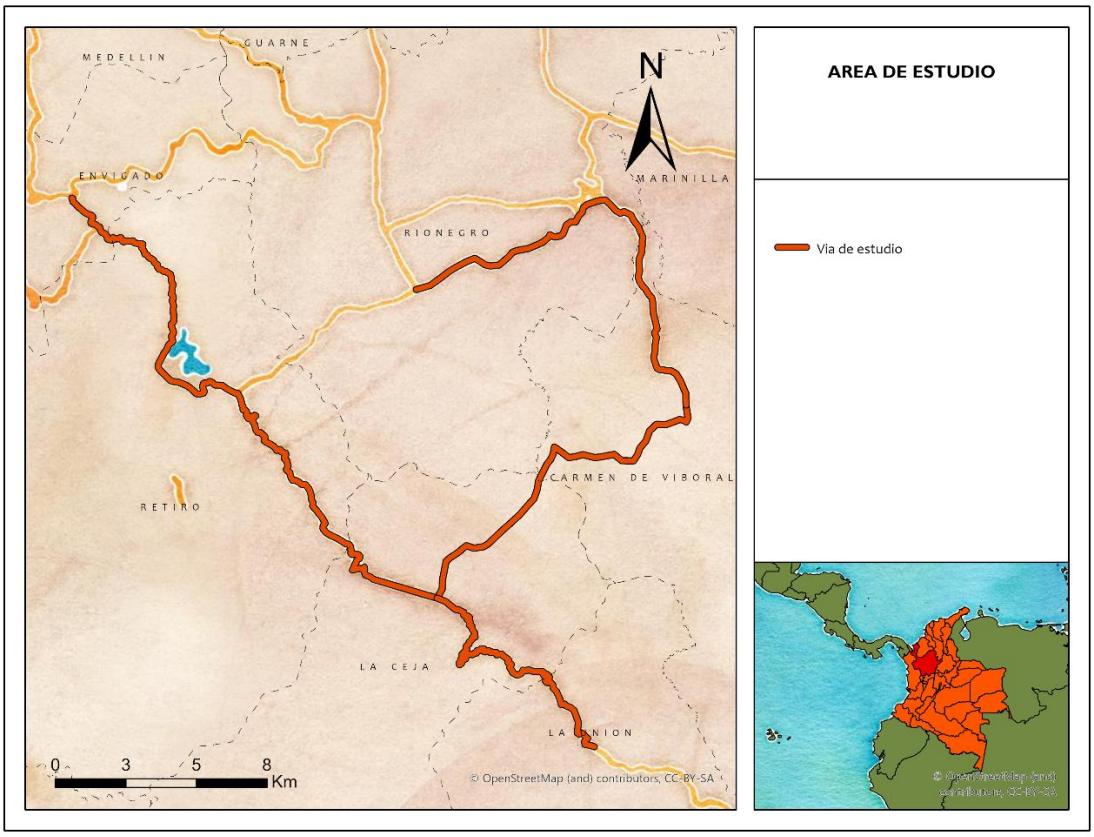
3.1. Caracterización del fenómeno del Atropellamiento de Fauna Silvestre

3.1.1. Área de estudio

Este proyecto tiene como área de estudio (Figura 8) las vías que comunican los municipios de Envigado (1575 m.s.n.m.), La Ceja (2200 m.s.n.m.), El Carmen de Viboral (2150 m.s.n.m.) y Rionegro (2135 m.s.n.m.), formando una red de 71 kilómetros lineales de carretera que según el Instituto Nacional de Vías (INVIAS), 56.48 km son considerados como Vías primarias y 13.71 km como secundarias. Estas vías presentan usualmente un alto flujo vehicular, especialmente en fines de semana que generan múltiples atropellamientos de fauna [183].

Los municipios mencionados están ubicados en el Valle de San Nicolás, también conocido como el Altiplano de Oriente u Oriente antioqueño. Sus temperaturas oscilan entre 9 y 24 ° C y se caracterizan por tener actividades semi agrícolas, entre las que se destaca el cultivo de flores y la producción de lácteos. Adicionalmente, es una región que ha tenido un crecimiento considerable de su economía en los últimos años, llegando incluso a duplicar el crecimiento del Departamento y la Nación, tal cual como lo reporta el artículo de prensa de la Cámara de Comercio del Oriente Antioqueño del 27 de enero de 2019: "El Oriente Antioqueño duplica el crecimiento de Antioquia y Colombia en la creación de empresas". Este crecimiento ocasionó una expansión de los cascos urbanos de múltiples municipios, lo cual ha llevado a un crecimiento de su infraestructura carretera [184].

Figura 8. Área de estudio, vías de los municipios de Envigado, El Retiro, La Ceja, El Carmen y Rionegro. Oriente antioqueño- Colombia. Fuente: autoría propia



Así mismo, esta zona se caracteriza por tener una predominancia de coberturas de vegetación secundaria, mosaicos de cultivo con espacios naturales, pastos, plantaciones forestales y bosques abiertos. Debido a lo anterior, es una zona con presencia constante de fauna. Las especies animales más avistadas en la zona son la Ardilla Colirroja (*Notosciurus granatensis*), el Agutí Centroamericano (*Dasyprocta punctata*), la Zarigüeya común (*Didelphis Marsupialis*), La Paca (*Cuniculus paca*), el Perro de Monte (*Potos flavus*), entre otros [185].

3.1.2. Recolección de datos de atropellamiento de fauna silvestre

Durante el desarrollo de esta investigación se extrajo los reportes de atropellamiento de fauna contenidos en la App Recosfa para el área de estudio [186].

Esta base de datos consta de 6204 reportes de atropellamiento recolectados entre 2016 y el 30 de agosto de 2020, cada uno con su respectiva ubicación en coordenadas planas (latitud longitud), la clase a la que pertenece, la especie (si es posible diferenciarla) y la fotografía. La información contenida en el aplicativo procede de diversos proyectos de atropellamiento de fauna realizados por el Instituto Tecnológico Metropolitano (ITM) en los cuales se hizo uso de la App Recosfa para la recolección de información. Sin embargo, debido al carácter abierto de la aplicación, ha sido ampliamente utilizada por la ciudadanía para reportar atropellamientos de fauna. Así mismo, gracias al trabajo conjunto entre el ITM y la Agencia Nacional de Infraestructura - ANI, la App Recosfa recibe reportes mensuales por parte de diferentes concesiones viales en todo el país, generando de esta manera, una base de datos robusta para el entendimiento, análisis y estudio del fenómeno del atropellamiento de fauna en las vías de Colombia.

3.1.3. Análisis de distribución de puntos

A partir de los puntos de atropellamiento de fauna recopilados, se realizaron análisis de patrones de puntos con el objetivo de identificar las zonas con acumulaciones estadísticamente significativas de atropellamiento. Para esto se hizo uso del análisis geo estadístico K Ripley, representando de manera gráfica las escalas espaciales en las cuales las agrupaciones de puntos difieren significativamente del azar, este análisis fue realizado con una distancia de observación inicial de 100 metros con incrementos de 100 metros hasta no obtener cambios en el valor de L(r) de la ecuación (2). Consecuentemente, se procedió a generar agrupaciones de puntos en la escala espacial identificada como significativa por el estadístico K Ripley para realizar pruebas de Autocorrelación espacial. Posteriormente se calculó el valor medio de agrupación de cada segmento y la varianza entre las agrupaciones generadas, calculando la diferencia entre la media global y cada segmento, obteniendo así la desviación de cada agrupación con respecto a la media de su vecindad y su universo de muestreo. Cuando el valor de este estadístico es mayor que la media se tendrá como resultado un valor positivo, por el contrario cuando los segmentos tienen valores de agrupación diferentes con respecto a la media global se tendrá un valor negativo, si estos valores positivos están espacialmente agrupados, el índice de Autocorrelación espacial será positivo, mientras que, si existen valores positivos y valores negativos en la misma vecindad, tendrá como resultado un índice de autocorrelación espacial negativo [187].

Por último, se realiza un análisis de puntos calientes a través del software Siriema [106]. Para este se utilizó la banda de distancia identificada como significativa por el

análisis de autocorrelación espacial y una división del tramo de estudio en 1000 segmentos equidistantes, de esto se obtuvo una distribución de puntos con su correspondiente valor de acumulación (HS) y los límites de confianza superior (UCL) e inferior (LCL), considerándose como punto caliente significativo si HS > UCL y dispersión significativa si HS < LCL. Para el propósito de este proyecto se generaron 2 clases de clasificación, la clase 1 son todos aquellos segmentos de vía cuyos valores de HS superen el límite de confianza superior, mientras que la clase 0 son los segmentos que no cumplan esta condición. Esto con el propósito de tener una base de datos con los puntos calientes y los puntos fríos no significativos.

3.1.4. Extracción de características

A partir de los puntos calientes identificados, se procedió a generar una matriz de características que describa el atropellamiento de fauna en el área de estudio, la cual fue usada durante la fase de entrenamiento de los modelos. Para esto se realizaron 2 fases principales: una fase de recolección de mapas e imágenes satelitales y una fase de selección de características.

Recolección de Mapas e Imágenes satelitales

Con el propósito de identificar las variables más relacionadas con el atropellamiento se descargaron mapas de la zona de estudio con las características ambientales propuestas. Inicialmente, una capa de vías y ríos fue descargada del Marco Geo estadístico Nacional (MGN) para el Departamento de Antioquia [188], esta incluye las vías presentes en el área de estudio actualizado al año 2017. A partir de estas vías y ríos se calculó la distancia euclidiana de cada píxel a la vía más cercana.

Adicionalmente, A partir de las capas de pérdida de cobertura vegetal publicada por la Global Forest Watch (GFW) [189] para el año 2018, y la capa de Bosque y No Bosque publicada por el IDEAM en el Sistema de Información Ambiental de Colombia [190] para el año 2016, se calculó la distancia a la zona más cercana de pérdida de cobertura y al bosque más cercano. Así mismo, se descargó una capa de coberturas vegetales del Mapa de Ecosistemas Continentales, Costeros y Marinos publicado por el Instituto de Hidrología, Meteorología y Estudios Ambientales, Pronósticos y Alertas Cambio Climático (IDEAM) en el año 2017 [191]. De igual manera, se descargó el Mapa de clasificación de las tierras por su vocación de uso a escala 1:100.000 para el departamento de Antioquia disponible en el GEOPORTAL del Instituto Geográfico

Agustín Codazzi (IGAC), actualizado el 4 de marzo de 2019 [192]. El modelo digital de elevaciones fue descargado de la plataforma EarthExplorer del Servicio Geológico de los Estados Unidos [193], este modelo tiene información de altitud para todo el planeta a una resolución de 1 Arc-segundo (~30 m). A partir de este, se realizó un modelo hidrográfico para la zona de estudio, lo cual dio como resultado un esquema de orden Horton-Strahler para los cuerpos de agua de la zona [194], Los mapas y RASTER resultantes fueron utilizados como insumo en el mapa de resistencias y conectividad ecológica que será descrito más adelante, igualmente fueron utilizados como información de entrenamiento.

Posteriormente, se obtuvieron imágenes satelitales tomadas por el satélite Landsat 8, y descargadas por medio de la plataforma de libre acceso: Google Earth Engine [195]. Previo a la descarga se realizó una imagen compuesta con las imágenes tomadas entre los años 2014 y 2018 restringiendo la cantidad máxima de nubes por píxel de 5% por medio del algoritmo simple composite disponible en la librería de algoritmos Landsat. A partir de la imagen satelital compuesta, se descargaron las 11 bandas multiespectrales y se calcularon los siguientes índices: Diferencia de Vegetación Normalizada (NDVI), Diferencia de Vegetación Verde Normalizada (GNDVI), Vegetación Mejorada (EVI), Vegetación Avanzada (AVI), Vegetación Ajustada Al Suelo (SAVI), Diferencia de Humedad Normalizada (NDMI), Estrés Hídrico (MSI), Cobertura Verde (GCI), Suelo Desnudo (BI) y Tasa de Quema Normalizada (NBRI) En la Tabla 2 se observan las capas descargadas, la fuente de la información, la resolución espacial y las variables calculadas a partir de esta. Estos índices fueron calculados con el objetivo de obtener la mayor información posible de la vegetación en el área, así como de diferentes patrones que, aunque pueden ser invisibles para nosotros, el algoritmo puede detectarlos y hacer uso de ellos.

Adicionalmente, se debe destacar que aunque existen características que pareciesen redundantes desde el punto de vista de la percepción remota (e.g. NDVI, GNDVI, SAVI, entre otros) es importante que estén en la matriz de características, para que de esta manera, sea el algoritmo de selección el que decida cuales índices son realmente relevantes para el ajuste del modelo. Por último, con el propósito de modelar el movimiento animal, se generó un modelo de conectividad estructural para el área de estudio siguiendo la metodología expuesta por Isaacs-Cubides, et al. [196]. Este modelo fue realizado por medio de la herramienta Linkage Mapper, [197]. Para realizarlo, fue necesario realizar una reclasificación de capas de acuerdo con la etología de la especie diana: zorro perro (*Cerdocyon thous*), la cual fue seleccionada debido a su alta frecuencia de atropellamiento en la zona de estudio, así como su presencia en la zona y

Tabla 2. Conjunto de capas descargadas y variables calculadas a partir de estas. Fuente: autoría propia

Capa de información	Fuente de descarga	Resolución espacial	Variables calculadas
Capa de vías 2017 Capa de ríos 2017	DANE [188]		distancia a vías, distancia a ríos
DEM	USGS [193]		orden Horton-Strahler
Capa de pérdida de cobertura vegetal 2018	Global Forest Watch (GFW) [189]		distancia a zonas con pérdida de cobertura
Capa de bosque/no-bosque 2016	IDEAM [190]		distancia a bosque
Capa de coberturas vegetales 2017	IDEAM [191]		coberturas vegetales
Capa de clasificación de las tierras por su vocación	IGAC [192]	30 m/px	vocación de suelos
Imagen Landsat 8	Google Earth Engine [195].		NDVI GNDVI EVI AVI SAVI NDMI MSI GCI BI NBRI

los requerimientos de campeo que esta especie necesita, siendo esta una especie carnívora y generalista. El zorro perro es el cánido silvestre con mayor rango de distribución en Suramérica, encontrándose en altitudes desde el nivel del mar hasta los 3000 m s.n.m., se encuentra principalmente en bosques montanos y tropicales, sabanas, humedales, áreas ganaderas, cultivos y, adicionalmente, se ha evidenciado que cada vez se adapta más a las áreas intervenidas [198], esta especie aprovecha los recursos alimenticios disponibles: pequeños mamíferos, ranas, reptiles, aves e insectos, adicionalmente, consume huevos, carroña y semillas y frutos [198].

Construcción de matriz de características

A partir de lo anterior, se le añadió a cada segmento de vía el valor asociado a cada variable espacial, las cuales fueron usadas como características para la fase de entrenamiento. Para realizar lo anterior fue necesario convertir los shapefiles de las características a capas RASTER, donde cada píxel de la imagen contiene la magnitud de la variable de medición. Dicha base de datos contiene solo información espacial, de modo que pueda ser recopilada en otro segmento de vía que no posea datos de atropellamiento, permitiendo realizar una predicción binaria a partir de características propias de la zona y no de la distribución de puntos. Para esto, se asoció el valor promedio de cada característica al interior de diferentes buffers de distancia: 90 m, 150 m y 300 m, por medio de la herramienta "Zonal statistics as table" disponible en ArcMap [199]. De este proceso se generará una matriz de características a diferentes escalas de observación.

3.1.5. Selección de características

Con el objetivo de identificar las variables más relacionadas con el atropellamiento de fauna, se calculó el valor de diferentes métricas de relevancia por medio de los test estadísticos de Información mutua, Chi cuadrado y F score de ANOVA. Después de realizar este proceso, se procedió a calcular el Área Bajo la Curva (AUC) de la Respuesta Característica de Funcionamiento del Receptor (ROC) del algoritmo de Bosques Aleatorios (RF) ajustado con diferentes cantidades de características. Esto permite determinar la capacidad de un modelo de predicción no optimizado con diferentes características para encontrar el número de características ideal.

Es importante mencionar que durante la etapa de caracterización de un fenómeno, es usual ingresar al algoritmo de selección la mayor cantidad de características posibles, incluso si estas contienen información que a simple vista pareciera redundante, lo anterior debido a que el algoritmo de selección debe hallar la combinación ideal de características para obtener el mejor ajuste posible, encontrando patrones entre las variables, así como el descarte de las características que no le brindan información al modelo[200].

Considerando que la matriz de características que fue utilizada en este proyecto es una distribución uniforme de puntos con el valor de acumulación de cada segmento de vía y que, adicionalmente, los datos espaciales cercanos tienen una mayor

correlación que los datos lejanos [201], se diseñó un algoritmo de validación para evaluar la capacidad real de predicción del modelo por medio de validación cruzada por bloques. Esto se realizó por medio de la técnica StratfiedKfold disponible en la librería de Scikit-learn [156], generando 4 bloques o pliegues de entrenamiento y validación, cada bloque consta de 9 segmentos de entrenamiento y 3 de validación, dando una proporción de 75% - 25% de entrenamiento - validación, los cuales rotarán hasta que los 12 segmentos hayan sido al menos 1 vez parte del conjunto de prueba [200].

3.2. Modelo de predicción de atropellamiento

Previo a realizar el entrenamiento de los modelos, se debe tener en cuenta que existe un desbalance entre las clases, dado que la mayor parte de una vía consiste en puntos no calientes su proporción será mayor que los puntos calientes, de este modo, fue necesario evaluar estrategias de aprendizaje en datos desbalanceados [202]. Entre las técnicas empleadas en esta tesis se encuentran: muestreo adaptativo sintético (ADASYN) [161], técnica de sobre muestreo sintético minoritario (SMOTE) [163], así mismo se usaron variaciones de este como lo es el KMeans SMOTE [165], el Borderline SMOTE [166] y SVM SMOTE [166], en este último caso, fue necesario realizar una estimación de hiperparámetros para el SVM embebido en el algoritmo.

A partir de lo anterior, se procedió a implementar algoritmos de aprendizaje de máquina como: K- Vecinos más cercanos (KNN), Máquinas de Soporte Vectorial (SVM), Redes Neuronales artificiales (RNA) y Bosques Aleatorios (RF). Estos métodos fueron seleccionados por su uso en aplicaciones similares a la predicción de puntos calientes de atropellamiento de fauna [32, 203]. Para cada uno de estos modelos se calculó el AUC-ROC, y la matriz de confusión de cada uno de los clasificadores propuestos, así como el estadístico Kappa de cada uno de los resultados.

Con el objetivo de encontrar los hiper parámetros ideales de los clasificadores, se implementaron diferentes técnicas de optimización, en el caso del clasificador KNN se realizó una búsqueda extensiva del valor de *K* vecinos variando desde 1 hasta 300, así mismo, para el clasificador SVM se utilizó la función GridSearchCV disponible en Scikit-learn [156] para los valores de *C* y *Gamma* del Kernel basado en formas radiales (rbf) entre 0.0625 y 8 para *C*, y entre 0.0005 y 100 para el valor *Gamma*. Del mismo modo, se calcularon los hiper parámetros para las redes neuronales: la cantidad de neuronas por capa, la función de activación y el optimizador fueron hallados por medio de GridSearchCV. Por último, el algoritmo RF fue optimizado mediante Algoritmos

Genéticos (GA) a través de la librería TPOT [204], esta librería es de acceso libre está diseñada para obtener el clasificador ideal de tipo pipeline según unos parámetros de selección previamente establecidos por el usuario [205]. A partir de los resultados de cada clasificador optimizado y balanceado se escogió el algoritmo con el mejor ajuste para la fase de transferencia de aprendizaje por medio de la prueba de Friedman.

3.3. Transferencia de aprendizaje en vías sin muestreo

Inicialmente, es importante aclarar que, aunque existe una metodología popularmente conocida como transfer learning, en el contexto de visión artificial, la cual consiste en transferir una arquitectura de redes neuronal previamente entrenada con imágenes de un elemento diferente al que se desea clasificar o detectar [206], este no es el método que fue utilizado en esta investigación. En este caso, se debe entender el ejercicio transferencia de aprendizaje como el uso de un modelo de clasificación optimizado y ajustado con datos de una zona de entrenamiento y posteriormente reajustado con un porcentaje de datos correspondientes a los segmentos del tramo de vía sobre el cual se realizará la predicción, permitiendo así, que el grueso del aprendizaje se tome de zonas comparables, completamente muestreadas y que se realice una predicción a partir de un refinamiento con un mínimo de datos disponibles en una zona de validación diferente.

Para realizar la partición inicial de los datos para esta etapa se utilizó el algoritmo StratfiedKfold con el propósito particionar los segmentos de validación cruzada (Figura 10) en 4 bloques o pliegues de entrenamiento y transferencia con una proporción de 3:1 respectivamente, los cuales al final de cada evaluación rotarán hasta que cada segmento haya sido al menos 1 vez parte del conjunto de prueba [200]. A partir de esto, se validó la viabilidad de implementación de la metodología de transferencia de aprendizaje descrita anteriormente por medio del algoritmo seleccionado como con mejor ajuste al atropellamiento en una zona sin datos. Para esto, se probó diferentes porcentajes de datos agregados para reajustar el modelo, con el objetivo de determinar el porcentaje ideal necesario para realizar predicciones en zonas con pocos datos de atropellamiento.

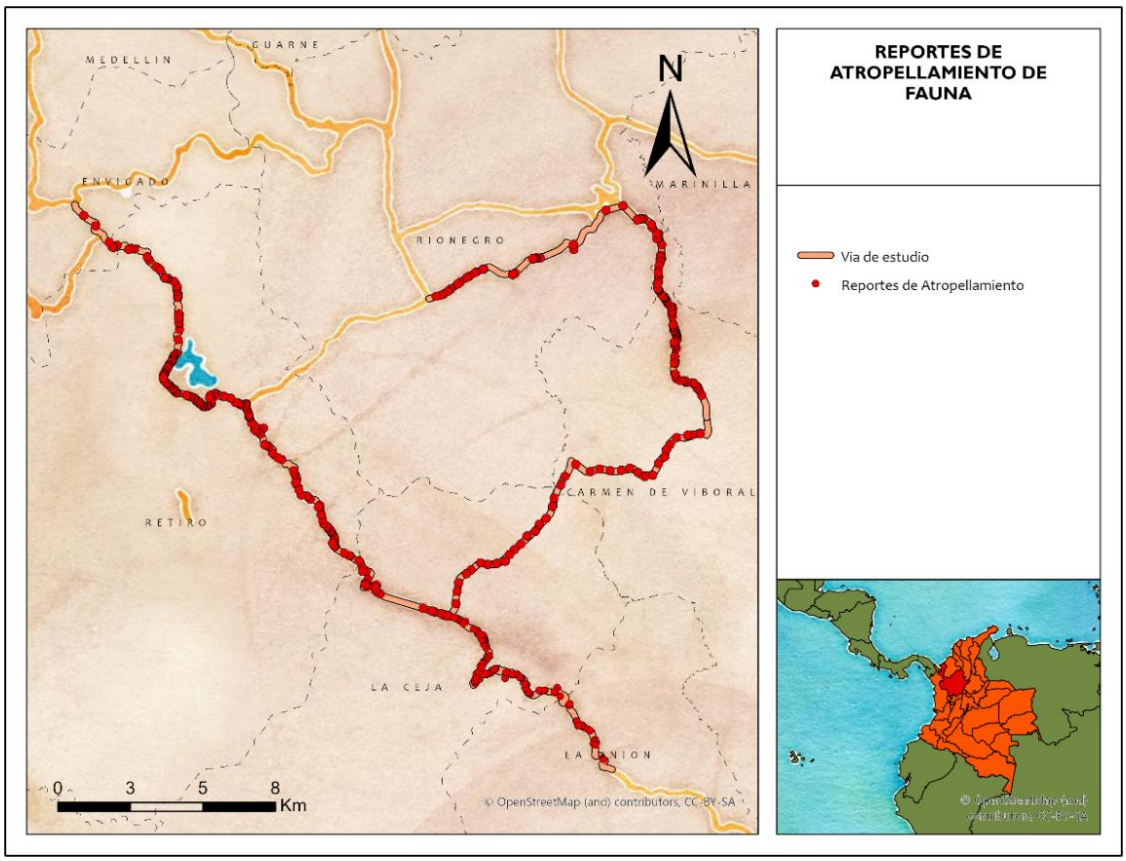
4. Resultados

4.1. Caracterización del fenómeno del Atropellamiento de Fauna Silvestre

4.1.1. Recolección de datos de atropellamiento de fauna silvestre

Al extraer la información contenida al interior de la App Recosfa, se logró recolectar 837 reportes de atropellamiento de fauna en el área de estudio, de los cuales 527 son Mamíferos, 178 Aves, 82 Anfibios, 47 Reptiles y 3 cuya clase no fue posible identificar. Así mismo, la familia *Didelphidae* fue el más reportado con 335 registros, seguido por la especie Rhinella *marina* (Sapo Gigante) y la *Notosciurus granatensis* (Ardilla Colirroja). Para esta investigación fue utilizada la distribución total de puntos, la cual puede observarse en la Figura 9.

Figura 9. Reportes de atropellamiento de fauna recolectados en el área de estudio – oriente de Antioquia, Colombia. Fuente: autoría propia



Es importante mencionar que 354 de los reportes contenidos en la App Recosfa fueron recolectados en el marco del proyecto: "Evaluación del impacto de la infraestructura vial sobre la mortalidad de vertebrados y posibles medidas para la conectividad ecológica del paisaje en el Valle de Aburrá" realizado por Juan Carlos Jaramillo Fayad, PhD, docente del Instituto Tecnológico Metropolitano. Adicionalmente, 399 reportes fueron recolectados en el marco de la tesis titulada: "Atropellamiento de Fauna Silvestre en el Área de Influencia de la Concesión DEVIMED – Oriente Antioqueño" desarrollada por el Biólogo Mateo Hernández y la Concesión DEVIMED, ambos casos tuvieron una metodología sistemática de muestreo. Los reportes restantes corresponden a datos de atropellamiento recolectados a conveniencia por personal de mantenimiento, así como de los usuarios de estas vías.

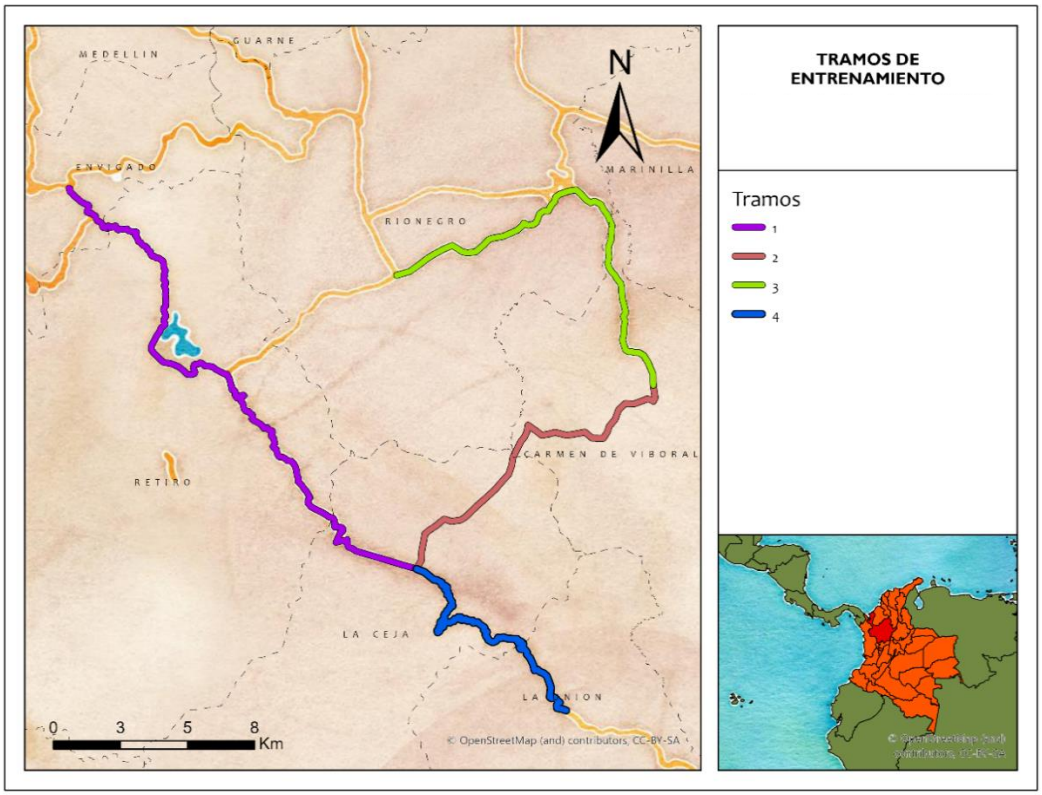
4.1.2. Análisis de distribución de puntos

Con el objetivo de realizar estos análisis, se evaluó la distribución de puntos en 4 tramos lineales de vías del área de estudio. Los tramos fueron seleccionados teniendo en cuenta la distancia y forma de la vía, prestando especial atención a evitar tramos de vía superpuestos, debido a que los análisis de K Ripley y puntos calientes 2D, basan su funcionamiento en la proyección de figuras circulares sobre la vía, las cuales, al aumentar su radio, pueden verse afectadas por este tipo de formas. Los tramos seleccionados pueden observarse en la Figura 10.

Al extraer los reportes de atropellamiento de fauna silvestre se encontraron 281 reportes correspondientes al tramo 1, 208 reportes en el tramo 2, 196 reportes en el tramo 3 y 152 en el tramo 4. Con respecto a las coberturas vegetales pudo observarse una predominancia de coberturas modificadas por el ser humano, así como una prevalencia de especies altamente generalistas y tolerantes a las actividades humanas.

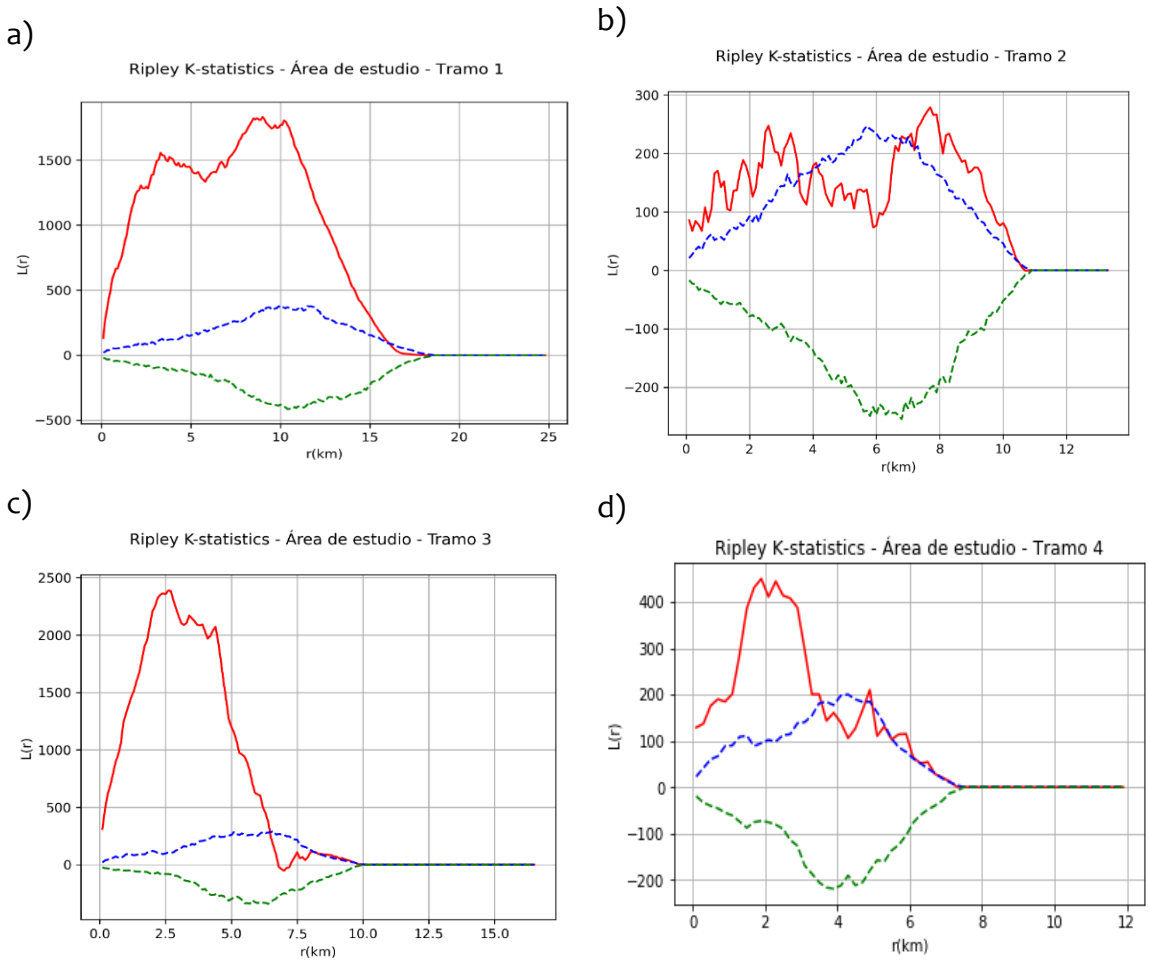
Adicionalmente, a partir del análisis K Ripley se identificaron bandas de distancia con agrupaciones significativas entre los 0.1 km(r) y 16 km(r) en el tramo 1, entre los 0.1 km(r) y 3.5 km(r) y entre las distancias 6.8 km(r) y 10.4 km(r) en el tramo 2, entre los 0.1 km(r) y 6.5 km (r) en el tramo 3, y entre los 0.1 km(r) y 3.7 km(r), los 5.1 km (r) y 5.3 km (r), 5.5 km (r) y 6.5 km (r) en el tramo 4, los cuales pueden ser observados en la Figura 11 generada a partir de la evaluación de los tramos en el software Siriema.

Figura 10. División de tramos para la realización de Análisis de patrones de puntos en el área de estudio - oriente de Antioquia, Colombia. Fuente: autoría propia



A partir de estas bandas de distancia se realizó el análisis de autocorrelación espacial, en el cual se identificó las bandas de distancia de 1.3 km y 269 m como bandas de distancia con patrones de agrupación significativos, con una autocorrelación espacial positiva débil (I = 0.18) para el tramo 1 y una autocorrelación espacial muy débil (I = 0.064) para el tramo 2, así mismo, se evidenció que en la banda de distancia de 1.3 km existen agrupaciones significativas con una autocorrelación espacial moderada (I = 0.47, I = 0.21) para los tramos 3 y 4, respectivamente.

Figura 11. Gráfica producto del análisis K Ripley para los tramos 1 (a), 2 (b), 3 (c) y 4 (d) del área de estudio - oriente de Antioquia, Colombia. En rojo se observa el valor de la función L(r), azul y verde corresponden a los límites de confianza superior e inferior, respectivamente. Fuente: autoría propia



Por último, se procedió a identificar los puntos calientes de atropellamiento de fauna con las bandas de distancia seleccionadas. En el tramo 1, se identificaron patrones de agrupación significativos entre los kilómetros 7.5 y 12.65, 13.5, 14, y entre los kilómetros 17.5 y 18.9, partiendo desde el inicio del tramo 1 (Glorieta de Las Palmas). En el tramo 2 fueron identificados patrones de agrupación significativos entre los kilómetros 0.5 a 0.6, 2.1 al 3.6 y kilómetros 5.6, 7.6, 8 y 11 partiendo desde el inicio del tramo 2 (La Ceja). Para el tramo 3 los kilómetros 2 y 7 como puntos de agrupación significativos partiendo desde el inicio del Tramo 3 (El Carmen de Viboral). Por último, en el tramo 4 se identificó que existen agrupaciones de puntos significativas entre los

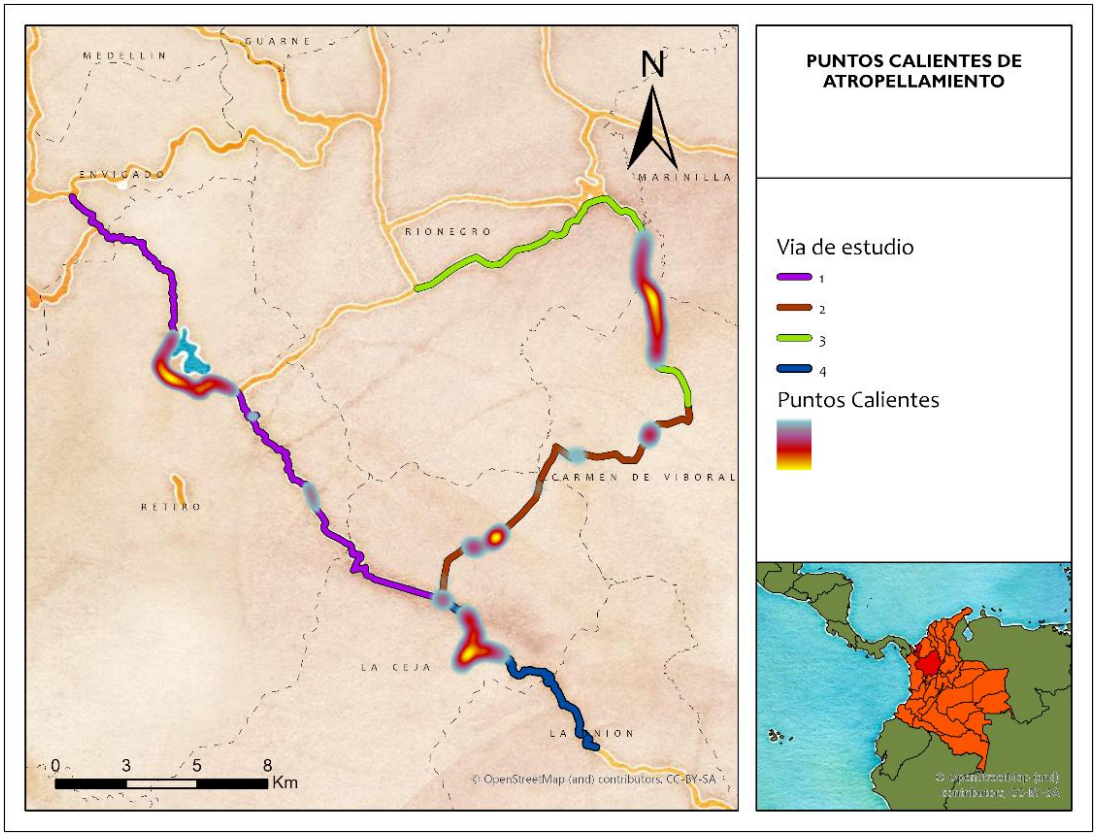
kilómetros 0 y 0.5, y entre los kilómetros 1 y 6, partiendo desde el inicio del Tramo 4 (Municipio de la Ceja). Los puntos calientes identificados pueden observarse en la Figura 12.

4.1.3. Extracción de características

Recolección de Mapas e Imágenes satelitales.

Con el propósito de identificar las variables más relacionadas con el atropellamiento se procedió a descargar mapas de la zona de estudio de las características ambientales propuestas.

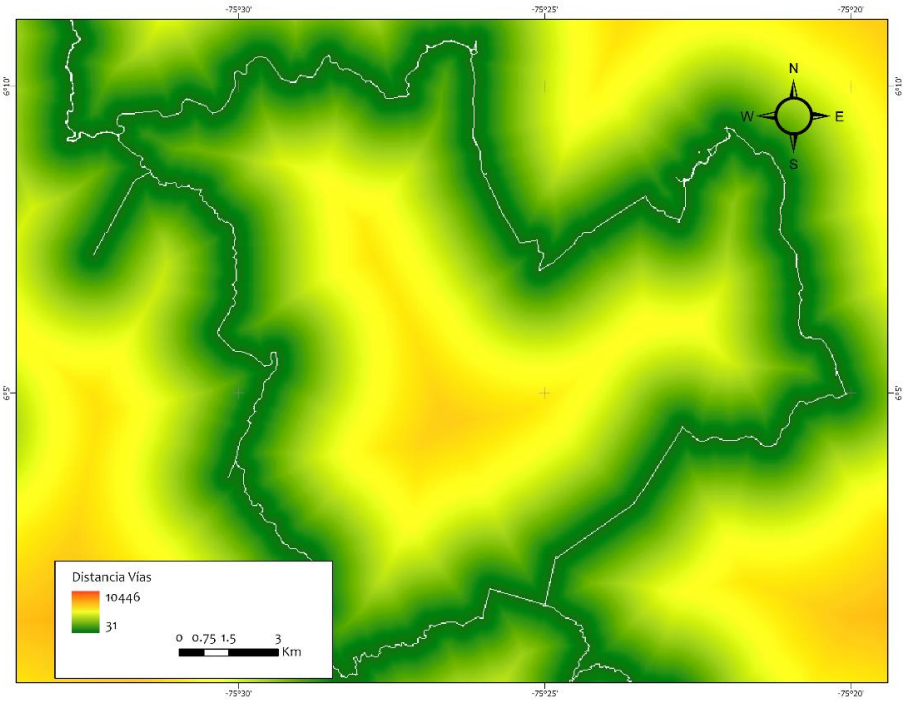
Figura 12. Mapa de puntos calientes en el área de estudio – oriente de Antioquia, Colombia. Fuente: autoría propia



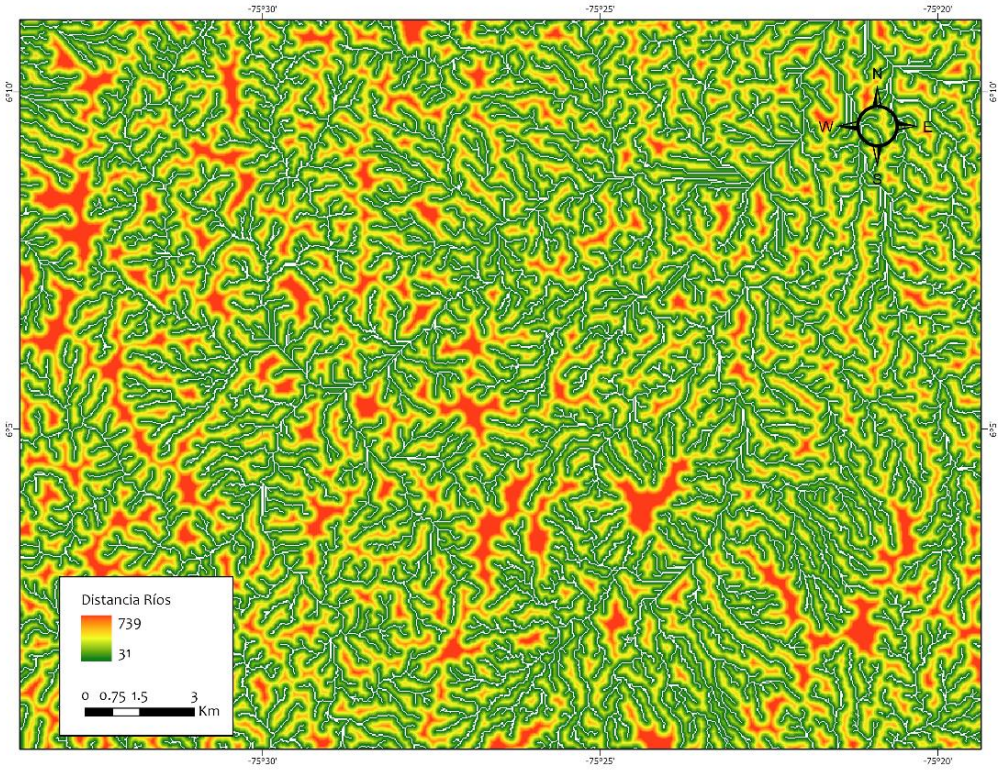
A partir de las capas descargadas se realizó un modelo de conectividad para la zona de estudio, para esto fue necesario realizar una reclasificación de los mapas descargados de acuerdo con el comportamiento de la especie diana: *Cerdocyon thous*. Adicionalmente, una capa de áreas protegidas fue descargada del repositorio digital de Parques Nacionales Naturales de Colombia [207] y fue utilizada como capa de núcleos. Estos núcleos actuarán como fuentes de corriente, la cual usará un mapa de resistencias como circuito, cada uno de los píxeles del mapa actuará como un nodo de corriente en el cual se hallará el camino con menor pérdida de voltaje. En el Anexo 2 (Sección 7.2) puede observarse la tabla de reclasificación necesaria para realizar el modelo. Como producto de esta sección, en la Figura 13 se presentan los mapas resultantes:

Figura 13. Mapas resultantes de la recolección de mapas para el área de estudio – oriente de Antioquia, Colombia: (A) Mapa de distancia a vías, (B) Mapa de distancia a Ríos, (C) Distancia a Pérdidas de Cobertura, (D) Distancia a Bosques, (E) Coberturas de suelo, (F) Vocación de uso de suelo, (G) Modelo Digital de Elevación, usado como elevación, (H) Modelo de cuencas hidrográficas, (I) Imagen compuesta de satélite RGB, (J) Modelo de resistencias para el *Cerdocyon thous*, (K) Modelo de Conectividad Ecológica. Fuente: autoría propia

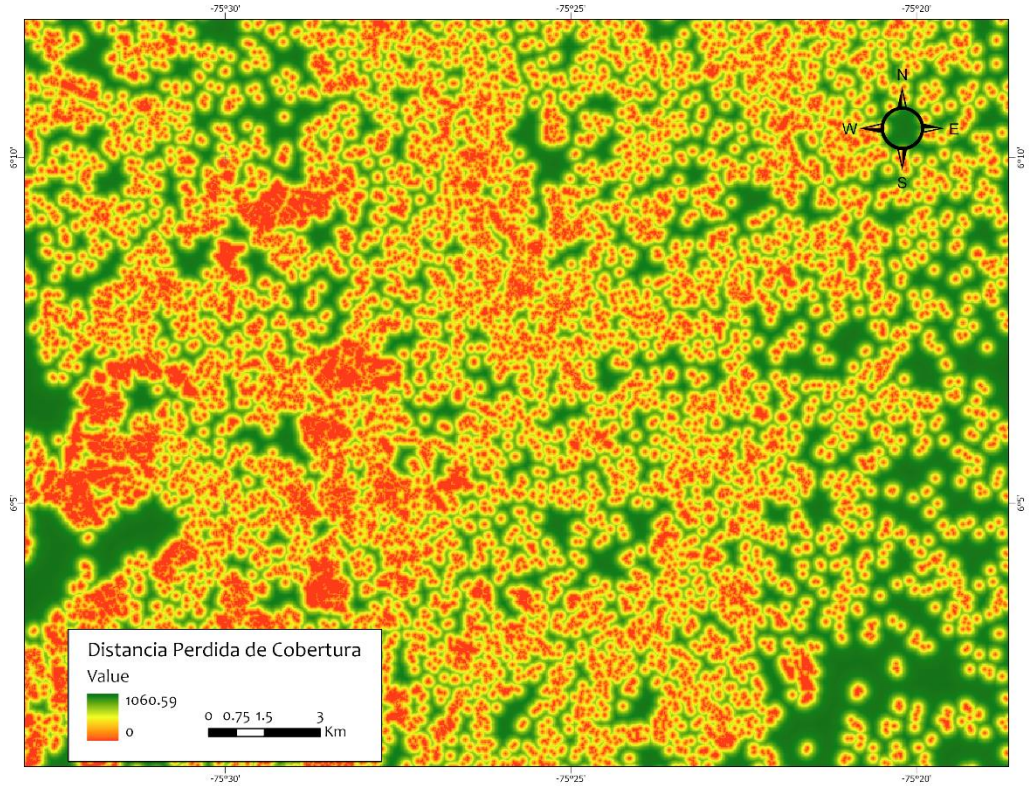
(A)



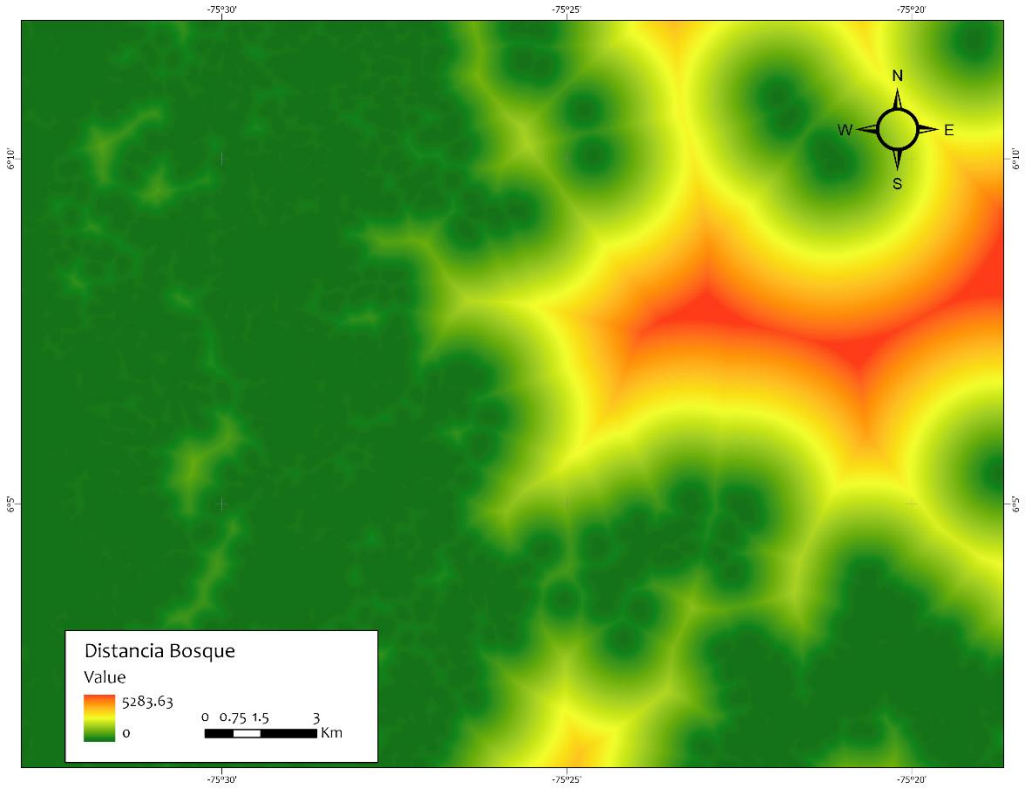
(B)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Distancia Ríos
739
31
0 0.75 1.5 3
Km

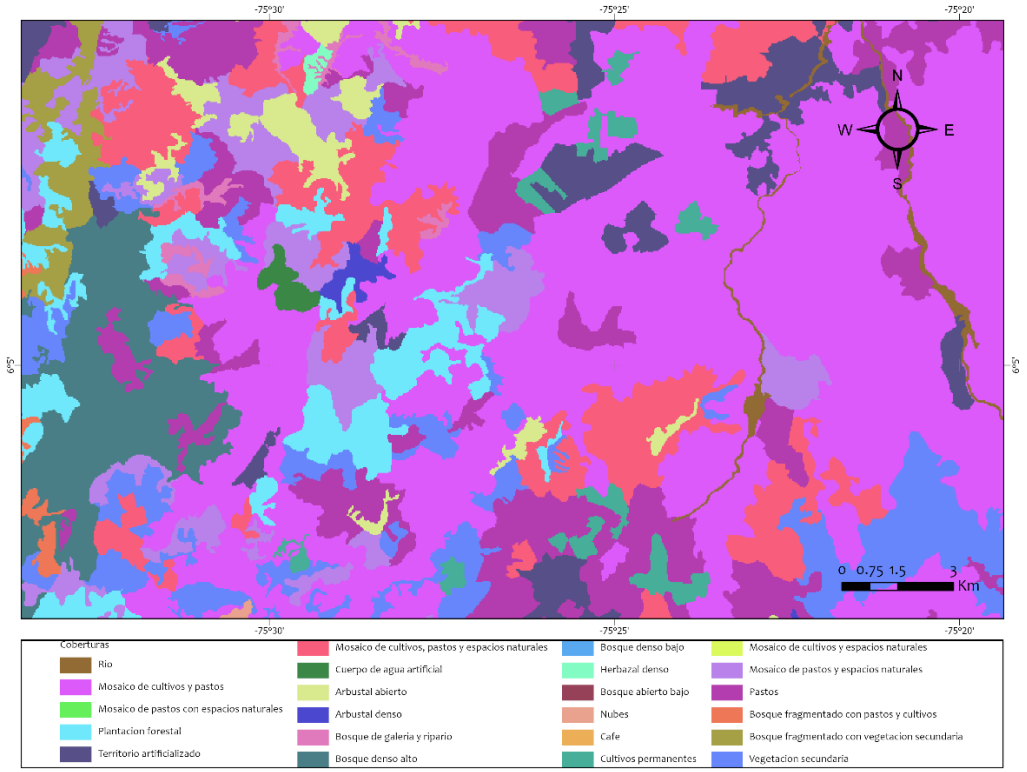
(C)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Distancia Perdida de Cobertura
Value
1060.59
0
0 0.75 1.5 3
Km

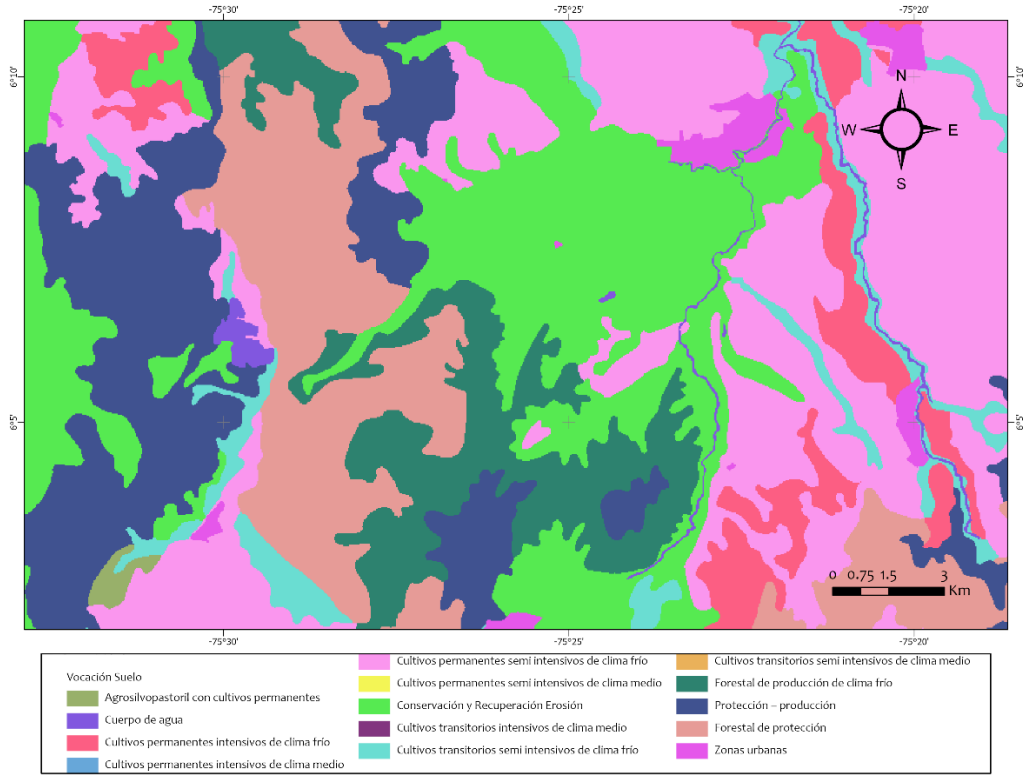
(D)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Distancia Bosque
Value
5283.63
0
0 0.75 1.5 3
Km

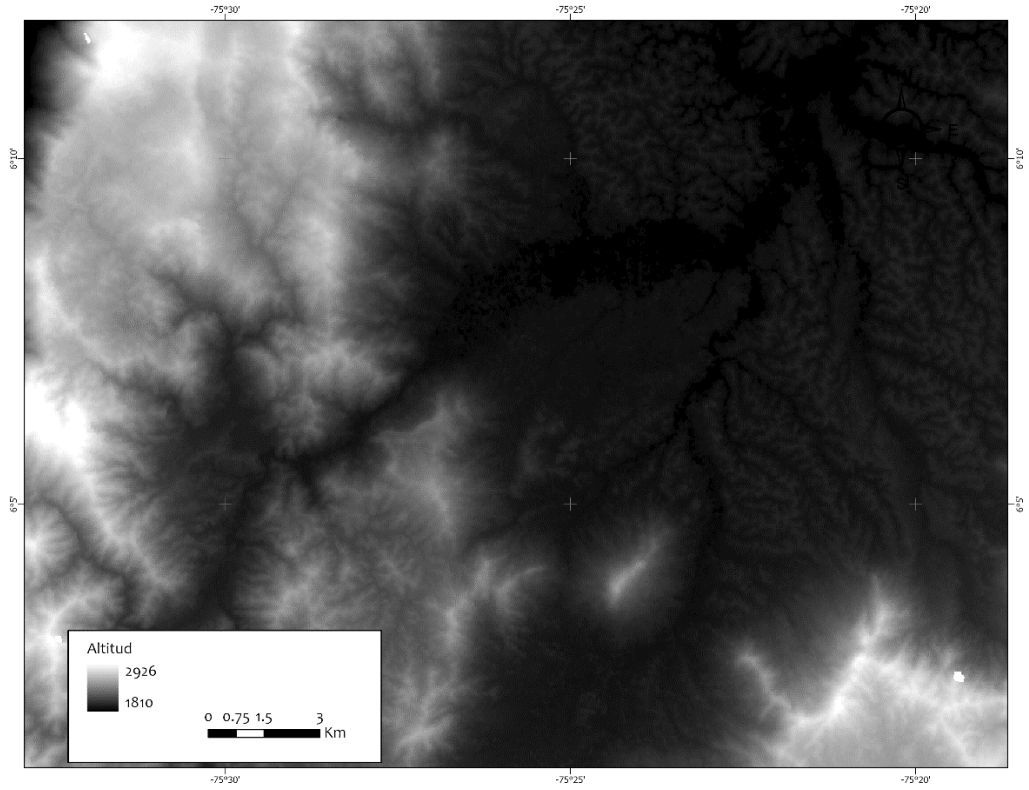
(E)


-75°30'
-75°25'
-75°20'
6°5'
N
W
E
S
0 0.75 1.5 3
Km
Coberturas
Rio
Mosaico de cultivos y pastos
Mosaico de pastos con espacios naturales
Plantacion forestal
Territorio artificializado
Mosaico de cultivos, pastos y espacios naturales
Cuerpo de agua artificial
Arbustal abierto
Arbustal denso
Bosque de galeria y ripario
Bosque denso alto
Bosque denso bajo
Herbazal denso
Bosque abierto bajo
Nubes
Cafe
Cultivos permanentes
Mosaico de cultivos y espacios naturales
Mosaico de pastos y espacios naturales
Pastos
Bosque fragmentado con pastos y cultivos
Bosque fragmentado con vegetacion secundaria
Vegetacion secundaria

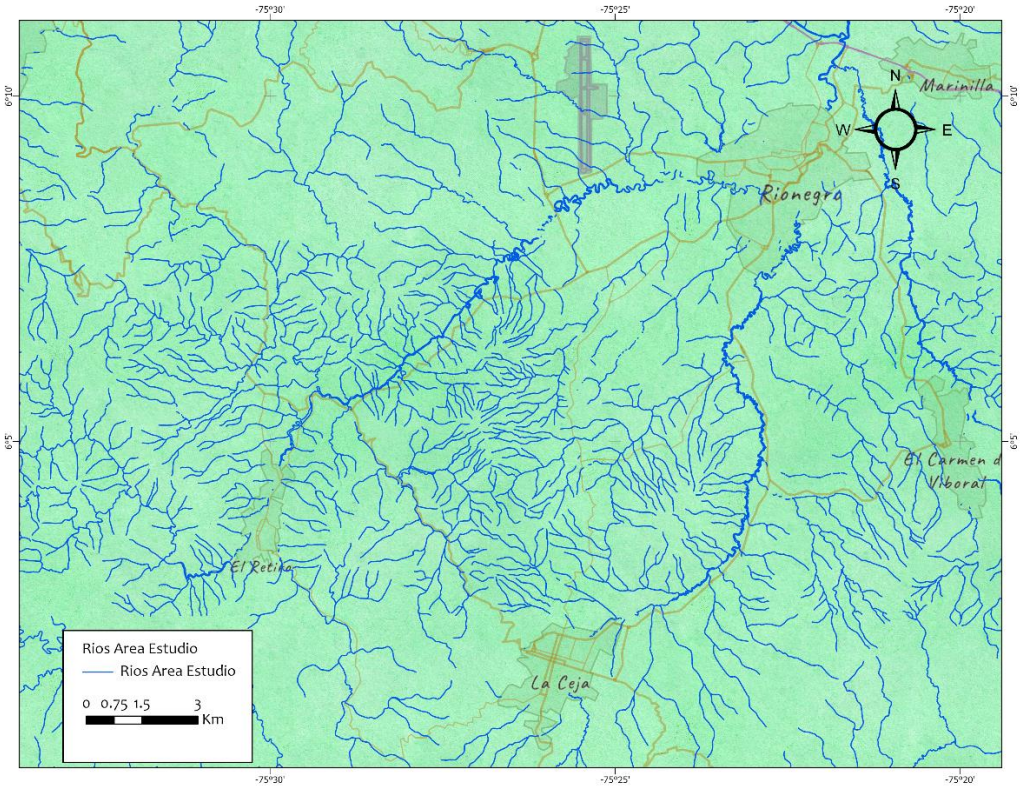
(F)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
0 0.75 1.5 3
Km
Vocación Suelo
Agrosilvopastoril con cultivos permanentes
Cuerpo de agua
Cultivos permanentes intensivos de clima frío
Cultivos permanentes intensivos de clima medio
Cultivos permanentes semi intensivos de clima frío
Cultivos permanentes semi intensivos de clima medio
Conservación y Recuperación Erosión
Cultivos transitorios intensivos de clima medio
Cultivos transitorios semi intensivos de clima frío
Cultivos transitorios semi intensivos de clima medio
Forestal de producción de clima frío
Protección – producción
Forestal de protección
Zonas urbanas

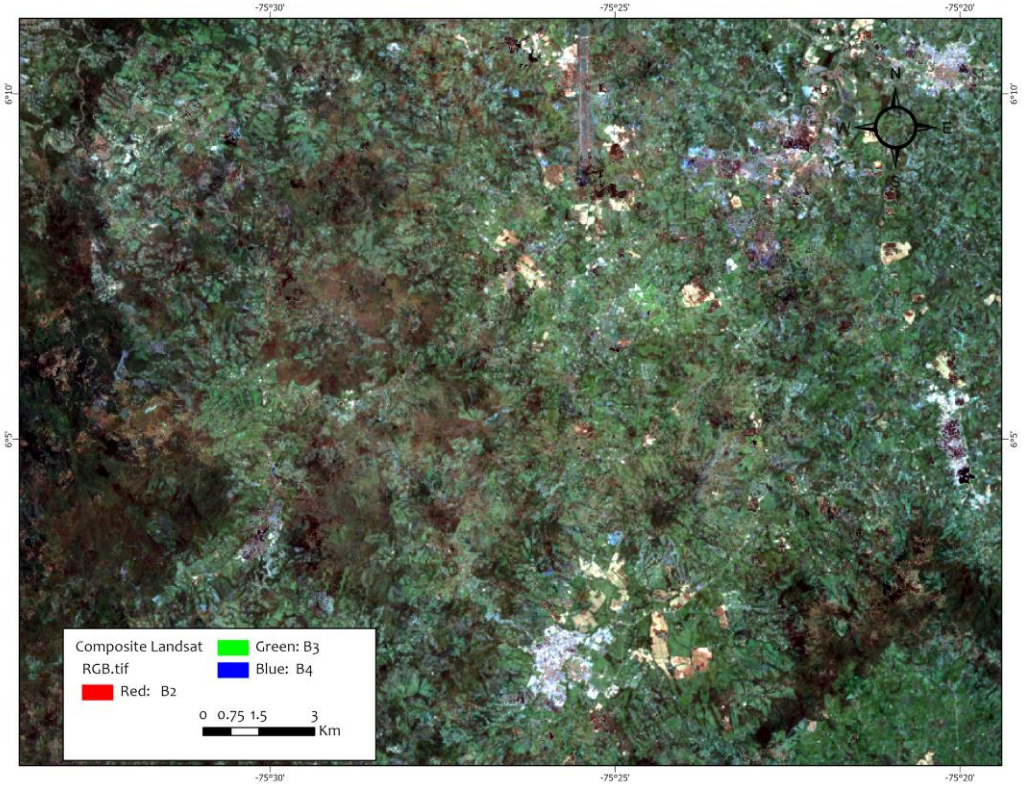
(G)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
Altitud
2926
1810
0 0.75 1.5 3
Km

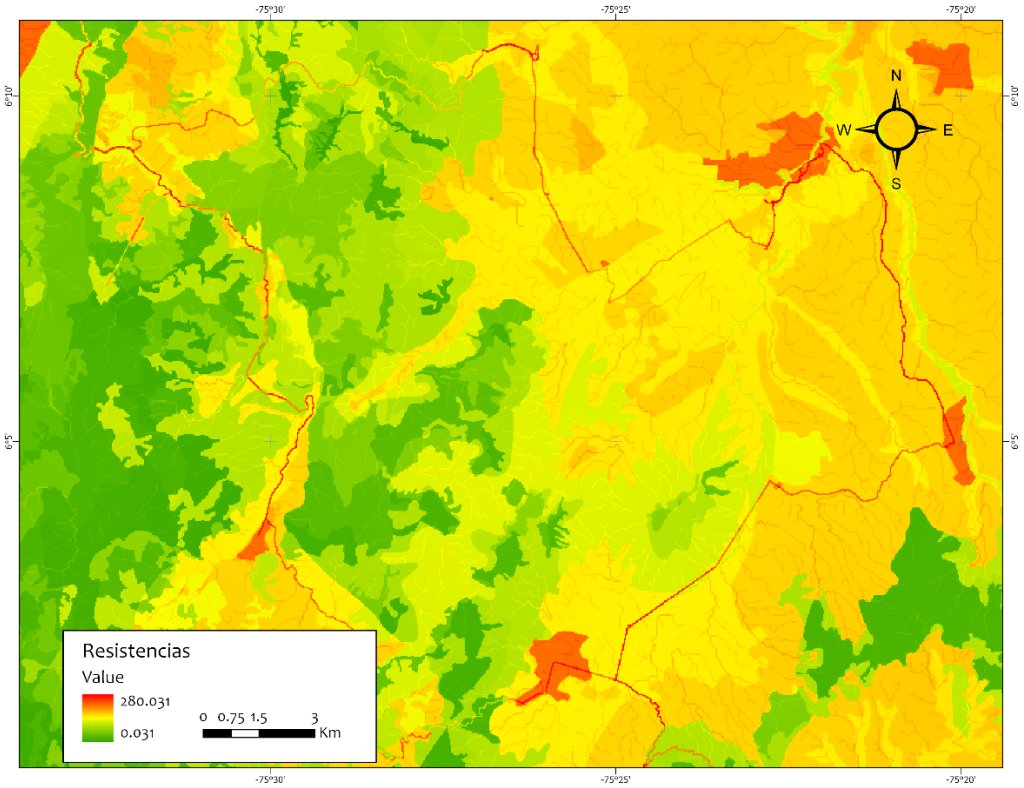
(H)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Marinilla
Rionegro
El Retiro
La Ceja
Rios Area Estudio
Rios Area Estudio
0 0.75 1.5 3
Km

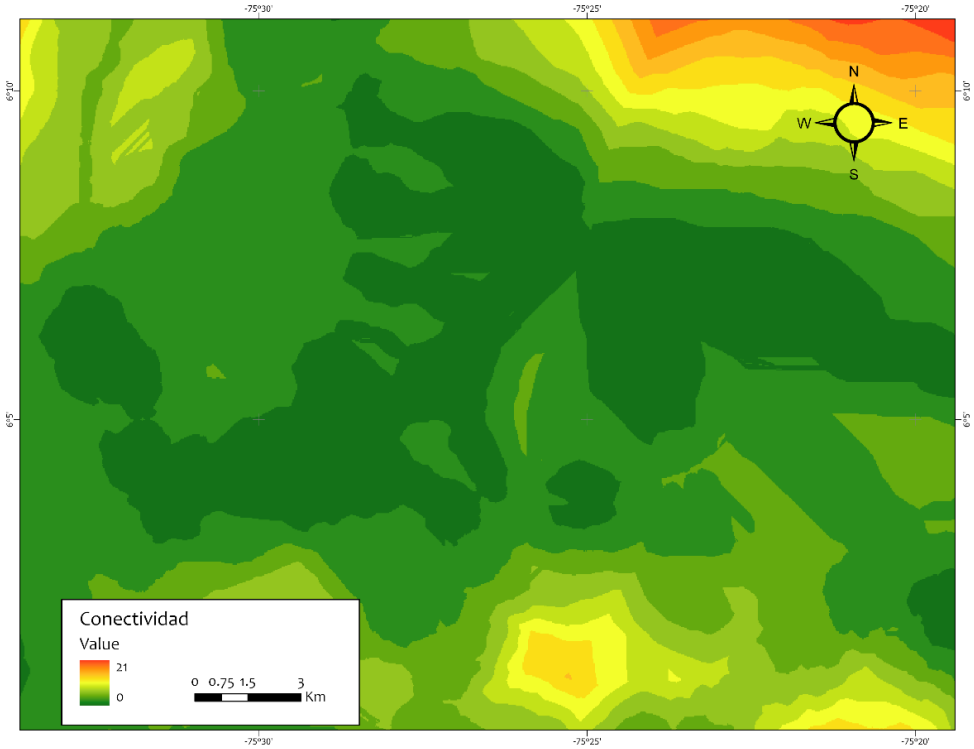
(I)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Composite Landsat
RGB.tif
Red: B2
Green: B3
Blue: B4
0 0.75 1.5 3
Km

(J)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Resistencias
Value
280.031
0.031
0 0.75 1.5 3
Km

(K)


-75°30'
-75°25'
-75°20'
6°10'
6°5'
N
W
E
S
Conectividad
Value
21
0
0 0.75 1.5 3
Km

Construcción de matriz de características

A partir de los índices, las imágenes satelitales y los mapas descritos anteriormente, se procedió a agregar las características a cada uno de los segmentos generados en la sección 4.1.2: Análisis de distribución de puntos. Las características fueron extraídas a diferentes distancias: 90 m, 150 m y 300 m, asociando el valor promedio de cada variable al interior del buffer, para esto se utilizó la herramienta "Zonal statistics as table"

De este proceso, se generó una matriz de características de 3000 segmentos de vía correspondiente a las filas y 96 columnas, cada una correspondiente a las características: Latitud, Longitud, HS, UCL, LCL, Pérdida de Cobertura, Ganancia de Cobertura, Porcentaje de Cobertura Boscosa, banda 1 (aerosoles 0.433 - 0.453 µm), banda 2 (azul visible 0.450 - 0.515 µm), banda 3 (verde visible 0.525 - 0.600 µm), banda 4 (rojo visible 0.630 - 0.680 µm), banda 5 (infrarrojo cercano 0.845 - 0.885 µm), banda 6 (infrarrojo de onda corta 1.56 - 1.66 µm SWIR 1), banda 7 (infrarrojo de onda corta 2.10 - 2.30 µm SWIR 2), banda 8 (pancromática 0.50 - 0.68 µm), banda 9 (nubes grises 1.36 - 1.39 µm), banda 10 (infrarrojo térmico 10.60 - 11.19 µm – TIRS 1), banda 11 (infrarrojo térmico 11.50 - 12.51 µm – TIRS 2) del satélite Landsat 8,los índices multiespectrales: NDWI, BSI, NBRI, GCI, MSI, NDMI, SAVI, EVI, GNDVI, AVI, NDVI, Altitud, Distancia a ríos, Distancia a zonas con pérdida de cobertura, Resistencia, Costo de Movimiento, Distancia a bosque, Distancia a corredor biológico, cada una evaluada a las escalas espaciales de 90, 150 y 300 m.

Para el propósito de este estudio las columnas: Latitud, Longitud, HS, UCL, LCL, solo serán usadas como guía para la ubicación de cada segmento y su respectivo etiquetamiento, más no serán usadas como características de entrenamiento, puesto que solo nos interesan las características ambientales que puedan predecir el fenómeno sin una distribución de puntos. Es importante mencionar que la matriz de características presenta desbalance de clases, lo cual debe ser tenido a consideración en las siguientes fases de trabajo, requiriendo aplicar técnicas que permitan no solo balancear de manera sintética la matriz de características, sino también, medir de manera adecuada el desempeño de los métodos propuestos [208, 209].

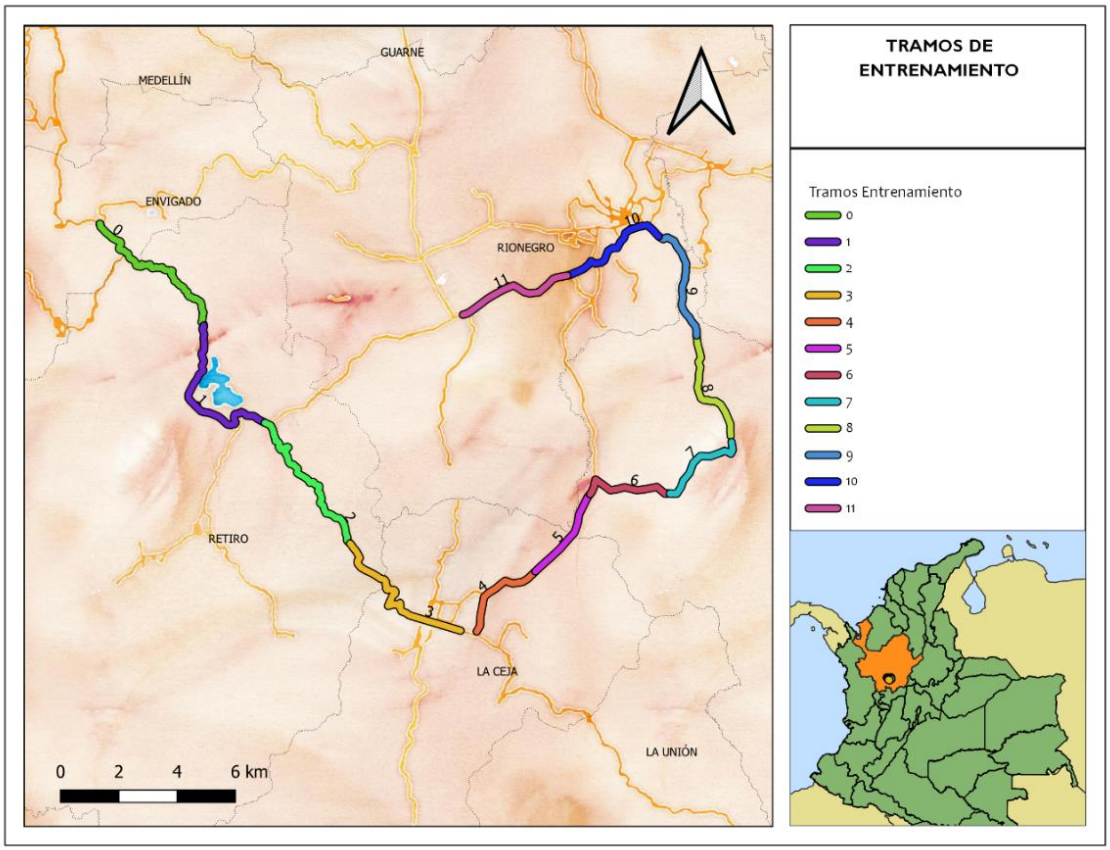
4.1.4. Selección de características

Como primera parte de la selección de características, fue necesario realizar una normalización de la base de datos, esto a raíz de que los valores de las diferentes características contenidas en la matriz tienen diferentes rangos, para lo cual, se buscó que a partir de este proceso, todas las características fueran comparables [210]. Posteriormente, se procedió a realizar la selección de características usando el método de selección univariante de las k mejores características utilizando como criterio de información Chi-cuadrado (χ^2) [211], Información Mutua (MI) [212], y el valor F de ANOVA (F-score) [213], los cuales están disponibles en la herramienta SelectKBest de Scikit-learn [156].

Con el objetivo de seleccionar el número de características que permitan obtener el mejor resultado de clasificación, se evaluó el área bajo la curva ROC del clasificador RF. Este clasificador fue elegido debido a que por su funcionamiento realiza una selección automática de hiper parámetros para una clasificación inicial [214]. Debido a la dependencia espacial entre los datos recolectados [201], se diseñó un algoritmo de validación en 12 segmentos, cada uno con 250 puntos distribuidos de manera secuencial en el área de entrenamiento, Eso permite evaluar la capacidad real de predicción del modelo empleando diferentes segmentos. En la Figura 14 se puede observar el mapa del área de entrenamiento con sus respectivos segmentos.

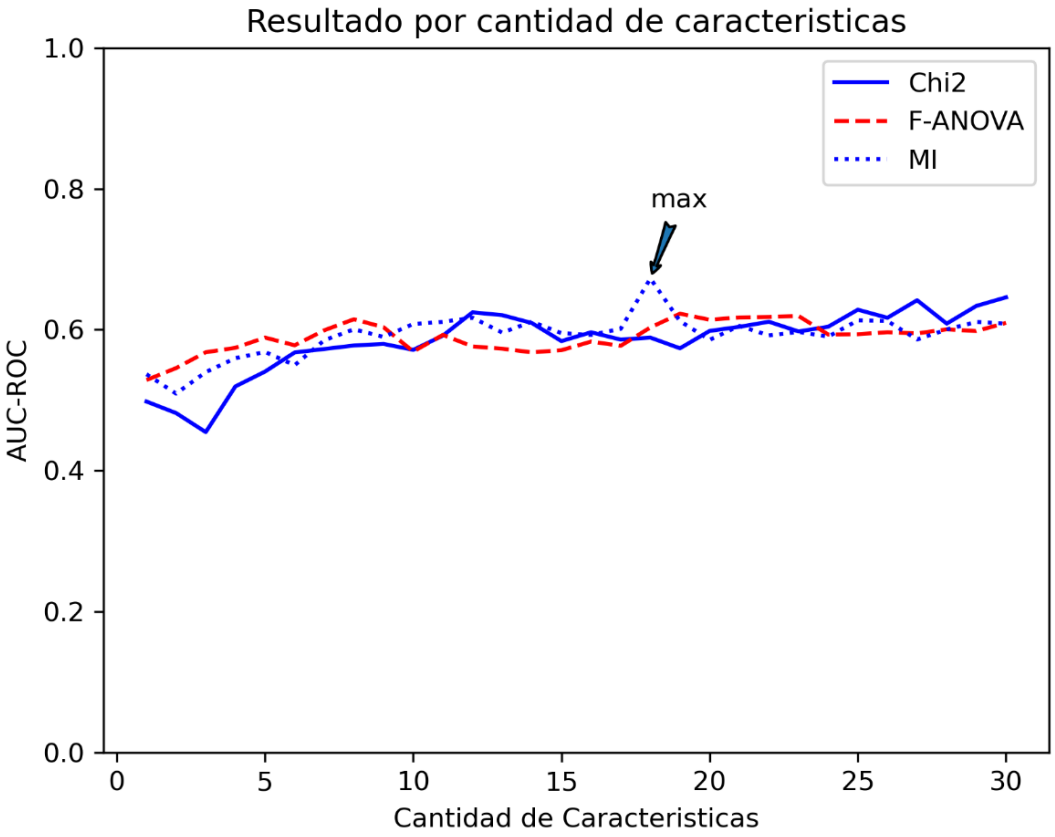
Con el objetivo de medir el desempeño del modelo con las características seleccionadas, se realizó una validación con 9 tramos de entrenamiento y 3 para validación, cambiando en cada iteración los tramos de validación hasta alcanzar 4 validaciones de 3 tramos, para un total de 12 validaciones. Al finalizar se presenta una gráfica en la cual se muestra el AUC-ROC medio del clasificador RF al ser entrenado con k características, variando este valor entre 1 y 30 características. En la Figura 15 se puede observar que la cantidad de características ideal, así como el método de selección que brinda una mayor generalización del fenómeno del atropellamiento de fauna, fueron 17 características con el método de selección de Información múltiple (MI), alcanzando un valor de AUC-ROC promedio entre las validaciones cruzadas de 0.672564

Figura 14. División de la zona de entrenamiento en segmentos de entrenamiento. Área de estudio – oriente de Antioquia, Colombia. Fuente: autoría propia



Las características seleccionadas pueden observarse en la Tabla 3, en la cual se listan con su respectivo valor de MI asociado y porcentaje de contribución en la clasificación. Es importante mencionar que el clasificador RF fue entrenado con la opción de balance de clases activa, permitiendo que el método se ajuste de forma inversamente proporcional a la frecuencia de aparición de cada clase, logrando así darle mayor ponderación a la clase minoritaria, lo anterior debido a que la proporción de la clase 1 (Punto Caliente) en comparación de la clase 0 (Punto no caliente) es de 1:4.

Figura 15. Área bajo la curva promedio de la Respuesta Característica de Funcionamiento del Receptor (mean AUC-ROC) del clasificador Bosques Aleatorios (RF) según el número de características. Fuente: autoría propia



4.1.5. Análisis de resultados y discusión

Esta sección de discusión tiene como propósito analizar y comparar los resultados obtenidos en la fase de caracterización del fenómeno del atropellamiento de fauna con respecto a otras investigaciones publicadas.

Tabla 3. Conjunto de características seleccionadas para la etapa de Entrenamiento y Transferencia de aprendizaje. Fuente: autoría propia

Característica	bits	Contribución
band_11_300m	0.22837	0.073298
Altitud_300m	0.210617	0.066922
Distancia a Bosque_300m	0.197259	0.067143
band_10_300m	0.19518	0.045025
Distancia a corredor biológico_300m	0.174479	0.095792
Resistencia_150m	0.172832	0.030807
Resistencia_300m	0.169023	0.043452
Altitud_150m	0.167544	0.074664
Distancia a Bosque_150m	0.16754	0.055238
Distancia a corredor biológico_150m	0.159924	0.074783
Distancia a corredor biológico_90m	0.152511	0.070427
Distancia a Bosque_90m	0.152351	0.059134
band_9_300m	0.150367	0.053969
band_11_150m	0.146185	0.044474
NBRI_300m	0.144953	0.042359
Altitud_90m	0.128334	0.03941
Costo de Movimiento_300m	0.122733	0.063102

Análisis de distribución de puntos

A partir de la recolección de datos de atropellamiento, se logró obtener una cantidad suficiente de datos para el proceso de análisis de distribución de puntos, estos concuerda con lo reportado por [183], siendo estos datos correspondientes a Colombia, donde se identificó a la zarigüeya (*Didelphis marsupialis*) como el animal más

atropellado, seguido por la ardilla colirroja (*Notosciurus granatensis*), los cuales también son identificados por este estudio como 2 especies altamente afectadas por el atropellamiento de fauna. Así mismo, concuerda con lo reportado por [20, 21] quien ha identificado a la zarigüeya como el animal más atropellado en las vías de envigado que colindan con el área de estudio.

Con respecto a la distribución de puntos se identificó que el tramo 1 es el segmento vial con mayor cantidad de reportes: 281, seguido por el tramo 2 con 208, el tramo 3 con 196 reportes y el tramo 4 con 152 reportes. Esto puede explicarse por la cercanía del tramo 1 y 4 con la Reserva del Río Nare, el DRMI San Miguel y el DRMI Cerros de San Nicolás, principales nodos ecológicos de la zona. En cambio, el tramo 3 tiene una mayor distancia a áreas protegidas y bosques, disminuyendo el impacto de las carreteras sobre estos ecosistemas protegidos [5, 12]. Aunque la longitud de los tramos es distinta se garantizó comparabilidad entre ellos al realizar la misma cantidad de divisiones para el cálculo de la intensidad de agrupación de puntos calientes H(s).

Luego de aplicar el análisis de patrones de puntos K Ripley a todos los tramos de atropellamiento, así como el análisis de autocorrelación espacial, fue posible determinar que el fenómeno de atropellamiento de fauna en el área de estudio no se encuentra mediada por el azar, tal como fue demostrado por Clevenger, et al. [104]. Se logró determinar que los puntos calientes de atropellamiento de fauna se encuentran relacionados con sus vecinos en términos de acumulación de puntos, siendo más similares con sus vecinos que con puntos de agregación distantes, tal como fue descrito por Chun y Griffith [137], [215, 216]. Por lo tanto, se puede decir que los puntos calientes identificados en este capítulo corresponden a puntos de agrupación significativos, no influidos por el azar y cuyo comportamiento tiene una autocorrelación espacial significativa, permitiendo identificarlos como puntos de especial atención para la autoridad ambiental y los administradores de las vías que cruzan por el área de estudio debido al cruce permanente de fauna por ellas [12].

Extracción de Características

Con respecto a la extracción de características, se utilizó la metodología expuesta por Amiri, et al. [30], Ghorbani, et al. [35], Jaafari, et al. [37], Ngoc Thach, et al. [38], Wang, et al. [42], Kantola, et al. [90], entre otros, que han utilizado variables espaciales como insumo para la predicción de diferentes fenómenos espaciales [199]. Adicionalmente, se buscó ampliar la base de datos de características al usar las bandas

espectrales del satélite Landsat, con el objetivo de obtener la mayor cantidad de información posible acerca de la cobertura vegetal, similar al uso que realizó Ascensão, et al. [113], con el índice NDVI e información de coberturas, trabajo que permitió tener una base a partir de la cual se propusieron algunos de los bloques de características presentados en este trabajo. Este trabajo tiene diferencias con lo reportado en la literatura en términos de las características planteadas, especialmente en que se integra la información acerca de las coberturas por medio de las bandas e índices espectrales, evitando usar variables categóricas. Las características planteadas en este proyecto toman como base los trabajos previos presentados por Kantola, et al. [90], Gonçalves, et al. [93], Ascensão, et al. [113], Fabrizio, et al. [128], Ha y Shilling [129], entre otros.

Selección de Características

A partir del conjunto de características antes descrito, se propuso un flujo de trabajo para la selección de las características que describan el fenómeno del atropellamiento por medio de métodos de selección univariantes [217]. Para esto se tomó como base los algoritmos de selección disponibles en el paquete de selección univariante de características de Scikit-learn. Así mismo, fue importante la selección de características a diferentes escalas, tal como fue mostrado por Ha y Shilling [129] obteniendo diferentes resultados de contribución a la predicción con la misma característica medida a diferentes escalas.

Con respecto al método de selección de características con mejor desempeño, la métrica de Información Mutua (MI) ha demostrado ser un método con resultados prometedores en diferentes áreas de las ciencias computacionales, permitiendo la selección de las características con mayor cantidad de información relevante para la salida del algoritmo de clasificación [217, 218]. Por esta razón, inicialmente se partió de la hipótesis de que las características seleccionadas por el modelo estarían relacionadas con el uso del suelo alrededor de los sucesos, la distancia a coberturas boscosas, la distancia a fuentes de agua, entre otras características que modelan el movimiento animal y por lo tanto, podrían modelar las zonas de cruce de fauna, tal como se demostró por Kantola, et al. [90], Gonçalves, et al. [93], Ascensão, et al. [113], Fabrizio, et al. [128], Ha y Shilling [129].

Las características seleccionadas para este proyecto coinciden con el supuesto inicial, dado que están relacionadas con la calidad de los ecosistemas [219]. Ejemplo de ello son: Distancia a Bosque_300m, Distancia a corredor biológico_300m,

Resistencia_150m, Resistencia_300m, Distancia a Bosque_150m, Distancia a corredor biológico_150m, Distancia a corredor biológico_90m, Distancia a Bosque_90m, Costo de Movimiento_300m, las cuales provienen del modelo de conectividad ecológica [196].

Las características: band_11_300m, band_10_300m, band_9_300m band_11_150m, NBRI_300m que corresponden a las bandas térmicas captadas por el Sensor Térmico Infrarrojo (TIRS) del satélite Landsat 8. Estas bandas captan la temperatura del suelo, la cual disminuye ante la presencia de vegetación, permitiendo observar de forma cuantitativa la regulación térmica provista por los bosques y coberturas vegetales [220], brindando zonas más atractivas para el movimiento animal como fue descrito por Maffei y Andrew [221].

4.2. Modelo de predicción de atropellamiento

Esta sección tiene como objetivo evaluar la viabilidad de los algoritmos de aprendizaje de máquina seleccionados anteriormente para la predicción del atropellamiento de fauna

4.2.1. Preprocesamiento y partición de la base de datos

A partir de la matriz de características seleccionadas en la sección anterior, se realizó el preprocesamiento de la información como preparación para la fase de entrenamiento de los algoritmos, este proceso tiene por objetivo particionar la matriz de características en bloques de entrenamiento y validación, siendo el bloque de entrenamiento el conjunto de datos a balancear en términos de desbalance de clases, mientras que el bloque de validación será el bloque de datos desconocidos sobre el que se evaluará el desempeño del algoritmo. Con el propósito de observar la capacidad real de los modelos ajustados se realizó validación cruzada en 4 bloques o pliegues de entrenamiento y validación. En la Tabla 4 puede observarse la distribución de los segmentos en cada pliegue.

Tabla 4. Distribución de los segmentos de entrenamiento en cada pliegue de entrenamiento y validación. Fuente: autoría propia

Pliegue	Entrenamiento	Validación
1	1,2,3,5,6,7,9,10,11	0,4,8
2	0,2,3,4,6,7,8,10,11	1,5,9
3	0,1,3,4,5,7,8,9,11	2,6,10
4	0,1,2,4,5,6,8,9,10	3,7,11

4.2.2. Comparación de algoritmos de clasificación

Con el objetivo de identificar el algoritmo con mejor ajuste al fenómeno del atropellamiento de fauna se realizó la comparación de los clasificadores KNN, SVM, RNA y RF, las cuales serán descritas a continuación.

El algoritmo KNN utilizado se optimizó mediante búsqueda exhaustiva variando el valor de k entre 1 y 300 vecinos, calculando para cada valor de k la matriz de confusión y la estadística Kappa. Así mismo, el algoritmo SVM fue optimizado mediante búsqueda exhaustiva empleando la herramienta GridSearchCV. El algoritmo RF fue optimizado mediante algoritmos genéticos (GA) por medio de la librería TPOT [204], donde se generaron 5 generaciones de 24 clasificadores, a partir de los cuales se eligieron los 12 mejores clasificadores de cada generación hasta llegar al valor óptimo. La optimización de hiperparámetros se realizó al interior de una validación cruzada en 4 pliegues, con una función de error de tipo exactitud balanceada. El algoritmo de Redes Neuronales fue construido por medio de la librería Keras [222] usando modelos de tipo secuencial. Este modelo recibe una estructura de capas y neuronas, el tipo de función de activación de cada neurona, el tipo de inicialización, la función de costo, el tipo de optimizador, el tamaño de cada bloque de entrenamiento y la cantidad de épocas a entrenar. Con el objetivo de minimizar el error de la red se implementó una búsqueda de exhaustiva para la cantidad de neuronas en la red, para esto se utilizó la función GridSearchCV. La estructura básica de la red sobre la cual se hizo la búsqueda fue una red con 17 características de entrada, 4 capas ocultas y una neurona de salida con una función sigmoide, esta estructura se observa en la Figura 16. La búsqueda del número de neuronas ideales tuvo como rango de búsqueda un n entre 1 y 100 neuronas. En la Tabla 5 se observan los resultados de cada uno de los métodos de clasificación entrenados, las técnicas de re-muestreo empleadas, así como las métricas de desempeño F1-score, Exactitud, Kappa y AUC-ROC. En dicha tabla puede además observarse que los mejores

resultados en términos de AUC de los algoritmos: RNA (0.63 ±0.11), RF (0.78 ± 0.12), SVM (0.59±0.17) y KNN (0.61 ± 0.1), fueron obtenidos con los métodos de re-muestreo Borderline SMOTE, ADASYN y KMEANS SMOTE, respectivamente.

Con el objetivo de identificar el algoritmo con mejor desempeñó se realizó la prueba estadística no paramétricas de Friedman y el test de comparación múltiple LSD para el algoritmo RF ADASYN, lo anterior debido a que este algoritmo fue el que presentó un mayor valor de Kappa promedio respecto a los demás algoritmos, los resultados del test de comparaciones múltiples se presentan en la Tabla 6, en esta puede observarse cada comparación entre RF ADASYN y los demás algoritmos utilizados, así mismo, se muestra el valor *p* de significancia estadística con un intervalo de confianza del 90%. A partir de estos se eligió el algoritmo RF en conjunto con el método de re-muestreo ADASYN como la técnica a ser utilizada para la predicción del fenómeno de atropellamiento, debido a que su comportamiento muestra diferencias estadísticas significativas con un intervalo de confianza del 90% con respecto a los clasificadores comparados. En términos estadísticos, el comportamiento de los modelos RF fueron similares, por lo cual pudo ser utilizado cualquiera, sin embargo, el RF ADASYN fue el método con mejor AUC promedio.

Figura 16. Estructura base de la red neuronal, previo a la búsqueda ideal de la estructura. Fuente: autoría propia

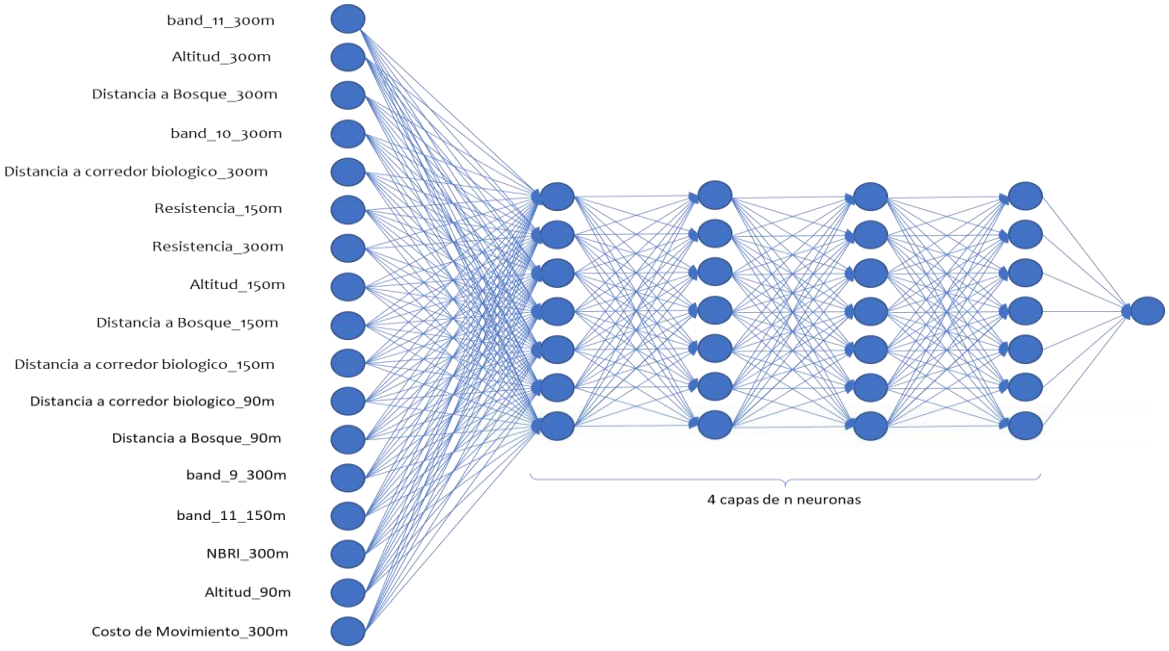


Tabla 5. Resultado consolidado de la etapa de Entrenamiento y validación en la zona de entrenamiento. Fuente: autoría propia

Algoritmo de clasificación	Método de Re-muestreo	F1-Score	Exactitud	Kappa	AUC- ROC
RNA	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

Tabla 6. Tabla de comparación múltiple entre el algoritmo RF ADASYN y los demás algoritmos probados con un intervalo de confianza al 90%. Fuente: autoría propia

Algoritmo Seleccionado	Algoritmo comparado	p valor
RF ADASYN	**SVM_SVMSMOTE**	**0.002787**
	RNA_SVMSMOTE	**0.006505**
	RF_SVMSMOTE	0.857609
	KNN_SVMSMOTE	**0.036329**
	SVM_SMOTE	**0.00928**
	RNA_SMOTE	**0.012009**
	RF_SMOTE	0.810931
	KNN_SMOTE	0.100037
	SVM_KMEANS	**0.013066**
	RNA_KMEANS	**0.072782**
	RF_KMEANS	0.369668
	KNN_KMEANS	0.10636
	SVM_BORDERLINE	**0.014204**
	RNA_BORDERLINE	0.188261
	RF_BORDERLINE	0.764916
	KNN_BORDERLINE	**0.059578**
	SVM ADASYN	**0.031316**
	RNA ADASYN	**0.059578**
	KNN ADASYN	**0.08285**

4.2.4. Análisis de resultados y discusión

Esta sección de discusión tiene como propósito analizar y comparar los resultados obtenidos con respecto a otras investigaciones publicadas.

Preprocesamiento y partición de la base de datos

De la etapa de preprocesamiento realizada en este proyecto fue realizada en 2 fases, una fase de normalización realizada durante la selección de características y una etapa de balance de clases por medio de técnicas de sobre muestreo. Durante la fase de normalización se utilizó la función MinMaxScaler de Scikit-learn [156], esta función tiene ventajas en cuanto a la facilidad de su implementación, así como en términos de

funcionamiento, debido a que ignora los datos perdidos o NaN, siendo estos problemáticos con otro tipo de funciones. De igual manera se realizó una fase de balance de clases, esto debido a que se identificó un desbalance de clases con una magnitud de 4:1 en favor de la clase 0 (puntos fríos), obteniendo resultados con valores máximos de 0.9 en términos de AUC-ROC, lo cual coincide con lo reportado por Haibo, et al. [161], Han, et al. [164], Douzas, et al. [165].

Adicionalmente, este trabajo usó una validación cruzada por grupos, garantizando así que los resultados no contienen sesgos relacionados con la alta correlación espacial que existe entre los segmentos vecinos, permitiendo evaluar el modelo en escenarios completamente desconocidos. Esto, según el conocimiento actual de los autores no ha sido utilizado para validar modelos de predicción de atropellamiento de fauna en el pasado, sin embargo la validación por grupos es una técnica usualmente utilizada para validar modelos de predicción espacial como fue reportado por Kajornrit y Wong [223].

Estimación de hiperparámetros y evaluación de desempeño

Como es conocido, el éxito en la implementación de un algoritmo de aprendizaje de máquina depende de la elección de hiperparámetros adecuados a la hora de entrenar un modelo [224]. Debido a esto, los algoritmos usados en este proyecto fueron optimizados durante la etapa de entrenamiento. Aunque el método de optimización de cada método fue seleccionado según la complejidad del problema de optimización, los métodos descritos son del tipo meta-heurístico, hallando los valores ideales de estos parámetros a partir de la evaluación de un set de iteraciones previamente configuradas por el usuario [225].

En el caso del algoritmo de vecinos más cercanos (KNN), solo se tiene un parámetro a optimizar, por lo que se decidió utilizar el método de búsqueda exhaustiva o fuerza bruta. Aunque existen métodos más rápidos como los propuestos por Fukunaga y Narendra [226], Moreno-Seco, et al. [227], Seongjoon y Koeng-Mo [228]. Debido a la baja cantidad de datos en la fase de entrenamiento y que el tiempo de ejecución del algoritmo fue menor a 5 minutos, no se consideró necesario el uso de otros algoritmos que permitieran disminuir el tiempo de entrenamiento.

Así mismo, en el caso del algoritmo de Máquina de Soporte Vectorial (SVM) aunque existe una gran variedad en los posibles Kernel a ser utilizados en la etapa de

entrenamiento, se decidió utilizar el Kernel de base radial debido a los resultados positivos que ha mostrado en diferentes aplicaciones como las presentadas por Huichuan y Xiyu [229], Ye y Li [230], entre otros. Adicionalmente, limitar el problema de optimización a un solo Kernel, permitió enfocarse en hallar los valores de *C* y *Gamma* por medio de métodos de búsqueda de exhaustiva con métodos de validación cruzada, garantizando una selección de hiperparámetros ajustada a diferentes conjuntos de datos, evitando sobre entrenar el modelo [231].

Con respecto a las redes neuronales artificiales (RNA) la optimización de la estructura de la red es un problema que para el entendimiento actual de los autores aún no se tiene una solución óptima. Por lo cual, en este proyecto se decidió realizar una optimización por búsqueda exhaustiva de los parámetros de estructura de la red mediante múltiples repeticiones del algoritmo de optimización. Por último, el algoritmo de bosques aleatorios fue optimizado mediante algoritmos genéticos, esta clase de algoritmos permite una optimización de múltiples hiperparámetros de una manera eficiente y eficaz [180]. Es importante mencionar que el algoritmo de bosques aleatorios permite ser implementado sin definir los hiperparámetros, puesto que el modo automático permite generar árboles de profundidad n hasta encontrar hojas puras [29], por lo cual el uso de métodos de optimización se realiza para garantizar una configuración ideal. Sin embargo, si estas técnicas de optimización no fuesen aplicadas sus resultados serían similares debido a su principio de operación. A pesar de esto, es positivo hacer uso de estos algoritmos para garantizar que no se esté generando un sobre entrenamiento del modelo.

Comparación de algoritmos de clasificación

Respecto a las métricas de desempeño no es recomendable comparar los métodos a través de métricas como la exactitud y otras métricas que no tienen en cuenta el desbalance, por lo cual se recomienda usar métricas como el estadístico Kappa, la métrica F1-Score y el AUC- ROC. Sin embargo, se debe tener en cuenta que estas medidas están influenciadas en gran medida por la exactitud del modelo en la predicción de la clase dominante [232]. El modelo KNN con el método de re-muestreo KMeans SMOTE fue el método con el mejor desempeño en términos de Kappa y AUC-ROC con 0.171 y 0.61, respectivamente. Adicionalmente, el algoritmo KNN tuvo un comportamiento promedio similar a un clasificador aleatorio (AUC-ROC = 0.5), excepto por el algoritmo KNN KMEANS, cuya tendencia permanece por encima de este límite en la mayor parte de la gráfica.

Aunque según el conocimiento actual de los autores el algoritmo de KNN no ha sido usado previamente en la predicción del atropellamiento de fauna, el desempeño de este clasificador no ha sido satisfactorio, debido a que no muestra un potencial de predicción mejor que el reportado por [113], donde un algoritmo de regresión logística binomial tuvo un AUC-ROC de 0.66 ± 0.09 y 0.72±0.14 para los segmentos de validación evaluados. Por último, es importante notar que el estadístico Kappa no tuvo variaciones significativas entre los métodos de re-muestreo evaluados presentando una máxima variación de 0.02 entre ellos.

Con respecto al clasificador SVM la búsqueda de hiperparámetros garantizó que los resultados presentados fueran con los valores óptimos. A pesar de esto, se evidencia en el comportamiento del clasificador que presenta una tendencia a predecir los puntos de validación hacia la clase 0, mostrando que el modelo no logra ajustarse al fenómeno del atropellamiento. A pesar de esto, el clasificador SVM según las métricas F1-score y exactitud es satisfactorio, sin embargo, estas métricas están altamente influenciadas por el desbalance de clase, en donde un clasificador que determine la clase dominante en todas las muestras, tendrá un F1 score y una Exactitud aproximada del 85%, según la proporción de clases en cada pliegue, evidenciando la importancia de realizar una comparación de los modelos planteados mediante la métrica Kappa y AUC ROC [232].

De acuerdo con los resultados del clasificador RF, este se caracterizó por un mejor ajuste al fenómeno del atropellamiento comparado con el resto de los algoritmos probados, especialmente con el método de re-muestreo ADASYN, obteniendo valores de Kappa promedio de 0.34. Respecto a los otros métodos de re-muestreo, el método KMeans SMOTE, tuvo el resultado más bajo con un valor de Kappa de 0.206 y un AUC-ROC de 0.70±0.11. Respecto al comportamiento del clasificador, es importante resaltar que el método de re-muestreo ADASYN generó una proporción de falsos negativos del 34.27% y falsos positivos del 19.15 %. Por lo cual, es importante buscar estrategias que permitan reducir la proporción de falsos negativos tomando a consideración la clasificación de las vecindades durante la implementación del modelo en situaciones prácticas, permitiendo así una mejor toma de decisiones e inversión de recursos.

Con respecto a la comparación de los métodos de balanceo se puede observar que han generado resultados similares a los publicados por otros autores (AUC-ROC > 0.7), especialmente los algoritmos con el método ADASYN, que brinda el mejor resultado (0.78 ± 12), mejorando el resultado obtenido por Ascensão, et al. [113] siendo el precedente que más se aproxima a la metodología aplicada en esta investigación, lo

cual podría indicar que sets de características con altos valores de información que han sido balanceados por vecindades con métodos como el ADASYN [161] permite que clasificadores basados en algoritmos de optimización por entropía, como el RF, tengan un mayor ajuste al fenómeno del atropellamiento. Así mismo, es importante destacar que este método de re-muestreo presentó el mejor resultado de esta investigación en términos del estadístico Kappa (0.343)

Por último, se observaron resultados positivos del estadístico Kappa para todos los métodos de re-muestreo implementados junto con las redes neuronales artificiales, por lo cual, según el conocimiento de los autores, las redes neuronales no han sido implementadas al problema del atropellamiento de fauna, requiriéndose una mayor cantidad de investigaciones que permitan determinar una estructura de red que permita mejorar los resultados aquí presentes. Lo anterior debido a que las redes neuronales han demostrado ser especialmente efectivas en la predicción de diferentes fenómenos [31-33, 35, 37].

Con el objetivo de elegir el algoritmo con un desempeño significativo sobre los demás, se aplicó el test de Friedman, a partir del cual se pudo observar que el algoritmo RF con el método de muestreo ADASYN fue el algoritmo con la media de desempeño más alta. Adicionalmente, por medio del test de comparación múltiple (Sección 0 – Anexo 4) se identificó que dicho algoritmo presenta diferencias significativas ($p<0.10$) con los algoritmos: SVM, RNA con los métodos de re-muestreo: SVM-SMOTE y SMOTE y KNN con el método de re-muestreo SVM-SMOTE. Por último, se identificó que este algoritmo presenta comportamientos estadísticamente similares con el algoritmo RF Borderline SMOTE, RF KMeans y RF SVM-SMOTE, KNN SMOTE, KNN KMEANS y RNA_BORDERLINE. Sin embargo, debido a que el algoritmo RF ADASYN tiene una media de desempeño más alta, es el algoritmo seleccionado para la fase de aprendizaje y validación por medio de transferencia de aprendizaje.

4.3. Transferencia de aprendizaje en vías sin muestreo

Con el objetivo de validar la metodología propuesta para la predicción de puntos calientes de atropellamiento de fauna, se implementó la metodología planteada para la etapa de validación. Para esto fue necesario realizar la partición inicial de la matriz de características en 4 bloques o pliegues de entrenamiento y transfer, los cuales al final de cada evaluación rotarán hasta que cada segmento haya sido al menos una vez parte

del conjunto de prueba [200], realizando el balance de clases en cada iteración. En la Tabla 7 puede observarse la distribución de los segmentos de entrenamiento en cada pliegue.

4.3.1. Validación por medio de transferencia de aprendizaje

A continuación, se presentan los resultados del modelo RF ADASYN, seleccionado como el modelo con mejor ajuste al fenómeno de atropellamiento en la sección anterior. Con el objetivo de evaluar la incidencia de la cantidad de datos de re-entrenamiento sobre el desempeño del modelo se evaluó diferentes porcentajes de

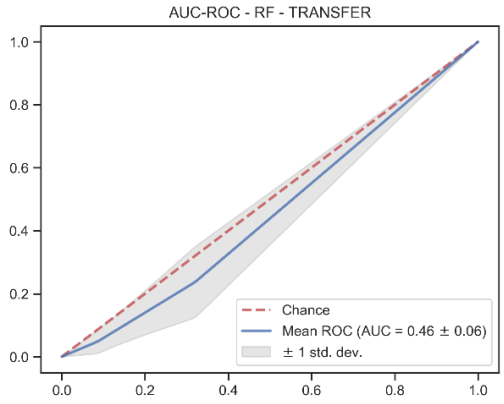
Tabla 7. Distribución de los segmentos de entrenamiento en cada pliegue de entrenamiento y validación. Fuente: autoría propia

Pliegue	Tramos de entrenamiento	Tramos de validación
1	1,2,3	4
2	2,3,4	1
3	1,3,4	2
4	1,2,4	3

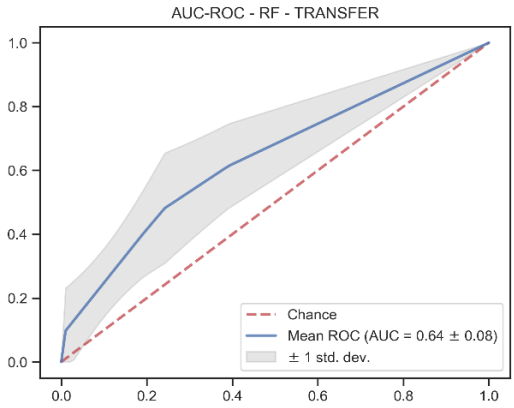
puntos correspondiente a los segmentos de vía agregados. En la Figura 17 se observa el comportamiento de las gráficas ROC con 0%, 1%, 5%, 10%, 15% y 20% de la longitud total de la vía a predecir.

Figura 17. AUC-ROC del clasificador RF ADASYN al ser sometido a diferentes % de datos agregados al conjunto de validación. A) 0%, B) 1%, C) 5%, D) 10%, E), 15%, F) 20%. Fuente: autoría propia

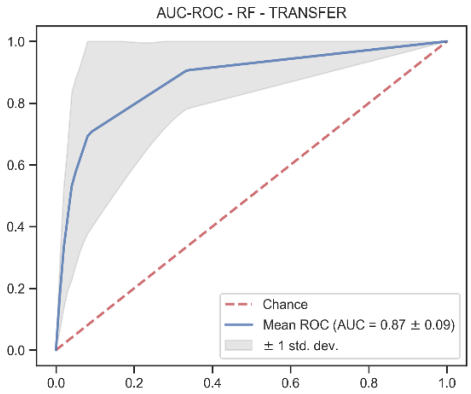
A)



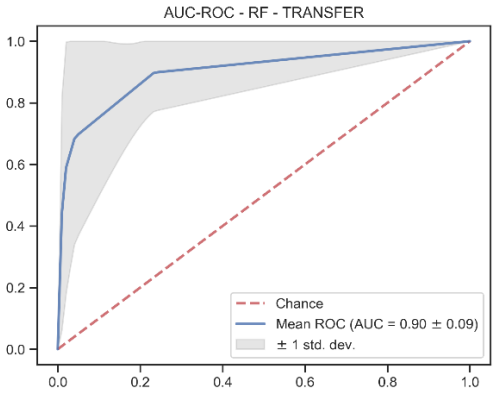
B)



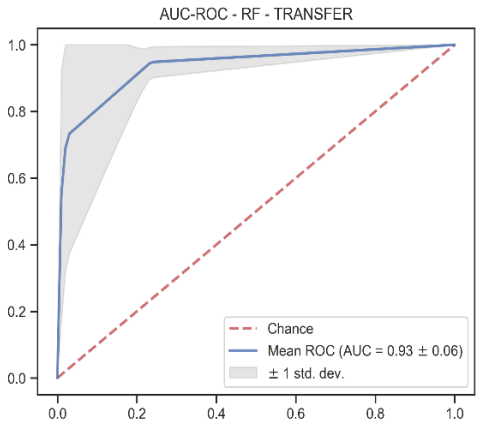
C)



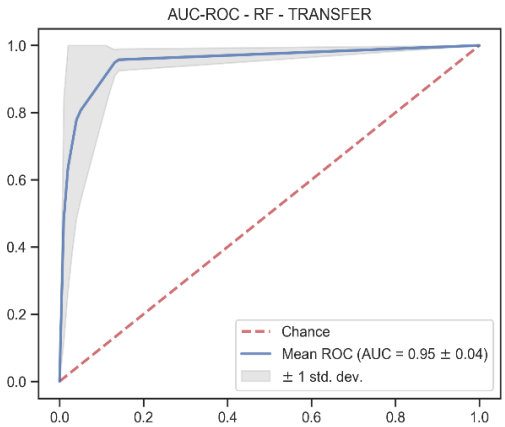
D)



E)

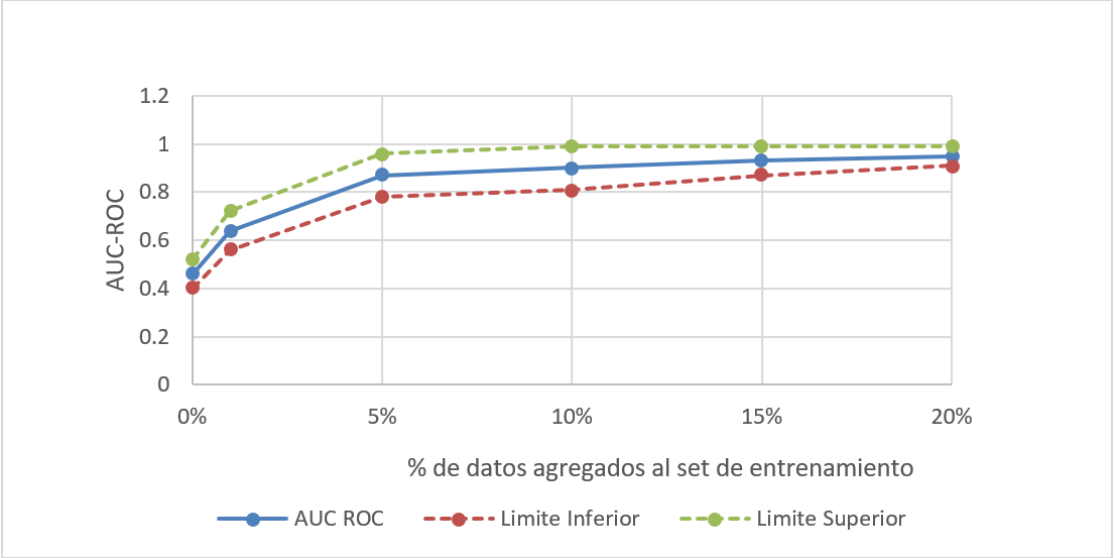


F)



Adicionalmente, en la Figura 18 se muestran los valores de AUC-ROC para cada uno de los porcentajes utilizados, en esta se observa que el valor de AUC-ROC se estabiliza a partir de un 5% de datos agregados a la base de datos de entrenamiento.

Figura 18. Valores AUC-ROC del modelo RF-ADASYN al ser reentrenado con múltiples porcentajes de datos. Fuente: autoría propia



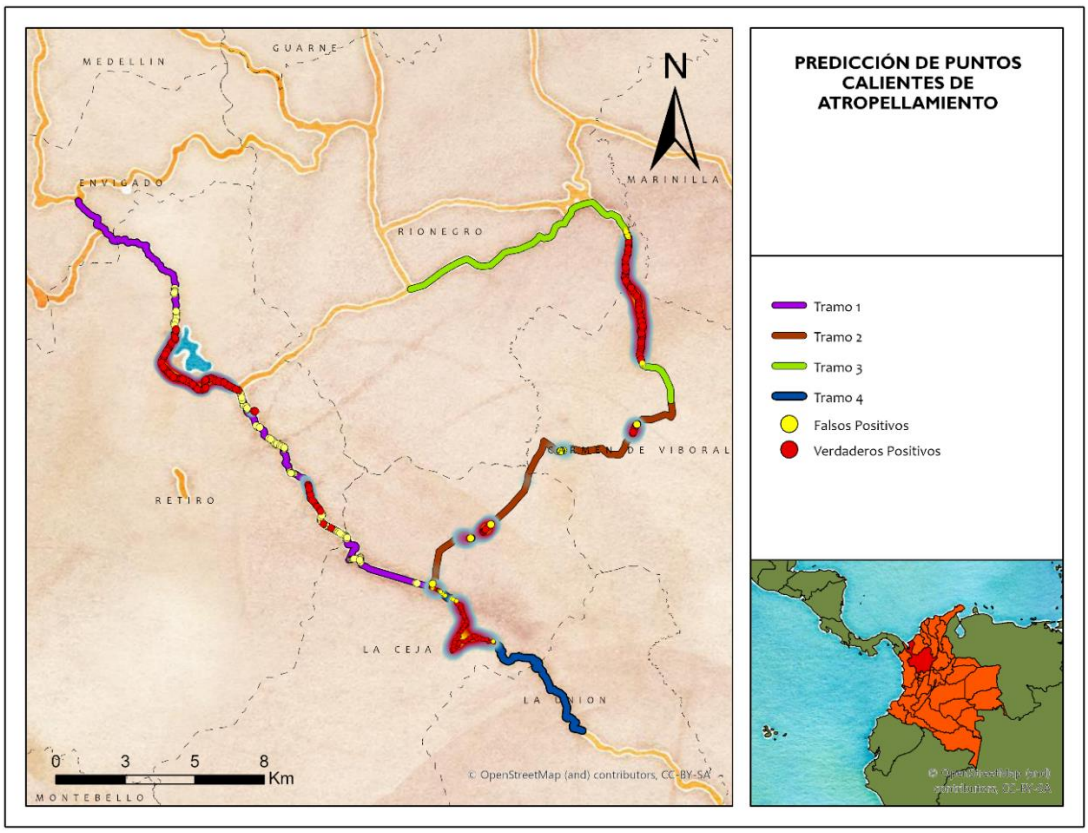
Por último, en la Tabla 8 se muestra el desempeño a detalle del clasificador RF ADASYN con el 5% de datos de re-entrenamiento.

Tabla 8. Métricas de desempeño promedio del clasificador RF ADASYN en la fase de Transfer Learning con un % de datos de entrenamiento adicionales del 5%. Fuente: autoría propia

Pliegue	F1-Score	Exactitud	Kappa	AUC-ROC
1	0.9497	0.9533	0.8975	0.9552
2	0.7498	0.8155	0.4618	0.7782
3	0.9176	0.9174	0.5726	0.746
4	0.9781	0.979	0.9475	0.9822
Media	0.8988	0.9163	0.7199	0.8654

Como resultado consolidado, en la Figura 19 se presenta el mapa con la predicción de todos los Tramos estudiados mediante la técnica de transferencia de aprendizaje con un porcentaje de datos adicionales de entrenamiento del 5%, así mismo, se presentan las métricas de desempeño promedio del clasificador.

Figura 19. Mapa resultante del algoritmo de clasificación para el Área de estudio. Fuente: autoría propia



4.3.2. Análisis de resultados y discusión

Con el propósito de realizar una validación de la metodología de predicción de puntos calientes de atropellamiento de fauna a partir de imágenes multiespectrales y sistemas de información geográfica, se propone realizar una adaptación de la técnica de Transfer Learning [233] usada en redes neuronales para implementarla en este problema de clasificación particular. Por lo cual, con el propósito de realizar una validación del algoritmo por medio de validación cruzada, se realiza una partición de la base de datos por bloques de entrenamiento y validación, generando un set de tramos de entrenamiento y dejando un tramo de validación por fuera. Es importante mencionar que todos los tramos del área de estudio poseen patrones de distribución del atropellamiento diferentes, por lo cual, al realizar el entrenamiento directamente con datos de zonas diferentes a la zona de validación, se espera un comportamiento similar a los observados en los experimentos de transferencia de aprendizaje si el algoritmo generaliza de manera adecuada el fenómeno de atropellamiento [234].

Durante el desarrollo metodológico se utilizaron diferentes porcentajes de los segmentos de vía del tramo de validación para alimentar el modelo. En la Figura 17, se puede observar los resultados del clasificador según la proporción de datos del tramo de validación usados durante el entrenamiento, puede observarse una mejoría directamente proporcional al porcentaje de datos utilizados, obteniendo valores de AUC ROC desde 0.64 ± 0.08, correspondiente al clasificador reentrenado con el 1% de los datos de validación, hasta 0.95 ± 0.04, correspondiente al clasificador reentrenado con el 20% de los datos de validación.

Teniendo en cuenta lo anterior y que el propósito práctico del algoritmo generado por esta investigación es el de optimizar los recursos económicos y humanos requeridos para la identificación de puntos calientes de atropellamiento de fauna silvestre en zonas que no exista abundancia de información, se decide seleccionar el clasificador entrenado con el 5% de los datos de validación como método ideal a ser utilizado en la práctica, esto debido a su bajo requerimiento de datos y su exitosa clasificación del fenómeno de atropellamiento.

4.4. Metodología propuesta para la predicción de puntos calientes de atropellamiento de fauna

Como resultado acumulado de este proyecto, en la Figura 20 se presenta el esquema metodológico validado por esta investigación para la predicción de puntos calientes de atropellamiento de fauna en vías del Oriente Antioqueño.

La metodología propuesta en este proyecto recolectó 837 reportes de atropellamiento, evidenciando la presencia del atropellamiento de fauna en el área de estudio, así mismo, por medio de técnicas geo estadísticas como el K de Ripley y el análisis de autocorrelación espacial de Moran se identificó que el patrón de puntos presenta diferencias significativas con respecto a una distribución normal en las bandas de distancia de 1.3 km para los tramos 1, 3 y 4, y 269 m para el tramo 2, a partir del cual se identificaron los puntos calientes de atropellamiento en cada uno de los tramos del área de estudio.

Figura 20. Metodología propuesta para la predicción de puntos calientes de atropellamiento de fauna. Fuente: autoría propia

Posteriormente, se identificó que las variables: distancia a Bosque, distancia a corredor biológico, resistencia del terreno al movimiento, costo de movimiento, las bandas 9, 10 y 11 del satélite Landsat 8 y el índice de quema normalizado (NBRI) brindan información relevante a los modelos de aprendizaje automático acerca del fenómeno de atropellamiento de fauna. Así mismo, al comparar los algoritmos RF, RNA, KNN y SVM, ajustados con las variables identificadas se obtuvo que el algoritmo RF con el método de re-muestreo ADASYN posee diferencias significativas con un intervalo de confianza del 90% con todos los métodos evaluados excepto consigo mismo, con el KNN re muestreado con los métodos SMOTE, KMEANS y con el algoritmo RNA con el método de re-muestreo BORDERLINE. Por último, al realizar predicciones en zonas sin datos se identificó que con datos correspondientes al 5% de la longitud total de la vía agregado a la base de datos de entrenamiento de forma aleatoria, se obtuvo un desempeño del algoritmo en términos de AUC-ROC de 0.87 ± 0.09, superando los resultados reportados en la literatura.

5. Conclusiones y recomendaciones

5.1. Conclusiones

- Esta investigación desarrolló una metodología para predicción de los sitios de mayor acumulación de atropellamiento de fauna, en vías del Oriente Antioqueño con base en algoritmos de inteligencia artificial, sistemas de información geográfica y procesamiento de imágenes multiespectrales. Esta metodología cuenta con una fase de caracterización del fenómeno de atropellamiento de fauna a partir de imágenes multiespectrales, una fase de selección de características por medio de métodos de selección univariantes, una fase de selección del modelo de aprendizaje automático con mejor ajuste y una fase de prueba en zonas desconocidas por el modelo. Con los experimentos se demostró que el algoritmo RF ADASYN fue el algoritmo con mejor desempeño (AUC-ROC = 0.78 ± 0.12) en zonas de aprendizaje y (AUC-ROC = 0.87 ± 0.09) en zonas de transferencia de aprendizaje al ser reentrenado con el 5% de información de la zona de predicción

- A partir del análisis de puntos calientes mediante técnicas geo estadísticas y geo espaciales, así como la estimación de la información mutua de las características recolectadas en la fase de caracterización, se determinó que las características espaciales más relacionadas con el atropellamiento de fauna silvestre en el Oriente Antioqueño son: Banda 9,10 y 11 del Satélite Landsat 8, Altitud, Distancia a Bosque, Distancia a corredor biológico, Resistencia al movimiento, Costo de movimiento y el índice multiespectral: NBRI.

- Al comparar diferentes algoritmos de inteligencia artificial se pudo identificar que el algoritmo con mejor ajuste a la predicción de puntos calientes de atropellamiento de fauna en carreteras del oriente antioqueño, teniendo en cuenta el desbalance de clases fue Boques Aleatorios (RF) con el método de re-muestreo ADASYN, el cual tuvo un desempeño promedio según la métrica AUC-ROC de 0.78 ± 0.12 en los sets de validación.

- Al validar la metodología propuesta para la predicción de puntos calientes de atropellamiento de fauna por medio de transferencia de aprendizaje en las vías del oriente antioqueño, se tuvo que el algoritmo de RF ADASYN reentrenado con información de acumulaciones de atropellamiento equivalentes al 5% de la longitud total de la vía a predecir tuvo un desempeño promedio medido por el AUC-ROC de 0.87 ± 0.09

- La metodología realizada en este trabajo tiene el potencial de disminuir los tiempos de respuesta de la academia, entes de control y operadores viales, al fenómeno del atropellamiento de fauna, permitiendo estimar las zonas con potenciales puntos calientes de atropellamiento para posteriormente ser validados por estudios diagnósticos que permitirán identificar medidas de mitigación al atropellamiento.

5.2. Recomendaciones

- Es fundamental el entrenamiento del modelo con información que haya sido plenamente identificada como significativa por medio de métodos geo estadísticos y geo espaciales, reduciendo de esta manera la influencia del azar sobre las predicciones del modelo.

- Para la fase de entrenamiento de los modelos de clasificación es importante usar métodos de optimización que permitan garantizar la elección de hiperparámetros ideales para cada set de entrenamiento.

- Aunque en esta investigación fueron utilizados métodos de selección de características univariantes, es posible utilizar otros métodos de selección para este tipo de problemáticas. Entre las opciones a utilizar se recomienda el uso de métodos de selección metaheurísticos como lo son los algoritmos genéticos y los métodos de optimización por enjambre de partículas (PSO), tal como fue mencionado en el Estado del Arte de esta investigación.

- Aunque no fue cubierto en esta investigación, existe la posibilidad de realizar predicciones espaciales por medio de algoritmos de aprendizaje profundo (Deep Learning), los cuales deben ser explorados en los años venideros y el cual es considerado el trabajo futuro de esta investigación.

- La metodología desarrollada en esta investigación puede extrapolarse a otras regiones, sin embargo, debe tenerse en cuenta el contexto biótico y abiótico de cada zona de estudio, modificando las características exploradas, así como la metodología de construcción de estas.

- Los puntos calientes de atropellamiento de fauna identificados por este trabajo corresponden a animales vertebrados sin hacer distinción por clase o especie, por lo cual, se abre la posibilidad de realizar como trabajo futuro la validación de estos modelos para la predicción de puntos calientes de clases o especies específicas teniendo en cuenta los patrones temporales y climáticos de la zona.

- Este trabajo muestra una aproximación teórica a la predicción de puntos calientes en zonas con pocos datos recolectados, sin embargo, es necesario realizar una validación en terreno de los resultados obtenidos por medio de esta metodología en otras zonas de estudio, lo cual se realizará como trabajo futuro al interior del ITM.

- El área en la que se realizaran las predicciones no necesariamente debe contener información estadísticamente significativa, debido a que recolectar esta información requeriría de un muestreo estandarizado. Aunque es recomendable realizar un muestreo de este tipo, la ciencia ciudadana y los diagnósticos rápidos participativos han demostrado ser valiosas herramientas a la hora de recolectar información de una manera rápida y costo-eficiente. Sin embargo, la información a partir de la cual se realiza el entrenamiento se recomienda que sea recolectada por medio de una metodología estandarizada que permita identificar el patrón de los atropellamientos en la zona.

6. Bibliografía

[1] J. A. G. Jaeger, "Improving Environmental Impact Assessment and Road Planning at the Landscape Scale," *Handbook of Road Ecology,* pp. 32-42, 2015/04/01 2015, doi: https://doi.org/10.1002/9781118568170.ch5.

[2] A. W. Coffin, "From roadkill to road ecology: A review of the ecological effects of roads," (in English), *J Transp Geogr,* vol. 15, no. 5, pp. 396-406, Sep 2007, doi: https://www.doi.org/10.1016/j.jtrangeo.2006.11.006.

[3] R. van der Ree, D. J. Smith, and C. Grilo, "The Ecological Effects of Linear Infrastructure and Traffic," in *Handbook of Road Ecology*, 2015, pp. 1-9. doi: https://doi.org/10.1002/9781118568170.ch1.

[4] P. Cramer, M. Olsson, M. E. Gadd, R. van der Ree, and L. E. Sielecki, "Transportation and Large Herbivores," in *Handbook of road ecology*, 2015. doi: https://doi.org/10.1002/9781118568170.ch42.

[5] R. van der Ree, D. J. Smith, and C. Grilo, *Handbook of road ecology*. Wiley, 2015, pp. xxvi, 522 pages. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118568170.

[6] J. Davenport and J. Davenport, *The Ecology of Transportation: Managing Mobility for the Environment.* 2006, p. 402. [Online]. Available: https://www.springer.com/gp/book/9781402045035.

[7] W. F. Laurance *et al.*, "A global strategy for road building," *Nature,* vol. 513, p. 229, 08/27/online 2014, doi: https://doi.org/10.1038/nature13717.

[8] A. P. Clevenger and N. Waltho, "Performance indices to identify attributes of highway crossing structures facilitating movement of large mammals," *Biological Conservation,* vol. 121, no. 3, pp. 453-464, 2005/02/01/ 2005, doi: https://doi.org/10.1016/j.biocon.2004.04.025.

[9] B. A. Crawford, J. C. Maerz, N. P. Nibbelink, K. A. Buhlmann, T. M. Norton, and S. E. Albeke, "Hot spots and hot moments of diamondback terrapin road-

crossing activity," *Journal of Applied Ecology,* vol. 51, pp. 367-375, 2014, doi: https://doi.org/10.1111/1365-2664.12195.

[10] J. C. Cureton and R. Deaton, "Hot moments and hot spots: Identifying factors explaining temporal and spatial variation in turtle road mortality," *Journal of Wildlife Management,* vol. 76, pp. 1047-1052, 2012, doi: https://doi.org/10.1002/jwmg.320.

[11] Z. D. Danks and W. F. Porter, "Temporal, Spatial, and Landscape Habitat Characteristics of Moose–Vehicle Collisions in Western Maine," *Journal of Wildlife Management,* vol. 74, pp. 1229-1241, 2010, doi: https://doi.org/10.2193/2008-358.

[12] R. T. T. Forman *et al.*, *Road ecology: Science and solutions* (Environmental Progress). 2003, pp. O16-O16. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/ep.670220307.

[13] X. Girardet, G. Conruyt-Rogeon, and J. C. Foltête, "Does regional landscape connectivity influence the location of roe deer roadkill hotspots?," *European Journal of Wildlife Research,* vol. 61, pp. 731-742, 2015, doi: https://doi.org/10.1007/s10344-015-0950-4.

[14] K. E. Gunson, G. Mountrakis, and L. J. Quackenbush, "Spatial wildlife-vehicle collision models: A review of current work and its application to transportation mitigation projects," *J. Environ. Manage.,* vol. 92, pp. 1074-1082, 2011, doi: https://doi.org/10.1016/j.jenvman.2010.11.027.

[15] A. B. Madsen, H. Strandgaard, and A. Prang, "Factors causing traffic killings of roe deer Capreolus capreolus in Denmark," *Wildlife Biology,* vol. 8, pp. 55-61, 2002, doi: https://doi.org/10.2981/wlb.2002.008.

[16] J. E. Malo, F. Suárez, and A. Díez, "Can we mitigate animal-vehicle accidents using predictive models?," *Journal of Applied Ecology,* vol. 41, no. 4, pp. 701-710, 2004, doi: https://doi.org/10.1111/j.0021-8901.2004.00929.x.

[17] M. A. Adárraga Caballero and L. C. Moreno Gutiérrez, "Mortalidad de vertebrados silvestres en dos segmentos de la carretera troncal del Caribe a su paso a través de dos ecosistemas de interés biológico en la costa Caribe

colombiana (Magdalena)," presented at the III Seminario Internacional de Ciencias Ambientales, SUE-Caribe, Barranquilla, Colombia, 2017.

[18] O. De La Ossa-Nadjar and J. De La Ossa V., "Vehicle collisions with wild fauna on the two roads that pass through the Montes de María, Sucre, Colombia," *Revista U.D.C.A Actualidad & Divulgación Científica,* vol. 18, pp. 503-511, 2015. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0123-42262015000200024&nrm=iso.

[19] J. De La Ossa-V and S. Galván-Guevara, "Registro de mortalidad de fauna silvestre por colisión vehicular en la carretera Toluviejo – ciénaga La Caimanera, Sucre, Colombia," (in Español), *Biota Colombiana,* vol. 16, no. 1, pp. 67-77, 2015. [Online]. Available: https://www.redalyc.org/articulo.oa?id=49142418007.

[20] C. A. Delgado Vélez, "Adiciones al atropellamiento vehicular de mamíferos en la vía de el escobero, envigado (Antioquia), Colombia," *Revista EIA,* pp. 147-153, 2014. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S1794-12372014000200012&nrm=iso.

[21] C. A. Delgado-V., "Muerte de mamíferos por vehículos en la vía del escobero, envigado (Antioquia), Colombia," *Actualidades Biológicas,* vol. 29, pp. 229-233, 2007. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0304-35842007000200007&nrm=iso.

[22] D. F. López-Herrera, M. León-Yusti, S. C. Guevara-Molina, and F. Vargas-Salinas, "Reptiles en corredores biológicos y mortalidad por atropellamiento vehicular en Barbas-Bremen, Quindío, Colombia," *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales,* vol. 40, pp. 484-493, 2016. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0370-39082016000300010&nrm=iso.

[23] M. C. Monroy, A. De La Ossa-Lacayo, and J. De La Ossa-V, "Tasa de atropellamiento de fauna silvestre en la vía San Onofre – María la baja, Caribe Colombiano," (in Spanish), *Revista de la asociación colombiana de ciencias biológicas,* vol. 1, no. 27, 2015. [Online]. Available:

http://www.ojs.asociacioncolombianadecienciasbiologicas.org/index.php/accb/article/view/106.

[24] E. Payan, C. Soto, A. Diaz-Pulido, A. Benitez, and A. Hernandez, "Wildlife road crossing and mortality: lessons for wildlife friendly road design in Colombia," in *International Conference on Ecology and Transportation (ICOET 2013)*, 2013. [Online]. Available: https://www.semanticscholar.org/paper/WILDLIFE-ROAD-CROSSING-AND-MORTALITY%3A-LESSONS-FOR-Pay%C3%A1n-Soto/733a37a19e04781f3c3563d1031aa6754617066b.

[25] A. Quintero Angel, D. Osorio, F. Vargas-Salinas, and C. Saavedra-Rodríguez, "Roadkill rate of snakes in a disturbed landscape of Central Andes of Colombia," *Herpetology Notes,* vol. 5, pp. 99-105, 01/01 2012. [Online]. Available: https://pdfs.semanticscholar.org/9b64/9c0bdb29a2f909cbe941bf5c73058bd1d205.pdf.

[26] E. Ramos Pallares and F. Meza-Joya, "Reptile road mortality in a fragmented landscape of the middle Magdalena Valley, Colombia," *Herpetology Notes,* vol. 11, 01/26 2018. [Online]. Available: https://www.biotaxa.org/hn/article/view/29825.

[27] F. Vargas and I. Delgado, "Mortalidad por atropello vehicular y distribución de anfibios y reptiles en un bosque subandino en el occidente de Colombia," *Caldasia,* vol. 33, no. 1, 2011-01-01 2011. [Online]. Available: https://revistas.unal.edu.co/index.php/cal/article/view/36380.

[28] F. Vargas-Salinas and F. López-Aranda, "Las carreteras pueden restringir el movimiento de pequeños mamíferos en bosques andinos de Colombia? Estudio de caso en el bosque de Yotoco, Valle del Cauca," *Caldasia,* vol. 34, pp. 409-420, 2012. [Online]. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0366-52322012000200011&nrm=iso.

[29] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python: A Guide for Data Scientists*, 1 ed. (A Guide for Data Scientists). O'Reilly Media, 2016, p. 534. [Online]. Available: https://books.google.com.co/books?id=vbQlDQAAQBAJ.

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

[31] D. T. Bui, Q. T. Bui, N. Quoc-Phi, P. Biswajeet, N. Haleh, and T. T. Phan, "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," *Agricultural and Forest Meteorology,* vol. 233, pp. 32-44, 2017, doi: https://www.doi.org/10.1016/j.agrformet.2016.11.002.

[32] D. T. Bui *et al.*, "A novel hybrid approach based on a swarm intelligence optimized extreme learning machine for flash flood susceptibility mapping," *CATENA,* vol. 179, pp. 184-196, 2019/08/01/ 2019, doi: https://doi.org/10.1016/j.catena.2019.04.009.

[33] D. T. Bui, B. Pradhan, H. Nampak, Q. T. Bui, Q. A. Tran, and Q. P. Nguyen, "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," *Journal of Hydrology,* vol. 540, pp. 317-330, 2016, doi: https://www.doi.org/10.1016/j.jhydrol.2016.06.027.

[34] S. S. Durduran, "A decision making system to automatic recognize of traffic accidents on the basis of a GIS platform," (in English), *Expert Syst Appl,* vol. 37, no. 12, pp. 7729-7736, Dec 2010, doi: https://doi.org/10.1016/j.eswa.2010.04.068.

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

[36] H. Harirforoush and L. Bellalite, "A new integrated GIS-based analysis to detect hotspots: A case study of the city of Sherbrooke," in *Accident Analysis and Prevention*, vol. 130: Elsevier Ltd, 2016, pp. 62-74. doi: https://doi.org/10.1016/j.aap.2016.08.015.

[37] A. Jaafari, E. K. Zenner, M. Panahi, and H. Shahabi, "Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability," *Agricultural and Forest Meteorology,* vol. 266-267, pp. 198-207, 2019, doi: https://www.doi.org/10.1016/j.agrformet.2018.12.015.

[38] N. Ngoc Thach *et al.*, "Spatial pattern assessment of tropical forest fire danger at Thuan Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study," *Ecological Informatics,* vol. 46, pp. 74-85, 2018, doi: https://www.doi.org/10.1016/j.ecoinf.2018.05.009.

[39] O. Rahmati, N. Tahmasebipour, A. Haghizadeh, H. R. Pourghasemi, and B. Feizizadeh, "Evaluation of different machine learning models for predicting and mapping the susceptibility of gully erosion," *Geomorphology,* vol. 298, pp. 118-137, 2017, doi: https://doi.org/10.1016/j.geomorph.2017.09.006.

[40] D. A. Tarasov, A. G. Buevich, A. P. Sergeev, and A. V. Shichkin, "High variation topsoil pollution forecasting in the Russian Subarctic: Using artificial neural networks combined with residual kriging," *Applied Geochemistry,* vol. 88, pp. 188-197, 2018, doi: https://doi.org/10.1016/j.apgeochem.2017.07.007.

[41] M. Valasik, "Gang violence predictability: Using risk terrain modeling to study gang homicides and gang assaults in East Los Angeles," *Journal of Criminal Justice,* vol. 58, pp. 10-21, 2018, doi: https://doi.org/10.1016/j.jcrimjus.2018.06.001.

[42] H. Wang, F. Qin, and X. Zhang, "A spatial exploring model for urban land ecological security based on a modified artificial bee colony algorithm," *Ecological Informatics,* vol. 50, pp. 51-61, 2019, doi: https://doi.org/10.1016/j.ecoinf.2018.12.009.

[43] H. E. McClure, "An Analysis of Animal Victims on Nebraska's Highways," *The Journal of Wildlife Management,* vol. 15, no. 4, pp. 410-420, 1951, doi: https://doi.org/10.2307/3796584.

[44] L. R. Jahn, "Highway Mortality as an Index of Deer-Population Change," *The Journal of Wildlife Management,* vol. 23, no. 2, pp. 187-197, 1959, doi: https://doi.org/10.2307/3797639.

[45] E. D. Bellis and H. B. Graves, "Deer Mortality on a Pennsylvania Interstate Highway," *The Journal of Wildlife Management,* vol. 35, no. 2, pp. 232-237, 1971, doi: https://doi.org/10.2307/3799596.

[46] D. J. Oxley, M. B. Fenton, and G. R. Carmody, "The Effects of Roads on Populations of Small Mammals," *Journal of Applied Ecology,* vol. 11, no. 1, pp. 51-59, 1974, doi: https://doi.org/10.2307/2402004.

[47] M. J. Puglisi, J. S. Lindzey, and E. D. Bellis, "Factors Associated with Highway Mortality of White-Tailed Deer," *The Journal of Wildlife Management,* vol. 38, no. 4, pp. 799-807, 1974, doi: https://doi.org/10.2307/3800048.

[48] T. M. Pojar, T. C. Reseigh, and D. F. Reed, ""DEER CROSSING" SIGNS MAY PROVE VALUABLE IN REDUCING ACCIDENTS AND ANIMAL DEATHS," *HIGHWAY RESEARCH NEWS,* no. 46, p. 55, 1972. [Online]. Available: https://trid.trb.org/view/115299.

[49] T. M. Pojar, R. A. Prosence, D. F. Reed, and T. N. Woodard, "Effectiveness of a Lighted, Animated Deer Crossing Sign," *The Journal of Wildlife Management,* vol. 39, no. 1, pp. 87-91, 1975, doi: https://doi.org/10.2307/3800469.

[50] G. L. Storm, R. D. Andrews, R. L. Phillips, R. A. Bishop, D. B. Siniff, and J. R. Tester, "Morphology, Reproduction, Dispersal, and Mortality of Midwestern Red Fox Populations," *Wildlife Monographs,* no. 49, pp. 3-82, 1976. [Online]. Available: www.jstor.org/stable/3830425.

[51] C. R. Ferris, "Effects of Interstate 95 on Breeding Birds in Northern Maine," *The Journal of Wildlife Management,* vol. 43, no. 2, pp. 421-427, 1979, doi: https://doi.org/10.2307/3800351.

[52] R. M. Case, "Interstate Highway Road-Killed Animals: A Data Source for Biologists," *Wildlife Society Bulletin (1973-2006),* vol. 6, no. 1, pp. 8-13, 1978. [Online]. Available: www.jstor.org/stable/3781058.

[53] M. S. Johnson and J. L. Brown, "Genetic Variation among Trait Groups and Apparent Absence of Close Inbreeding in Grey-Crowned Babblers," *Behavioral Ecology and Sociobiology,* vol. 7, no. 2, pp. 93-98, 1980. [Online]. Available: www.jstor.org/stable/4599312.

[54] S. G. Ford, "Evaluation of highway deer kill mitigation on SIE/LAS-395.Final report, report no. FHWA/CA/TP-80/01.," U.S. Department of Transportation, Federal Highway Administration and California Department of Transportation, Sacramento, CA., 1980. [Online]. Available: https://bit.ly/3de3klf

[55] D. F. Reed and T. N. Woodard, "Effectiveness of Highway Lighting in Reducing Deer-Vehicle Accidents," *The Journal of Wildlife Management,* vol. 45, no. 3, pp. 721-726, 1981, doi: https://doi.org/10.2307/3808706.

[56] J. A. Ludwig and T. Bremicker, "EVALUATION OF 2.4-M FENCES AND ONE-WAY GATES FOR REDUCING DEER-VEHICLE COLLISIONS IN MINNESOTA," in *Transportation Research Record 913*, Washington, D.C., 1983: Transportation Research Board, National Research Council, pp. 19-22. [Online]. Available: http://onlinepubs.trb.org/Onlinepubs/trr/1983/913/913-006.pdf.

[57] J. A. Curatolo and S. M. Murphy, "The effects of pipelines, roads, and traffic on the movement of caribou, Rangifer tarandus," *Canadian Field-Naturalist,* vol. 100, no. 2, p. 224, 1986. [Online]. Available: https://www.frames.gov/catalog/3857.

[58] B. J. Smith, H. W. Browers, T. E. Dahl, D. E. Nomsen, and K. F. Higgins, "Indirect wetland drainage in association with Federal highway projects in the Prairie Pothole Region," *Wetlands,* journal article vol. 9, no. 1, pp. 27-39, June 01 1989, doi: https://doi.org/10.1007/bf03160766.

[59] H. J. Mader, "Animal habitat isolation by roads and agricultural fields," *Biological Conservation,* vol. 29, no. 1, pp. 81-96, 1984/01/01/ 1984, doi: https://doi.org/10.1016/0006-3207(84)90015-6.

[60] I. M. Mansergh and D. J. Scotts, "Habitat Continuity and Social Organization of the Mountain Pygmy-Possum Restored by Tunnel," *The Journal of Wildlife Management,* vol. 53, no. 3, pp. 701-707, 1989, doi: https://doi.org/10.2307/3809200.

[61] A. F. Reeve and S. H. Anderson, "Ineffectiveness of Swareflex Reflectors at Reducing Deer-Vehicle Collisions," *Wildlife Society Bulletin (1973-2006),* vol. 21, no. 2, pp. 127-132, 1993. [Online]. Available: www.jstor.org/stable/3782912.

[62] M. Yanes, J. M. Velasco, and F. Suárez, "Permeability of roads and railways to vertebrates: The importance of culverts," *Biological Conservation,* vol. 71, no. 3, pp. 217-222, 1995/01/01/ 1995, doi: https://doi.org/10.1016/0006-3207(94)00028-O.

[63] A. Rodriguez, G. Crema, and M. Delibes, "Use of Non-Wildlife Passages Across a High Speed Railway by Terrestrial Vertebrates," *Journal of Applied Ecology,* vol. 33, no. 6, pp. 1527-1540, 1996, doi: https://doi.org/10.2307/2404791.

[64] L. A. Romin and J. A. Bissonette, "Deer: Vehicle Collisions: Status of State Monitoring Activities and Mitigation Efforts," *Wildlife Society Bulletin (1973-2006),* vol. 24, no. 2, pp. 276-283, 1996. [Online]. Available: www.jstor.org/stable/3783118.

[65] M. E. Lehnert and J. A. Bissonette, "Effectiveness of Highway Crosswalk Structures at Reducing Deer-Vehicle Collisions," *Wildlife Society Bulletin (1973-2006),* vol. 25, no. 4, pp. 809-818, 1997. [Online]. Available: www.jstor.org/stable/3783727.

[66] D. J. Decker, K. M. Loconti Lee, and N. A. Connelly, "Incidence and costos of deer-related vehicular accidents in Tompkins County, New York: Insights on an increasingly important aspect of deer management," *HDRU,* vol. 7, no. 89, p. 22, 1990. [Online]. Available: https://core.ac.uk/download/pdf/196231992.pdf.

[67] W. Reh and A. Seitz, "The influence of land use on the genetic structure of populations of the common frog Rana temporaria," *Biological Conservation,* vol. 54, no. 3, pp. 239-249, 1990/01/01/ 1990, doi: https://doi.org/10.1016/0006-3207(90)90054-S.

[68] M. Conover, W. Pitt, K. Kessler, T. DuBow, and W. Sanborn, "Review of human injuries, illnesses, and economic losses caused by wildlife in the United States," *Wildlife Society Bulletin,* vol. 23, pp. 407-414, 01/01 1995. [Online]. Available: https://www.jstor.org/stable/3782947?seq=1.

[69] G. W. T. A. G. Bruinderink and E. Hazebroek, "Ungulate Traffic Collisions in Europe," *Conservation Biology,* vol. 10, no. 4, pp. 1059-1067, 1996. [Online]. Available: http://www.jstor.org/stable/2387142.

[70] S. A. May and T. W. Norton, "Influence of fragmentation and disturbance on the potential impact of feral predators on native fauna in Australian forest ecosystems," *Wildlife Research,* vol. 23, no. 4, pp. 387-400, 1996, doi: https://doi.org/10.1071/WR9960387.

[71] M. R. Conover, "Monetary and Intangible Valuation of Deer in the United States," *Wildlife Society Bulletin (1973-2006),* vol. 25, no. 2, pp. 298-305, 1997. [Online]. Available: www.jstor.org/stable/3783447.

[72] A. Rodriguez, G. Crema, and M. Delibes, "Factors affecting crossing of red foxes and wildcats through non-wildlife passages across a high-speed railway," *Ecography,* vol. 20, no. 3, pp. 287-294, 1997/06/01 1997, doi: https://doi.org/10.1111/j.1600-0587.1997.tb00373.x.

[73] K. Gunther, "Factors Influencing the Frequency of Road-killed Wildlife in Yellowstone National Park," in *International Conference on Wildlife Ecology and Transportation*, Fort Meyers, FL., 1998. [Online]. Available: https://www.researchgate.net/publication/269393716_Factors_Influencing_the_Frequency_of_Road-killed_Wildlife_in_Yellowstone_National_Park.

[74] T. T. F. Richard and L. E. Alexander, "Roads and Their Major Ecological Effects," *Annual Review of Ecology and Systematics,* vol. 29, pp. 207-C2, 1998. [Online]. Available: www.jstor.org/stable/221707.

[75] B. Iuell *et al.*, *Wildlife and Traffic: A European Handbook for Identifying Conflicts and Designing Solutions*. European Co-operation in the Field of Scientific and Technical Research, 2003. [Online]. Available: http://www.iene.info/wp-content/uploads/COST341_Handbook.pdf.

[76] J. D. Quintero, *Guía de buenas prácticas para carreteras ambientalmente amigables*. The Nature Conservancy, 2016, p. 104. [Online]. Available: http://fcds.org.co/site/wp-content/uploads/2018/09/carreteras-ambientalmente-amigables_WEB_02_2016-1.pdf.

[77] E. Pomareda García, D. Araya Gamboa, Y. Ríos Montero, E. Arévalo Huezo, M. C. Aguilar Ruiz, and R. M. Menacho Odio, *Guía Ambiental "Vías Amigables con la Vida Silvestre"*. Comité Científico de la Comisión Vías y Vida Silvestre., 2014, p.

75. [Online]. Available: https://www.researchgate.net/publication/307946704_Guia_Ambiental_Vias_Amigables_con_la_Vida_Silvestre_Environmental_Guide_Wildlife_Friendly_Roads.

[78] Ministerio de Medio Ambiente y Medio Rural y Marino, *Indicadores de fragmentación de hábitats causada por infraestructuras lineales de transporte*. Madrid: Organismo Autónomo Parques Nacionales (in Spanish), 2010. [Online]. Available: https://www.miteco.gob.es/es/biodiversidad/publicaciones/4_indicadores_fragmentac_habitat_tcm30-195795.pdf.

[79] A. y. M. A. Ministerio de Agricultura, *Prescripciones técnicas para el diseño de pasos de fauna y vallados perimetrales*. Madrid, España, 2015. [Online]. Available: https://www.mitma.gob.es/recursos_mfom/PrescripPasosFaunaVallados.pdf.

[80] A. Clevenger and M. Huijser, "Wildlife Crossing Structure Handbook, Design and Evaluation in North America," FHWA-CFL/TD-11-003, March 2011 2011. [Online]. Available: https://roadecology.ucdavis.edu/files/content/projects/DOT-FHWA_Wildlife_Crossing_Structures_Handbook.pdf

[81] A. Bager, *Infraestrutura Viária & Biodiversidade: Métodos e Diagnósticos*. 2018. [Online]. Available: https://www.amazon.ca/Infraestrutura-Vi%C3%A1ria-Biodiversidade-Diagn%C3%B3sticos-Portuguese-ebook/dp/B086BTCBGP.

[82] J. Golay and M. Kanevski, "A new estimator of intrinsic dimension based on the multipoint Morisita index," *Pattern Recognition,* vol. 48, pp. 4070-4081, 2015, doi: https://doi.org/10.1016/j.patcog.2015.06.010.

[83] Centers for Disease Control and Prevention. "John Snow: A Legacy of Disease Detectives." Centers for Disease Control and Prevention. [Online]. Available: https://blogs.cdc.gov/publichealthmatters/2017/03/a-legacy-of-disease-detectives/

[84] A. S. Fotheringham and A. P. Rogerson, *Spatial Analysis* (The SAGE Handbook of). SAGE Publications Ltd, 2009, pp. 1-4. [Online]. Available: https://uk.sagepub.com/en-gb/eur/the-sage-handbook-of-spatial-analysis/book227940.

[85] NASA, "Sputnik 1," *NASA,* 2015. [Online]. Available: https://www.nasa.gov/multimedia/imagegallery/image_feature_924.html.

[86] N. Andrienko, G. Andrienko, and S. Rinzivillo, "Leveraging spatial abstraction in traffic analysis and forecasting with visual analytics," *Information Systems,* vol. 57, pp. 172-194, 2016/04/01/ 2016, doi: https://doi.org/10.1016/j.is.2015.08.007.

[87] C. Li *et al.*, "Using NDVI percentiles to monitor real-time crop growth," *Computers and Electronics in Agriculture,* vol. 162, pp. 357-363, 2019/07/01/ 2019, doi: https://doi.org/10.1016/j.compag.2019.04.026.

[88] C. Grilo, J. A. Bissonette, and M. Santos-Reis, "Spatial–temporal patterns in Mediterranean carnivore road casualties: Consequences for mitigation," *Biological Conservation,* vol. 142, no. 2, pp. 301-313, 2009/02/01/ 2009, doi: https://doi.org/10.1016/j.biocon.2008.10.026.

[89] T. L. Joyce and S. P. Mahoney, "Spatial and Temporal Distributions of Moose-Vehicle Collisions in Newfoundland," *Wildlife Society Bulletin (1973-2006),* vol. 29, no. 1, pp. 281-291, 2001. [Online]. Available: www.jstor.org/stable/3784010.

[90] T. Kantola, J. L. Tracy, K. A. Baum, M. A. Quinn, and R. N. Coulson, "Spatial risk assessment of eastern monarch butterfly road mortality during autumn migration within the southern corridor," *Biological Conservation,* vol. 231, pp. 150-160, 2019/03/01/ 2019, doi: https://doi.org/10.1016/j.biocon.2019.01.008.

[91] A. P. Kirilenko, S. O. Stepchenkova, and J. M. Hernandez, "Comparative clustering of destination attractions for different origin markets with network and spatial analyses of online reviews," *Tourism Management,* vol. 72, pp. 400-410, 2019, doi: https://doi.org/10.1016/j.tourman.2019.01.001.

[92] N. Sillero, "Amphibian mortality levels on Spanish country roads: descriptive and spatial analysis," (in English), vol. 29, no. 3, p. 337, 2008, doi: https://doi.org/10.1163/156853808785112066.

[93] L. O. Gonçalves *et al.*, "Reptile road-kills in Southern Brazil: Composition, hot moments and hotspots," *Science of The Total Environment,* vol. 615, pp. 1438-1445, 2018/02/15/ 2018, doi: https://doi.org/10.1016/j.scitotenv.2017.09.053.

[94] D. Ramp, J. Caldwell, K. A. Edwards, D. Warton, and D. B. Croft, "Modelling of wildlife fatality hotspots along the Snowy Mountain Highway in New South Wales, Australia," *Biological Conservation,* vol. 126, no. 4, pp. 474-490, 2005/12/01/ 2005, doi: https://doi.org/10.1016/j.biocon.2005.07.001.

[95] F. Z. Teixeira *et al.*, "Are Road-Kill Hotspots Coincident among Different Vertebrate Groups?," *Oecologia Australis,* vol. 17, no. 1, pp. 36-47, 2013, doi: http://doi.org/10.4257/oeco.2013.1701.04.

[96] F. Zimmermann Teixeira, A. Kindel, S. M. Hartz, S. Mitchell, and L. Fahrig, "When road-kill hotspots do not indicate the best sites for road-kill mitigation," *Journal of Applied Ecology,* vol. 54, no. 5, pp. 1544-1551, 2017, doi: https://www.doi.org/10.1111/1365-2664.12870.

[97] W. F. Laurance, "Bad Roads, Good Roads," in *Handbook of Road Ecology*, 2015, pp. 10-15. doi: https://doi.org/10.1002/9781118568170.ch2.

[98] R. van der Ree, J. A. G. Jaeger, E. A. van der Grift, and A. P. Clevenger, "Effects of Roads and Traffic on Wildlife Populations and Landscape Function: Road Ecology is Moving toward Larger Scales," *Ecology and Society; Vol. 16, No. 1 (2011),* 01/01 2011. [Online]. Available: http://www.ecologyandsociety.org/vol16/iss1/art48/.

[99] D. J. Smith and R. van der Ree, "Field Methods to Evaluate the Impacts of Roads on Wildlife," in *Handbook of Road Ecology*, 2015, pp. 82-95. doi: https://doi.org/10.1002/9781118568170.ch11.

[100] P. Sunnucks and N. Balkenhol, "Incorporating Landscape Genetics into Road Ecology," in *Handbook of Road Ecology*, 2015, pp. 110-118. doi: https://doi.org/10.1002/9781118568170.ch14.

[101] Y. E. Chee, "Principles Underpinning Biodiversity Offsets and Guidance on their Use," in *Handbook of Road Ecology*, 2015, pp. 51-59. doi: https://doi.org/10.1002/9781118568170.ch7.

[102] K. Gunson and F. Z. Teixeira, "Road–Wildlife Mitigation Planning can be Improved by Identifying the Patterns and Processes Associated with Wildlife-

Vehicle Collisions," in *Handbook of Road Ecology*, 2015, pp. 101-109. doi: https://www.doi.org/10.1002/9781118568170.ch13.

[103] N. Selva, A. Switalski, S. Kreft, and P. L. Ibisch, "Why Keep Areas Road-Free? The Importance of Roadless Areas," in *Handbook of Road Ecology*, 2015, pp. 16-26. doi: https://doi.org/10.1002/9781118568170.ch3.

[104] A. P. Clevenger, B. Chruszcz, and K. E. Gunson, "Spatial patterns and factors influencing small vertebrate fauna road-kill aggregations," *Biological Conservation,* vol. 109, no. 1, pp. 15-26, 2003/01/01/ 2003, doi: https://doi.org/10.1016/S0006-3207(02)00127-1.

[105] A. Bager, P. d. S. Lucas, A. Bourscheit, A. Kuczach, and B. Maia, "Os Caminhos da Conservação da Biodiversidade Brasileira frente aos Impactos da Infraestrutura Viária," *Biodiversidade Brasileira,* vol. 6, no. 1, p. 12, 2016. [Online]. Available: https://www.icmbio.gov.br/revistaeletronica/index.php/BioBR/article/download/530/456.

[106] A. V. Coelho, I. P. Coelho, A. Kindel, and F. Z. Teixeira, *Road mortality software Siriema: road mortality software. User's Manual V. 2.0.*, Porto Alegre - Brazil: Universidade Federal do Rio Grande do Sul, 2014, p. 34. [Online]. Available: http://www.ufrgs.br/siriema/index.php?lang=en.

[107] F. Z. Teixeira, A. V. P. Coelho, I. B. Esperandio, and A. Kindel, "Vertebrate road mortality estimates: Effects of sampling methods and carcass removal," *Biological Conservation,* vol. 157, pp. 317-323, 2013/01/01/ 2013, doi: https://doi.org/10.1016/j.biocon.2012.09.006.

[108] J. Parchizadeh *et al.*, "Roads threaten Asiatic cheetahs in Iran," *Current Biology,* vol. 28, no. 19, pp. R1141-R1142, 2018/10/08/ 2018, doi: https://doi.org/10.1016/j.cub.2018.09.005.

[109] D. C. Wilkins, K. M. Kockelman, and N. Jiang, "Animal-vehicle collisions in Texas: How to protect travelers and animals on roadways," *Accident Analysis & Prevention,* vol. 131, pp. 157-170, 2019/10/01/ 2019, doi: https://doi.org/10.1016/j.aap.2019.05.030.

[110] J. M. Kolowski and C. K. Nielsen, "Using Penrose distance to identify potential risk of wildlife–vehicle collisions," *Biological Conservation,* vol. 141, no. 4, pp. 1119-1128, 2008/04/01/ 2008, doi: https://doi.org/10.1016/j.biocon.2008.02.011.

[111] R. A. L. Santos, M. Mota-Ferreira, L. M. S. Aguiar, and F. Ascensão, "Predicting wildlife road-crossing probability from roadkill data using occupancy-detection models," *Science of The Total Environment,* vol. 642, pp. 629-637, 2018/11/15/ 2018, doi: https://doi.org/10.1016/j.scitotenv.2018.06.107.

[112] R. R. Jensen, R. A. Gonser, and C. Joyner, "Landscape factors that contribute to animal–vehicle collisions in two northern Utah canyons," *Applied Geography,* vol. 50, pp. 74-79, 2014/06/01/ 2014, doi: https://doi.org/10.1016/j.apgeog.2014.02.007.

[113] F. Ascensão, D. Yogui, M. Alves, E. P. Medici, and A. Desbiez, "Predicting spatiotemporal patterns of road mortality for medium-large mammals," *J. Environ. Manage.,* vol. 248, p. 109320, 2019/10/15/ 2019, doi: https://doi.org/10.1016/j.jenvman.2019.109320.

[114] M. P. Huijser *et al.*, "Effectiveness of short sections of wildlife fencing and crossing structures along highways in reducing wildlife–vehicle collisions and providing safe crossing opportunities for large mammals," *Biological Conservation,* vol. 197, pp. 61-68, 2016/05/01/ 2016, doi: https://doi.org/10.1016/j.biocon.2016.02.002.

[115] R. A. Finder, J. L. Roseberry, and A. Woolf, "Site and landscape conditions at white-tailed deer/vehicle collision locations in Illinois," *Landscape and Urban Planning,* vol. 44, no. 2, pp. 77-85, 1999/05/10/ 1999, doi: https://doi.org/10.1016/S0169-2046(99)00006-7.

[116] H. Gundersen and H. P. Andreassen, "The risk of moose Alces alces collision: A predictive logistic modelfor moose-train accidents," *Wildlife Biology,* vol. 4, no. 2, pp. 103-110, 8, 1998, doi: https://doi.org/10.2981/wlb.1998.007.

[117] C. F. Jaarsma, F. van Langevelde, and H. Botma, "Flattened fauna and mitigation: Traffic victims related to road, traffic, vehicle, and species characteristics," *Transportation Research Part D: Transport and Environment,*

vol. 11, no. 4, pp. 264-276, 2006/07/01/ 2006, doi: https://doi.org/10.1016/j.trd.2006.05.001.

[118] R. Found and M. S. Boyce, "Predicting deer–vehicle collisions in an urban area," *J. Environ. Manage.,* vol. 92, no. 10, pp. 2486-2493, 2011/10/01/ 2011, doi: https://doi.org/10.1016/j.jenvman.2011.05.010.

[119] Y. Lao, G. Zhang, Y.-J. Wu, and Y. Wang, "Modeling animal–vehicle collisions considering animal–vehicle interactions," *Accident Analysis & Prevention,* vol. 43, no. 6, pp. 1991-1998, 2011/11/01/ 2011, doi: https://doi.org/10.1016/j.aap.2011.05.017.

[120] L. Nelli, J. Langbein, P. Watson, and R. Putman, "Mapping risk: Quantifying and predicting the risk of deer-vehicle collisions on major roads in England," *Mammalian Biology,* vol. 91, pp. 71-78, 2018/07/01/ 2018, doi: https://doi.org/10.1016/j.mambio.2018.03.013.

[121] M. J. Gould, W. R. Gould, J. W. Cain, and G. W. Roemer, "Validating the performance of occupancy models for estimating habitat use and predicting the distribution of highly-mobile species: A case study using the American black bear," *Biological Conservation,* vol. 234, pp. 28-36, 2019/06/01/ 2019, doi: https://doi.org/10.1016/j.biocon.2019.03.010.

[122] A. Seiler, "Predicting locations of moose-vehicle collisions in Sweden," *Journal of Applied Ecology,* vol. 42, no. 2, pp. 371-382, 2005, doi: https://doi.org/10.1111/j.1365-2664.2005.01013.x.

[123] R. Wynn-Grant, J. R. Ginsberg, C. W. Lackey, E. J. Sterling, and J. P. Beckmann, "Risky business: Modeling mortality risk near the urban-wildland interface for a large carnivore," *Global Ecology and Conservation,* vol. 16, p. e00443, 2018/10/01/ 2018, doi: https://doi.org/10.1016/j.gecco.2018.e00443.

[124] G. Chirici *et al.*, "A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data," *Remote Sensing of Environment,* vol. 176, pp. 282-294, 2016/04/01/ 2016, doi: https://doi.org/10.1016/j.rse.2016.02.001.

[125] M. Hosseinalizadeh, N. Kariminejad, O. Rahmati, S. Keesstra, M. Alinejad, and A. Mohammadian Behbahani, "How can statistical and artificial intelligence approaches predict piping erosion susceptibility?," *Science of The Total Environment,* vol. 646, pp. 1554-1566, 2019/01/01/ 2019, doi: https://doi.org/10.1016/j.scitotenv.2018.07.396.

[126] J. Xu, G. Feng, T. Zhao, X. Sun, and M. Zhu, "Remote sensing image classification based on semi-supervised adaptive interval type-2 fuzzy c-means algorithm," *Computers & Geosciences,* vol. 131, pp. 132-143, 2019/10/01/ 2019, doi: https://doi.org/10.1016/j.cageo.2019.06.005.

[127] F. Harrell, "Classification vs. Prediction," (in English), *Statistical Thinking,* 2020. [Online]. Available: https://www.fharrell.com/post/classification.

[128] M. Fabrizio, M. Di Febbraro, and A. Loy, "Where will it cross next? Optimal management of road collision risk for otters in Italy," *J. Environ. Manage.,* vol. 251, p. 109609, 2019/12/01/ 2019, doi: https://doi.org/10.1016/j.jenvman.2019.109609.

[129] H. Ha and F. Shilling, "Modelling potential wildlife-vehicle collisions (WVC) locations using environmental factors and human population density: A case-study from 3 state highways in Central California," *Ecological Informatics,* vol. 43, 10/18 2017, doi: https://www.doi.org/10.1016/j.ecoinf.2017.10.005.

[130] T. R. Board and N. R. Council, *Assessing and Managing the Ecological Impacts of Paved Roads*. Washington, DC: The National Academies Press (in English), 2005, p. 324. doi: https://doi.org/10.17226/11535.

[131] N. J. Silvy, *The Wildlife Techniques Manual: Volume 1: Research. Volume 2: Management 2-vol. Set*. Johns Hopkins University Press, 2012. [Online]. Available: https://books.google.com.co/books?id=PL2IHTdzSeAC.

[132] S. M. Santos, R. Lourenço, A. Mira, and P. Beja, "Relative Effects of Road Risk, Habitat Suitability, and Connectivity on Wildlife Roadkills: The Case of Tawny Owls (Strix aluco)," *PLoS One,* vol. 8, no. 11, p. e79967, 2013, doi: https://doi.org/10.1371/journal.pone.0079967.

[133] B. D. Ripley, *Spatial Statistics*, First ed. (Wiley series in probability and mathematical statistics). London: Wiley, 1981, p. 268. doi: https://doi.orgx.

[134] N. A. C. Cressie, *Statistics for spatial data* (Probability and mathematical statistics). U.S.A: J. Wiley, 1993. [Online]. Available: https://books.google.com.co/books?id=4SdRAAAAMAAJ.

[135] B. D. Ripley, *Spatial statistics*, Second ed. (Wiley series in probability and statistics). Hoboken: Wiley, 2004, p. 268. [Online]. Available: https://www.wiley.com/en-co/Spatial+Statistics-p-9780471691167.

[136] A. V. Coelho, I. P. Coelho, A. Kindel, and F. Z. Teixeira, *Road mortality software Siriema: road mortality software. User's Manual V. 2.0*, Porto Alegre - Brazil: Universidade Federal do Rio Grande do Sul, 2014, p. 34. [Online]. Available: https://www.researchgate.net/profile/Andreas_Kindel/publication/319069206_Siriema_road_mortality_software_User's_Manual_V_20/links/598e408baca2721d9b4d501f/Siriema-road-mortality-software-Users-Manual-V-20.pdf.

[137] Y. Chun and D. Griffith, *Spatial Statistics and Geostatistics: Basic Concepts*, First Edit ed. (Theory and Applications for Geographic Information Science and Technology). California: SAGE, 2013, pp. 1-16. doi: https://www.doi.org/10.1007/978-3-319-23519-6_1650-1.

[138] Esri Co. "Understanding Euclidean distance analysis—Help | ArcGIS for Desktop." [Online]. Available: https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/understanding-euclidean-distance-analysis.htm#ESRI_SECTION1_29048F6D811B40D0A0B7E2BA0F36E92E

[139] D. Massonnet and K. L. Feigl, "Radar interferometry and its application to changes in the Earth's surface," *Reviews of Geophysics*, vol. 36, no. 4, pp. 441-500, 1998/11/01 1998, doi: https://doi.org/10.1029/97RG03139.

[140] USGS, "Synthetic Aperture Radar (SAR) Processing System," ed, 2020. [Online]. Available: https://www.usgs.gov/centers/eros/science/usgs-eros-archive-radar-synthetic-aperture-radar-sar-processing-system?qt-science_center_objects=0#qt-science_center_objects

[141] K. G. Nikolakopoulos, E. K. Kamaratakis, and N. Chrysoulakis, "SRTM vs ASTER elevation products. Comparison for two regions in Crete, Greece," *Int. J. Remote Sens.*, vol. 27, no. 21, pp. 4819-4838, 2006/11/01 2006, doi: https://doi.org/10.1080/01431160600835853.

[142] USGS. "Landsat 8 Data Users Handbook." [Online]. Available: https://www.usgs.gov/media/files/landsat-8-data-users-handbook

[143] R. D. Jackson, "Spectral indices in N-Space," *Remote Sensing of Environment,* vol. 13, no. 5, pp. 409-421, 1983/11/01/ 1983, doi: https://doi.org/10.1016/0034-4257(83)90010-X.

[144] M. E. Arboit and D. S. Maglione, "The current situation and recent changes in vegetation indices (VIS) in forested cities with dry climates: the case of the Mendoza metropolitan area, Argentina," (in Spanish), *Urbano,* Article vol. 21, no. 38, pp. 18-35, Nov 2018, doi: 10.22320/07183607.2018.21.38.02.

[145] A. A. Gitelson, Y. J. Kaufman, and M. N. Merzlyak, "Use of a green channel in remote sensing of global vegetation from EOS-MODIS," *Remote Sensing of Environment,* vol. 58, no. 3, pp. 289-298, 1996/12/01/ 1996, doi: https://doi.org/10.1016/S0034-4257(96)00072-7.

[146] USGS, "Landsat Enhanced Vegetation Index," ed, 2020. [Online]. Available: https://www.usgs.gov/land-resources/nli/landsat/landsat-enhanced-vegetation-index?qt-science_support_page_related_con=0#qt-science_support_page_related_con

[147] Z. Azizi, A. Najafi, and H. Sohrabi, *FOREST CANOPY DENSITY ESTIMATING, USING SATELLITE IMAGES*. 2014. [Online]. Available: https://www.isprs.org/proceedings/XXXVII/congress/8_pdf/11_WG-VIII-11/21.pdf.

[148] J. Baynes, "Assessing forest canopy density in a highly variable landscape using Landsat data and FCD Mapper software," *Australian Forestry,* vol. 67, no. 4, pp. 247-253, 2004/01/01 2004, doi: https://doi.org/10.1080/00049158.2004.10674942.

[149] A. Rikimaru, P. S. Roy, and S. Miyatake, "Tropical forest cover density mapping," *Tropical Ecology,* vol. 43, no. (1), pp. 39-47, 2002 2002. [Online]. Available: http://www.tropecol.com/pdf/open/PDF_43_1/43104.pdf.

[150] G. Rondeaux, M. Steven, and F. Baret, "Optimization of soil-adjusted vegetation indices," *Remote Sensing of Environment,* vol. 55, no. 2, pp. 95-107, 1996/02/01/ 1996, doi: https://doi.org/10.1016/0034-4257(95)00186-7.

[151] B.-c. Gao, "NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space," *Remote Sensing of Environment,* vol. 58, no. 3, pp. 257-266, 1996/12/01/ 1996, doi: https://doi.org/10.1016/S0034-4257(96)00067-3.

[152] C. H. Key, N. Benson, D. Ohlen, S. Howard, R. McKinley, and Z. Z., "The normalized burn ratio and relationships to burn severity: ecology, remote sensing and implementation " in *Proceedings of the Ninth Forest Service Remote Sensing Applications Conference. American Society for Photogrammetry and Remote Sensing, Bethesda, MD* San Diego, CA A. S. f. P. a. R. Sensing, Ed., 8-12 April, 2002 / 2002 2002. [Online]. Available: https://www.yumpu.com/en/document/read/24226870/the-normalized-burn-ratio-and-relationships-to-burn-severity-.

[153] E. R. Hunt and B. N. Rock, "Detection of changes in leaf water content using Near- and Middle-Infrared reflectances," *Remote Sensing of Environment,* vol. 30, no. 1, pp. 43-54, 1989/10/01/ 1989, doi: https://doi.org/10.1016/0034-4257(89)90046-1.

[154] A. A. Gitelson, A. Viña, T. J. Arkebauer, D. C. Rundquist, G. Keydan, and B. Leavitt, "Remote estimation of leaf area index and green leaf biomass in maize canopies," *Geophys. Res. Lett.,* vol. 30, no. 5, 2003/03/01 2003, doi: https://doi.org/10.1029/2002GL016450.

[155] Faqs.org, "Should I normalize/standardize/rescale the," ed, 2020. [Online]. Available: http://www.faqs.org/faqs/ai-faq/neural-nets/part2/section-16.html

[156] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, pp. 2825-2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html.

[157] P. E. Latham and Y. Roudi, "Mutual information," *Scholarpedia,* vol. 4, no. 1, pp. 1658-1658, 2009, doi: https://doi.org/10.4249/scholarpedia.1658.

[158] C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal,* vol. 27, no. 3, pp. 379-423, 1948/07/01 1948, doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[159] B. C. Ross, "Mutual Information between Discrete and Continuous Data Sets," *PLoS One,* vol. 9, no. 2, p. e87357, 2014, doi: https://doi.org/10.1371/journal.pone.0087357.

[160] G. E. P. Box, "NON-NORMALITY AND TESTS ON VARIANCES," *Biometrika,* vol. 40, no. 3-4, pp. 318-335, 1953, doi: https://doi.org/10.1093/biomet/40.3-4.318.

[161] H. Haibo, B. Yang, E. A. Garcia, and L. Shutao, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1-8 June 2008 2008, pp. 1322-1328, doi: https://doi.org/10.1109/IJCNN.2008.4633969.

[162] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation method," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 535-539, 1997, doi: https://doi.org/10.1109/34.589216.

[163] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, no. Vol. 16 (2002), Jun 1, 2002 2002, doi: https://doi.org/10.1613/jair.953.

[164] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning," in *Advances in Intelligent Computing*, Berlin, Heidelberg, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., 2005 2005: Springer Berlin Heidelberg, pp. 878-887, doi: https://doi.org/10.1007/11538059_91.

[165] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences,* vol. 465, pp. 1-20, 2018, doi: https://doi.org/10.1016/j.ins.2018.06.056.

[166] H. M. Nguyen, E. W. Cooper, and K. Kamei, "Borderline over-sampling for imbalanced data classification," *IJKESDP,* vol. 3, pp. 4-21, 2011, doi: https://doi.org/10.1504/IJKESDP.2011.039875.

[167] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition*. Packt Publishing, 2017. [Online]. Available: https://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow-ebook/dp/B0742K7HYF.

[168] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995/09/01 1995, doi: https://doi.org/10.1007/BF00994018.

[169] T. Hofmann, B. Scholkopf, and A. J. Smola, "Kernel methods in machine learning," (in en), *Ann. Statist.,* vol. 36, no. 3, pp. 1171-1220, 2008/06 2008, doi: https://doi.org/10.1214/009053607000000677.

[170] R. A. Fisher, "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS," *Annals of Eugenics,* vol. 7, no. 2, pp. 179-188, 1936/09/01 1936, doi: https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[171] S. Haykin, *Neural Networks: A Comprehensive Foundation (3rd Edition)*. Prentice-Hall, Inc., 2007. [Online]. Available: http://dai.fmph.uniba.sk/courses/NN/haykin.neural-networks.3ed.2009.pdf.

[172] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review,* vol. 65, no. 6, pp. 386-408, 1958, doi: https://doi.org/10.1037/h0042519.

[173] E. R. David and L. M. James, *Parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models*. MIT Press, 1986. [Online]. Available: https://www.researchgate.net/profile/James_Mcclelland/publication/243672900_A_General_Framework_for_Parallel_Distributed_Processing/links/00b7d531ea5f9d6928000000/A-General-Framework-for-Parallel-Distributed-Processing.pdf.

[174] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment,* vol. 62, no. 1, pp. 77-89, 1997/10/01/ 1997, doi: https://doi.org/10.1016/S0034-4257(97)00083-7.

[175] A. Ben-David, "A lot of randomness is hiding in accuracy," *Engineering Applications of Artificial Intelligence,* vol. 20, no. 7, pp. 875-885, 2007/10/01/ 2007, doi: https://doi.org/10.1016/j.engappai.2007.01.001.

[176] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer, 2018. doi: https://doi.org/10.1007/978-3-319-98074-4.

[177] J. A. Swets, "Measuring the accuracy of diagnostic systems," (in eng), *Science,* vol. 240, no. 4857, pp. 1285-93, Jun 3 1988, doi: https://doi.org/10.1126/science.3287615.

[178] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition,* vol. 30, no. 7, pp. 1145-1159, 1997/07/01/ 1997, doi: https://doi.org/10.1016/S0031-3203(96)00142-2.

[179] J. H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. MIT Press, 1992. [Online]. Available: https://ieeexplore.ieee.org/book/6267401.

[180] O. Kramer, *Genetic Algorithm Essentials* (Studies in Computational Intelligence). 2017. doi: https://doi.org/10.1007/978-3-319-52156-5.

[181] G. Syswerda, "Simulated Crossover in Genetic Algorithms," in *Foundations of Genetic Algorithms*, vol. 2, L. D. Whitley Ed.: Elsevier, 1993, pp. 239-255. doi: https://doi.org/10.1016/B978-0-08-094832-4.50021-0.

[182] O. Kramer, "Evolutionary self-adaptation: a survey of operators and strategy parameters," *Evolutionary Intelligence,* vol. 3, no. 2, pp. 51-65, 2010/08/01 2010, doi: https://doi.org/10.1007/s12065-010-0035-y.

[183] J. C. Jaramillo, J. L. González Manosalva, M. M. Velásquez López, C. Correa-Ayram, and P. Isaacs-Cubides., "Los animales atropellados de Colombia:

Estrategias para mitigar los efectos de la infraestructura vial.," in *Biodiversidad 2017. Estado y tendencias de la biodiversidad continental de Colombia.*, vol. 2017, L. A. Moreno, C. Rueda, and G. I. Andrade Eds., (Reporte BIO. Bogotá, D. C., Colombia.: Instituto de Investigación de Recursos Biológicos Alexander von Humboldt., 2018, ch. 12. [Online]. Available: http://reporte.humboldt.org.co/biodiversidad/2017/cap2/206/index.html#seccion12.

[184] Cámara de Comercio Oriente Antioqueño, "El Oriente Antioqueño duplica el crecimiento de Antioquia y Colombia en la creación de empresas," ed, 2020. [Online]. Available: https://www.ccoa.org.co/noticia/el-oriente-antioqueno-duplica-el-crecimiento-de-antioquia-y-colombia-en-la-creacion-de-empresas

[185] Y. García-Morera and L. Giraldo-Iral, "Recopilación de Información de FAUNA en la Jurisdicción de CORNARE, hasta el año 2015 v1.1.," C. A. R. d. l. c. d. l. r. N. y. N.-. CORNARE., Ed., ed, 2018. [Online]. Available: https://ipt.biodiversidad.co/sib/resource?r=cornarefauna2015#downloads

[186] Recosfa, "Red Colombiana de Seguimiento de Fauna Atropellada," ed, 2019. [Online]. Available: http://www.recosfa.com

[187] Esri Co. "How Spatial Autocorrelation Works." Esri. [Online]. Available: http://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm

[188] DANE, "Geoportal DANE - Descarga del Marco Geoestadistico Nacional (MGN)," ed, 2020. [Online]. Available: https://geoportal.dane.gov.co/servicios/descarga-y-metadatos/descarga-mgn-marco-geoestadistico-nacional

[189] Global Forest Watch, "Forest Monitoring, Land Use & Deforestation Trends," ed, 2020. [Online]. Available: https://www.globalforestwatch.org

[190] IDEAM, "Catálogo de mapas - IDEAM," ed, 2020. [Online]. Available: http://www.siac.gov.co/catalogo-de-mapas

[191] IDEAM. "Mapa ecosistemas continentales, costeros y marinos." [Online]. Available: http://www.ideam.gov.co/web/ecosistemas/mapa-ecosistemas-continentales-costeros-marinos

[192] IGAC, "Datos Abiertos Agrología | GEOPORTAL," ed, 2020. [Online]. Available: https://geoportal.igac.gov.co/contenido/datos-abiertos-agrologia

[193] U.S. Geological Service, "EarthExplorer," ed: USGS - U.S. Geological Survey, 2020. [Online]. Available: https://earthexplorer.usgs.gov

[194] S. R. Chavan and V. V. Srinivas, "Effect of DEM source on equivalent Horton–Strahler ratio based GIUH for catchments in two Indian river basins," *Journal of Hydrology,* vol. 528, pp. 463-489, 2015/09/01/ 2015, doi: https://doi.org/10.1016/j.jhydrol.2015.06.049.

[195] Google Co., "Google Earth Engine," ed, 2020. [Online]. Available: https://earthengine.google.com

[196] P. J. Isaacs-Cubides, L. Trujillo, and V. Jaimes, "Zonificación de alternativas de conectividad ecológica, restauración y conservación en las microcuencas Curubital, Mugroso, Chisacá y Regadera, cuenca del río Tunjuelo (Distrito Capital de Bogotá), Colombia," *Biota Colombiana,* vol. 18, no. 2017, 1, pp. 70 - 88, 2017. [Online]. Available: http://repository.humboldt.org.co/handle/20.500.11761/32531.

[197] B. G. Dickson *et al.*, "Circuit-theory applications to connectivity science and conservation," *Conservation Biology,* vol. 33, no. 2, pp. 239-249, 2019/04/01 2019, doi: http://doi.org/10.1111/cobi.13230.

[198] M. Lucherini, M. Hoffmann, and C. Sillero-Zubiri. "IUCN Red List of Threatened Species: Crab-eating Fox." [Online]. Available: https://www.iucnredlist.org/species/4248/81266293#external-data

[199] S. H. Siyal, D. Mentis, and M. Howells, "Mapping key economic indicators of onshore wind energy in Sweden by using a geospatial methodology," *Energy Conversion and Management,* vol. 128, pp. 211-226, 2016/11/15/ 2016, doi: https://doi.org/10.1016/j.enconman.2016.09.055.

[200] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media,

2017. [Online]. Available: https://books.google.com.co/books?id=khpYDgAAQBAJ.

[201] W. R. Tobler, "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic Geography,* vol. 46, pp. 234-240, 1970, doi: http://doi.org/10.2307/143141.

[202] T. Pan, J. Zhao, W. Wu, and J. Yang, "Learning imbalanced datasets based on SMOTE and Gaussian distribution," *Information Sciences,* vol. 512, pp. 1214-1233, 2020/02/01/ 2020, doi: https://doi.org/10.1016/j.ins.2019.10.048.

[203] L. Peng, R. Niu, B. Huang, X. Wu, Y. Zhao, and R. Ye, "Landslide susceptibility mapping based on rough set theory and support vector machines: A case of the Three Gorges area, China," *Geomorphology,* vol. 204, pp. 287-301, 2014/01/01/ 2014, doi: https://doi.org/10.1016/j.geomorph.2013.08.013.

[204] R. S. Olson, N. Bartley, R. J. Urbanowicz, and J. H. Moore, "Evaluation of a Tree-based Pipeline Optimization Tool for Automating Data Science," presented at the Proceedings of the Genetic and Evolutionary Computation Conference 2016, Denver, Colorado, USA, 2016. doi: https://doi.org/10.1145/2908812.2908918.

[205] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics,* vol. 36, no. 1, pp. 250-256, 2019, doi: https://doi.org/10.1093/bioinformatics/btz470.

[206] M. A. Morid, A. Borjali, and G. Del Fiol, "A scoping review of transfer learning research on medical image analysis using ImageNet," *Computers in Biology and Medicine,* vol. 128, p. 104115, 2021/01/01/ 2021, doi: https://doi.org/10.1016/j.compbiomed.2020.104115.

[207] Parques Nacionales Naturales de Colombia, "RUNAP," ed, 2020. [Online]. Available: https://runap.parquesnacionales.gov.co

[208] T. R. Hoens and N. V. Chawla, "Imbalanced Datasets: From Sampling to Classifiers," in *Imbalanced Learning*, (Wiley Online Books, 2013, pp. 43-59. [Online]. Available: https://doi.org/10.1002/9781118646106.ch3.

[209] N. Japkowicz, "Assessment Metrics for Imbalanced Learning," in *Imbalanced Learning*, (Wiley Online Books, 2013, pp. 187-206. [Online]. Available: https://doi.org/10.1002/9781118646106.ch8.

[210] S. Jain, S. Shukla, and R. Wadhvani, "Dynamic selection of normalization techniques using data complexity measures," *Expert Syst Appl,* vol. 106, pp. 252-262, 2018/09/15/ 2018, doi: https://doi.org/10.1016/j.eswa.2018.04.008.

[211] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," *Journal of King Saud University - Computer and Information Sciences,* vol. 32, no. 2, pp. 225-231, 2020/02/01/ 2020, doi: https://doi.org/10.1016/j.jksuci.2018.05.010.

[212] W. Qian, J. Huang, Y. Wang, and W. Shu, "Mutual information-based label distribution feature selection for multi-label learning," *Knowledge-Based Systems,* vol. 195, p. 105684, 2020/05/11/ 2020, doi: https://doi.org/10.1016/j.knosys.2020.105684.

[213] S. Güneş, K. Polat, and Ş. Yosunkaya, "Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome," *Expert Syst Appl,* vol. 37, no. 2, pp. 998-1004, 2010/03/01/ 2010, doi: https://doi.org/10.1016/j.eswa.2009.05.075.

[214] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001/10/01 2001, doi: http://doi.org/10.1023/A:1010933404324.

[215] A. Getis and J. K. Ord, "The Analysis of Spatial Association by Use of Distance Statistics," *Geographical Analysis,* vol. 24, no. 3, pp. 189-206, 1992, doi: https://doi.org/10.1111/j.1538-4632.1992.tb00261.x.

[216] A. Getis and J. K. Ord, "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application," *Geographical Analysis,* vol. 27, no. 4, p. 21, 1995, doi: https://doi.org/10.1111/j.1538-4632.1995.tb00912.x.

[217] P. Drotár, J. Gazda, and Z. Smékal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Computers in Biology and*

Medicine, vol. 66, pp. 1-10, 2015/11/01/ 2015, doi: https://doi.org/10.1016/j.compbiomed.2015.08.010.

[218] H. Zhou, X. Wang, and Y. Zhang, "Feature selection based on weighted conditional mutual information," *Applied Computing and Informatics,* 2020/01/07/ 2020, doi: https://doi.org/10.1016/j.aci.2019.12.003.

[219] M. C. Hansen *et al.*, "High-Resolution Global Maps of 21st-Century Forest Cover Change," *Science,* vol. 342, no. 6160, pp. 850-853, 2013, doi: https://doi.org/10.1126/science.1244693.

[220] W. Shen, M. Li, C. Huang, T. He, X. Tao, and A. Wei, "Local land surface temperature change induced by afforestation based on satellite observations in Guangdong plantation forests in China," *Agricultural and Forest Meteorology,* vol. 276-277, p. 107641, 2019/10/15/ 2019, doi: https://doi.org/10.1016/j.agrformet.2019.107641.

[221] L. Maffei and T. Andrew, B. , "Área de acción, actividad y uso de hábitat del zorro patas negras, Cerdocyon thous, en un bosque seco," *Mastozoología Neotropical,* 2003. [Online]. Available: https://www.redalyc.org/articulo.oa?id=45710113.

[222] Chollet, François, and others, "Keras," ed, 2015. [Online]. Available: https://keras.io

[223] J. Kajornrit and K. W. Wong, "Cluster validation methods for localization of spatial rainfall data in the northeast region of Thailand," in *2013 International Conference on Machine Learning and Cybernetics*, 14-17 July 2013 2013, vol. 04, pp. 1637-1642, doi: https://doi.org/10.1109/ICMLC.2013.6890861.

[224] K. Smets, B. Verdonk, and E. M. Jordaan, "Evaluation of Performance Measures for SVR Hyperparameter Selection," in *2007 International Joint Conference on Neural Networks*, 12-17 Aug. 2007 2007, pp. 637-642, doi: https://doi.org/10.1109/IJCNN.2007.4371031.

[225] J. Kurniasih, E. Utami, and S. Raharjo, "Heuristics and Metaheuristics Approach for Query Optimization Using Genetics and Memetics Algorithm," in *2019 1st International Conference on Cybernetics and Intelligent System (ICORIS)*, 22-23

Aug. 2019 2019, vol. 1, pp. 168-172, doi: https://doi.org/10.1109/ICORIS.2019.8874909.

[226] K. Fukunaga and P. M. Narendra, "A Branch and Bound Algorithm for Computing k-Nearest Neighbors," *IEEE Transactions on Computers,* vol. C-24, no. 7, pp. 750-753, 1975, doi: https://doi.org/10.1109/T-C.1975.224297.

[227] F. Moreno-Seco, L. Micó, and J. Oncina, "Extending LAESA Fast Nearest Neighbour Algorithm to Find the k Nearest Neighbours," in *Structural, Syntactic, and Statistical Pattern Recognition*, Berlin, Heidelberg, T. Caelli, A. Amin, R. P. W. Duin, D. de Ridder, and M. Kamel, Eds., 2002 2002: Springer Berlin Heidelberg, pp. 718-724, doi: https://doi.org/10.1007/3-540-70659-3_75.

[228] B. Seongjoon and S. Koeng-Mo, "Fast k-nearest-neighbour search algorithm for nonparametric classification," *Electronics Letters,* vol. 36, no. 21, pp. 1821-1822, 2000, doi: https://doi.org/10.1049/el:20001249.

[229] D. Huichuan and L. Xiyu, "Lower C limits in support vector machines with radial basis function kernels," in *2012 International Symposium on Information Technologies in Medicine and Education*, 3-5 Aug. 2012 2012, vol. 2, pp. 768-771, doi: https://doi.org/10.1109/ITiME.2012.6291416.

[230] Z. Ye and H. Li, "Based on Radial Basis Kernel function of Support Vector Machines for speaker recognition," in *2012 5th International Congress on Image and Signal Processing*, 16-18 Oct. 2012 2012, pp. 1584-1587, doi: https://doi.org/10.1109/CISP.2012.6469807.

[231] T. Wang, X. Ye, L. Wang, and H. Li, "Grid Search Optimized SVM Method for Dish-like Underwater Robot Attitude Prediction," in *2012 Fifth International Joint Conference on Computational Sciences and Optimization*, 23-26 June 2012 2012, pp. 839-843, doi: https://doi.org/10.1109/CSO.2012.189.

[232] H. He *et al.*, *Imbalanced Learning*. 2013. doi: https://doi.org/10.1002/9781118646106.

[233] L. Y. Pratt, "Discriminability-Based Transfer between Neural Networks," presented at the Advances in Neural Information Processing Systems 5, [NIPS

Conference], 1992. [Online]. Available: http://papers.neurips.cc/paper/641-discriminability-based-transfer-between-neural-networks.pdf.

[234] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big Data,* vol. 3, no. 1, p. 9, 2016/05/28 2016, doi: https://doi.org/10.1186/s40537-016-0043-6.

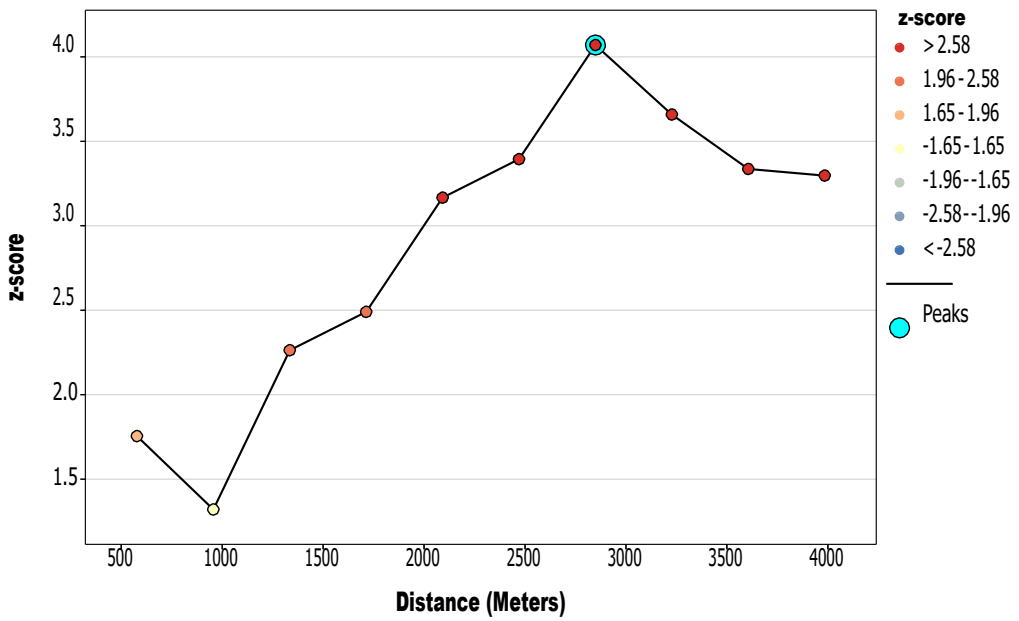
7. Anexos

7.1. Anexo 1 – Resultados Autocorrelación espacial

A continuación, se muestran los resultados obtenidos a partir del uso de la herramienta "Incremental Spatial Autocorrelation" con los reportes de atropellamiento de fauna silvestre presentes en cada tramo Estos reportes de atropellamiento fueron agrupados en segmentos, cada segmento posee una variable de conteo de puntos agrupados llamada ICOUNT la cual fue la variable de evaluación para este análisis.

7.1.1. Tramo 1

Figura 21. Autocorrelación espacial para el Tramo 1. Fuente: Autoría propia.



Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
579.00	0.256729	-0.023810	0.025555	1.754918	0.079273
957.49	0.130172	-0.023810	0.013597	1.320532	0.186657
1335.98	0.184243	-0.023810	0.008452	2.263103	0.023629
1714.48	0.164575	-0.023810	0.005722	2.490463	0.012758
2092.97	0.187658	-0.023810	0.004460	3.166447	0.001543
2471.46	0.176416	-0.023810	0.003480	3.394241	0.000688
2849.95	0.193879	-0.023810	0.002861	4.069813	0.000047
3228.44	0.154372	-0.023810	0.002372	3.658458	0.000254
3606.94	0.124838	-0.023810	0.001985	3.336194	0.000849
3985.43	0.112525	-0.023810	0.001710	3.297026	0.000977

First Peak (Distance; Value): 2849.95; 4.069813

Max Peak (Distance; Value): 2849.95; 4.069813

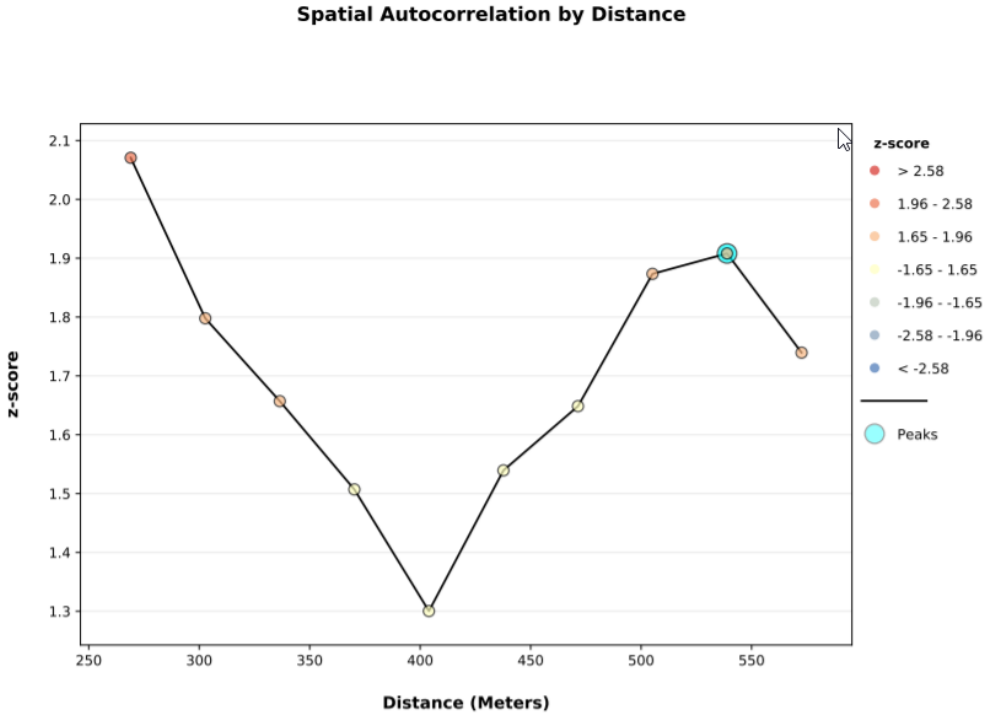
Distance measured in Meters

Incremental Autocorrelation Parameters

Parameter Name	Input Value
Input Features	Tramo1_Puntos_CollectEvents2
Input Field	ICOUNT
Number of Distance Bands	10
Beginning Distance	579.000000
Distance Increment	378.491987
Distance Method	EUCLIDEAN
Row Standardization	True
Selection Set	False

7.1.2. Tramo 2

Figura 22. Autocorrelación espacial para el Tramo 2. Fuente: Autoría propia.



Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
269.00	0.064382	-0.005236	0.001130	2.070742	0.038383
302.74	0.050720	-0.005236	0.000969	1.797817	0.072206
336.48	0.043367	-0.005236	0.000860	1.656978	0.097524
370.21	0.036478	-0.005236	0.000766	1.506972	0.131818
403.95	0.029007	-0.005236	0.000694	1.300291	0.193501
437.69	0.033232	-0.005236	0.000624	1.539373	0.123713
471.43	0.033717	-0.005236	0.000558	1.648503	0.099250
505.16	0.037318	-0.005236	0.000516	1.873402	0.061013
538.90	0.036641	-0.005236	0.000482	1.908075	0.056381
572.64	0.031481	-0.005236	0.000446	1.739424	0.081960

First Peak (Distance; Value): 538.90; 1.908075

Max Peak (Distance; Value): 538.90; 1.908075

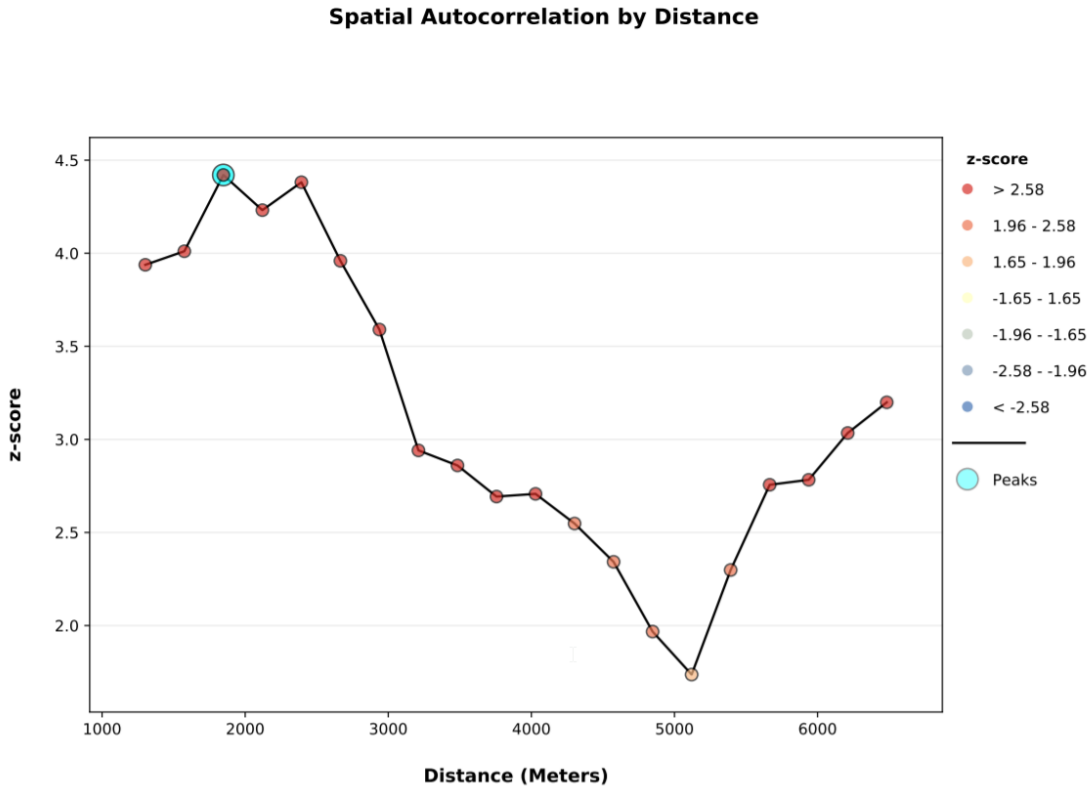
Distance measured in Meters

Incremental Autocorrelation Parameters

Parameter Name	Input Value
Input Features	Datos_Aprendizaje_Tramo_2_CopyFeatures6_CollectEvents
Input Field	ICOUNT
Number of Distance Bands	10
Beginning Distance	269.000000
Distance Increment	33.737512
Distance Method	EUCLIDEAN
Row Standardization	True
Selection Set	False

7.1.3. Tramo 3

Figura 23. Autocorrelación espacial para el Tramo 3. Fuente: Autoría propia.



Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
1302.00	0.477409	-0.037037	0.017067	3.937823	0.000082
1574.70	0.439029	-0.037037	0.014090	4.010694	0.000061
1847.40	0.424357	-0.037037	0.010895	4.420418	0.000010
2120.11	0.362729	-0.037037	0.008925	4.231519	0.000023
2392.81	0.329670	-0.037037	0.007005	4.381406	0.000012
2665.51	0.272254	-0.037037	0.006103	3.959230	0.000075
2938.21	0.225237	-0.037037	0.005338	3.589841	0.000331
3210.92	0.163975	-0.037037	0.004672	2.940845	0.003273
3483.62	0.151421	-0.037037	0.004342	2.859962	0.004237
3756.32	0.131261	-0.037037	0.003906	2.692883	0.007084
4029.02	0.123995	-0.037037	0.003538	2.707358	0.006782
4301.73	0.104980	-0.037037	0.003107	2.548015	0.010834
4574.43	0.088347	-0.037037	0.002866	2.342235	0.019169
4847.13	0.063273	-0.037037	0.002599	1.967530	0.049122
5119.83	0.045434	-0.037037	0.002255	1.736711	0.082438
5392.54	0.066613	-0.037037	0.002034	2.298406	0.021539
5665.24	0.079459	-0.037037	0.001787	2.756169	0.005848
5937.94	0.076931	-0.037037	0.001677	2.782920	0.005387

Global Moran's I Summary by Distance (Cont.)

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
6210.64	0.076132	-0.037037	0.001391	3.034318	0.002411
6483.35	0.069539	-0.037037	0.001110	3.199157	0.001378

First Peak (Distance; Value): 1847.40; 4.420418

Max Peak (Distance; Value): 1847.40; 4.420418

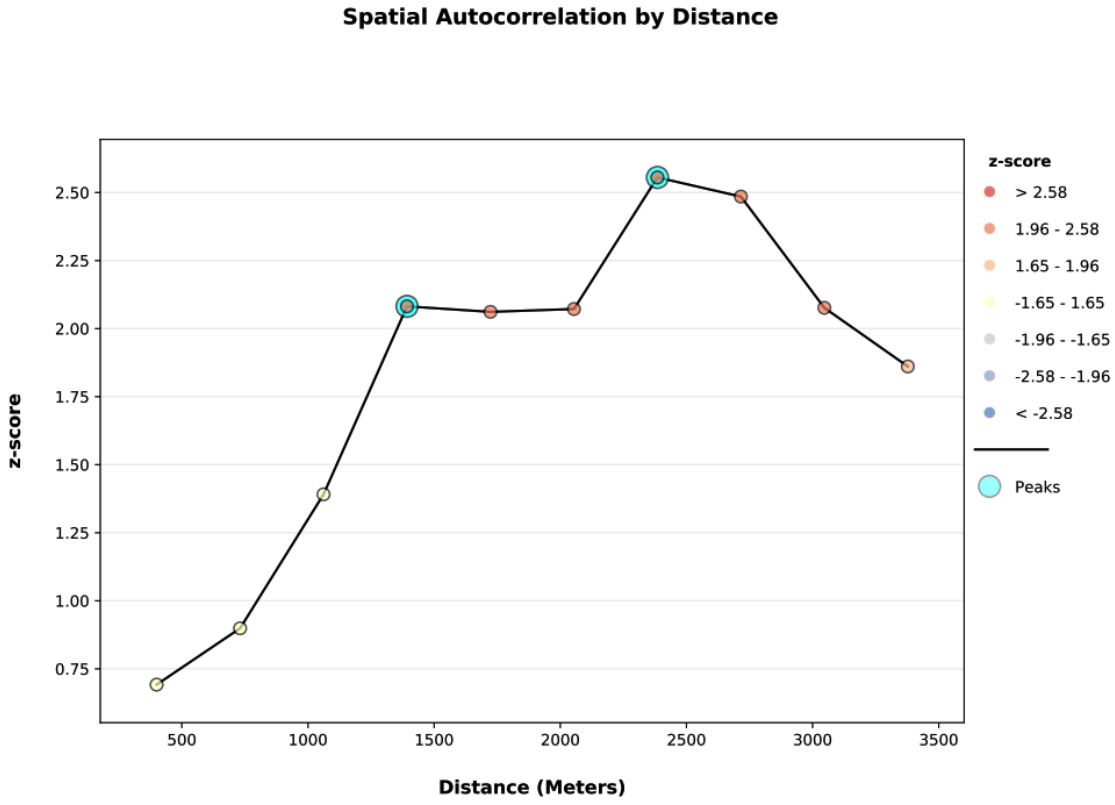
Distance measured in Meters

Incremental Autocorrelation Parameters

Parameter Name	Input Value
Input Features	Datos_Aprendizaje_Tramo_3_CopyFeatures_CollectEvents
Input Field	ICOUNT
Number of Distance Bands	20
Beginning Distance	1302.000000
Distance Increment	272.702379
Distance Method	EUCLIDEAN
Row Standardization	True
Selection Set	False

7.1.4. Tramo 4

Figura 24. Autocorrelación espacial para el Tramo 4. Fuente: Autoría propia.



Global Moran's I Summary by Distance

Distance	Moran's Index	Expected Index	Variance	z-score	p-value
400.00*	0.183333	-0.166667	0.255907	0.691874	0.489016
730.82	0.137972	-0.058824	0.047931	0.898886	0.368714
1061.64	0.179223	-0.058824	0.029284	1.391053	0.164209
1392.46	0.216448	-0.058824	0.017485	2.081776	0.037363
1723.28	0.163144	-0.058824	0.011593	2.061559	0.039250
2054.10	0.142317	-0.058824	0.009424	2.072006	0.038265
2384.93	0.148907	-0.058824	0.006608	2.555440	0.010605
2715.75	0.117597	-0.058824	0.005040	2.485152	0.012950
3046.57	0.066003	-0.058824	0.003614	2.076332	0.037863
3377.39	0.038420	-0.058824	0.002731	1.860958	0.062750

First Peak (Distance; Value): 1392.46; 2.081776

Max Peak (Distance; Value): 2384.93; 2.555440

Distance measured in Meters

* At least one distance increment resulted in features with no neighbors which may invalidate the significance of the corresponding results.

Incremental Autocorrelation Parameters

Parameter Name	Input Value
Input Features	Datos_Validacion_Tramo_4_CopyFeatures_CollectEvents
Input Field	ICOUNT
Number of Distance Bands	10
Beginning Distance	400.000000
Distance Increment	330.820917
Distance Method	EUCLIDEAN
Row Standardization	True
Selection Set	False

7.2. Anexo 2 – Reclasificación de mapas, modelo de conectividad ecológica

Descripción	Valor de Resistencia
Mapa de vocación de suelo	
Agrosilvopastoril con cultivos permanentes	40
Cuerpo de agua	50
Cultivos permanentes intensivos de clima frío	60
Cultivos permanentes intensivos de clima medio	60
Cultivos permanentes semi intensivos de clima frío	60
Cultivos permanentes semi intensivos de clima medio	60
Conservación y Recuperación Erosión	20
Cultivos transitorios intensivos de clima frío	50
Cultivos transitorios intensivos de clima medio	50
Cultivos transitorios semi intensivos de clima frío	50
Cultivos transitorios semi intensivos de clima medio	50

Forestal de producción de clima frío	30
Protección – producción	15
Forestal de protección	10
Zonas urbanas	90
Modelo Hidrográfico de la zona de estudio	
Cuerpos de agua - Orden 1	0
Cuerpos de agua - Orden 2	0
Cuerpos de agua - Orden 3	0
Quebradas - Orden 4	10
Quebradas - Orden 5	10
Ríos Principales - Orden 6	20
Capa de vías en la zona de estudio	
secondary	50
tertiary	30
residential	30
trunk	80
secondary link	50
primary	80
trunk link	70
tertiary link	30
living street	20
primary link	80
service	20
footway	20
steps	20
unclassified	50
track	20
path	30
road	30
pedestrian	20
cycleway	30
proposed	30
Capa de distancia de a vías	
Distancia a vías 0 -100 m	90
Distancia a vías 101 -300 m	80
Distancia a vías 301 -500 m	65
Distancia a vías 501 -1000 m	40
Distancia a vías > 1001 m	20
Capa de distancia a ríos	
Distancia a ríos 0 -100 m	20
Distancia a ríos 101 -300 m	40
Distancia a ríos 301 -500 m	65
Distancia a ríos 501 -1000 m	80
Distancia a ríos > 1001 m	90
Mapa de coberturas de suelo en el área de estudio	
Rio	60
Mosaico de cultivos y pastos	80

Mosaico de pastos con espacios naturales	70
Plantación forestal	40
Territorio artificializado	90
Mosaico de cultivos, pastos y espacios naturales	80
Cuerpo de agua artificial	50
Arbustal abierto	60
Arbustal denso	60
Bosque de galería y ripario	20
Bosque denso alto	20
Bosque denso bajo	20
Herbazal denso	45
Bosque abierto bajo	30
Café	70
Cultivos permanentes	80
Mosaico de cultivos y espacios naturales	50
Mosaico de pastos y espacios naturales	60
Pastos	80
Bosque fragmentado con pastos y cultivos	40
Bosque fragmentado con vegetación secundaria	40
Vegetación secundaria	30

7.3. Anexo 3 – Resultado de la prueba de comparaciones múltiples

Tabla 9. Tabla resultante de la prueba de comparaciones múltiples. Fuente: Realización propia

Algoritmo Seleccionado	Algoritmo comparado	Media	p valor
SVM_SVMSMOTE	RNA_SVMSMOTE	7.067951	0.787831
SVM_SVMSMOTE	RF_SVMSMOTE	-3.55705	0.00494
SVM_SVMSMOTE	KNN_SVMSMOTE	4.442951	0.369668
SVM_SVMSMOTE	SVM_SMOTE	6.567951	0.697467
SVM_SVMSMOTE	RNA_SMOTE	6.192951	0.632329
SVM_SVMSMOTE	RF_SMOTE	-3.30705	0.00594
SVM_SVMSMOTE	KNN_SMOTE	2.567951	0.178417
SVM_SVMSMOTE	SVM_KMEANS	6.067951	0.611205
SVM_SVMSMOTE	RNA_KMEANS	3.192951	0.231646
SVM_SVMSMOTE	RF_KMEANS	-0.55705	0.036329
SVM_SVMSMOTE	KNN_KMEANS	2.442951	0.168962

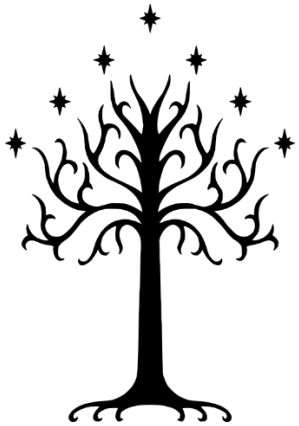
SVM_SVMSMOTE	SVM_BORDERLINE	5.942951	0.590399
SVM_SVMSMOTE	RNA_BORDERLINE	1.192951	0.094017
SVM_SVMSMOTE	RF_BORDERLINE	-3.05705	0.007118
SVM_SVMSMOTE	KNN_BORDERLINE	3.567951	0.268545
SVM_SVMSMOTE	SVM ADASYN	4.692951	0.40243
SVM_SVMSMOTE	RNA ADASYN	3.567951	0.268545
SVM_SVMSMOTE	RF ADASYN	-4.30705	0.002787
SVM_SVMSMOTE	KNN ADASYN	2.942951	0.20914
RNA_SVMSMOTE	RF_SVMSMOTE	-2.43205	0.011029
RNA_SVMSMOTE	KNN_SVMSMOTE	5.567951	0.530025
RNA_SVMSMOTE	SVM_SMOTE	7.692951	0.90479
RNA_SVMSMOTE	RNA_SMOTE	7.317951	0.834197
RNA_SVMSMOTE	RF_SMOTE	-2.18205	0.013066
RNA_SVMSMOTE	KNN_SMOTE	3.692951	0.281697
RNA_SVMSMOTE	SVM_KMEANS	7.192951	0.810931
RNA_SVMSMOTE	RNA_KMEANS	4.317951	0.353927
RNA_SVMSMOTE	RF_KMEANS	0.567951	0.068138
RNA_SVMSMOTE	KNN_KMEANS	3.567951	0.268545
RNA_SVMSMOTE	SVM_BORDERLINE	7.067951	0.787831
RNA_SVMSMOTE	RNA_BORDERLINE	2.317951	0.159887
RNA_SVMSMOTE	RF_BORDERLINE	-1.93205	0.015429
RNA_SVMSMOTE	KNN_BORDERLINE	4.692951	0.40243
RNA_SVMSMOTE	SVM ADASYN	5.817951	0.569926
RNA_SVMSMOTE	RNA ADASYN	4.692951	0.40243
RNA_SVMSMOTE	RF ADASYN	-3.18205	0.006505
RNA_SVMSMOTE	KNN ADASYN	4.067951	0.323738
RF_SVMSMOTE	KNN_SVMSMOTE	16.19295	0.055645
RF_SVMSMOTE	SVM_SMOTE	18.31795	0.015429
RF_SVMSMOTE	RNA_SMOTE	17.94295	0.019677
RF_SVMSMOTE	RF_SMOTE	8.442951	0.95231
RF_SVMSMOTE	KNN_SMOTE	14.31795	0.142851
RF_SVMSMOTE	SVM_KMEANS	17.81795	0.021304
RF_SVMSMOTE	RNA_KMEANS	14.94295	0.10636
RF_SVMSMOTE	RF_KMEANS	11.19295	0.472956
RF_SVMSMOTE	KNN_KMEANS	14.19295	0.151186
RF_SVMSMOTE	SVM_BORDERLINE	17.69295	0.023048
RF_SVMSMOTE	RNA_BORDERLINE	12.94295	0.255822
RF_SVMSMOTE	RF_BORDERLINE	8.692951	0.90479
RF_SVMSMOTE	KNN_BORDERLINE	15.31795	0.088291
RF_SVMSMOTE	SVM ADASYN	16.44295	0.048426
RF_SVMSMOTE	RNA ADASYN	15.31795	0.088291
RF_SVMSMOTE	RF ADASYN	7.442951	0.857609

RF_SVMSMOTE	KNN ADASYN	14.69295	0.119954
KNN_SVMSMOTE	SVM_SMOTE	10.31795	0.611205
KNN_SVMSMOTE	RNA_SMOTE	9.942951	0.675476
KNN_SVMSMOTE	RF_SMOTE	0.442951	0.063739
KNN_SVMSMOTE	KNN_SMOTE	6.317951	0.653758
KNN_SVMSMOTE	SVM_KMEANS	9.817951	0.697467
KNN_SVMSMOTE	RNA_KMEANS	6.942951	0.764916
KNN_SVMSMOTE	RF_KMEANS	3.192951	0.231646
KNN_SVMSMOTE	KNN_KMEANS	6.192951	0.632329
KNN_SVMSMOTE	SVM_BORDERLINE	9.692951	0.719716
KNN_SVMSMOTE	RNA_BORDERLINE	4.942951	0.436874
KNN_SVMSMOTE	RF_BORDERLINE	0.692951	0.072782
KNN_SVMSMOTE	KNN_BORDERLINE	7.317951	0.834197
KNN_SVMSMOTE	SVM ADASYN	8.442951	0.95231
KNN_SVMSMOTE	RNA ADASYN	7.317951	0.834197
KNN_SVMSMOTE	RF ADASYN	-0.55705	0.036329
KNN_SVMSMOTE	KNN ADASYN	6.692951	0.719716
SVM_SMOTE	RNA_SMOTE	7.817951	0.928518
SVM_SMOTE	RF_SMOTE	-1.68205	0.018159
SVM_SMOTE	KNN_SMOTE	4.192951	0.338616
SVM_SMOTE	SVM_KMEANS	7.692951	0.90479
SVM_SMOTE	RNA_KMEANS	4.817951	0.419444
SVM_SMOTE	RF_KMEANS	1.067951	0.088291
SVM_SMOTE	KNN_KMEANS	4.067951	0.323738
SVM_SMOTE	SVM_BORDERLINE	7.567951	0.881146
SVM_SMOTE	RNA_BORDERLINE	2.817951	0.1985
SVM_SMOTE	RF_BORDERLINE	-1.43205	0.021304
SVM_SMOTE	KNN_BORDERLINE	5.192951	0.472956
SVM_SMOTE	SVM ADASYN	6.317951	0.653758
SVM_SMOTE	RNA ADASYN	5.192951	0.472956
SVM_SMOTE	RF ADASYN	-2.68205	0.00928
SVM_SMOTE	KNN ADASYN	4.567951	0.385836
RNA_SMOTE	RF_SMOTE	-1.30705	0.023048
RNA_SMOTE	KNN_SMOTE	4.567951	0.385836
RNA_SMOTE	SVM_KMEANS	8.067951	0.976144
RNA_SMOTE	RNA_KMEANS	5.192951	0.472956
RNA_SMOTE	RF_KMEANS	1.442951	0.10636
RNA_SMOTE	KNN_KMEANS	4.442951	0.369668
RNA_SMOTE	SVM_BORDERLINE	7.942951	0.95231
RNA_SMOTE	RNA_BORDERLINE	3.192951	0.231646
RNA_SMOTE	RF_BORDERLINE	-1.05705	0.026909
RNA_SMOTE	KNN_BORDERLINE	5.567951	0.530025

RNA_SMOTE	SVM ADASYN	6.692951	0.719716
RNA_SMOTE	RNA ADASYN	5.567951	0.530025
RNA_SMOTE	RF ADASYN	-2.30705	0.012009
RNA_SMOTE	KNN ADASYN	4.942951	0.436874
RF_SMOTE	KNN_SMOTE	14.06795	0.159887
RF_SMOTE	SVM_KMEANS	17.56795	0.024914
RF_SMOTE	RNA_KMEANS	14.69295	0.119954
RF_SMOTE	RF_KMEANS	10.94295	0.510621
RF_SMOTE	KNN_KMEANS	13.94295	0.168962
RF_SMOTE	SVM_BORDERLINE	17.44295	0.026909
RF_SMOTE	RNA_BORDERLINE	12.69295	0.281697
RF_SMOTE	RF_BORDERLINE	8.442951	0.95231
RF_SMOTE	KNN_BORDERLINE	15.06795	0.100037
RF_SMOTE	SVM ADASYN	16.19295	0.055645
RF_SMOTE	RNA ADASYN	15.06795	0.100037
RF_SMOTE	RF ADASYN	7.192951	0.810931
RF_SMOTE	KNN ADASYN	14.44295	0.134873
KNN_SMOTE	SVM_KMEANS	11.69295	0.40243
KNN_SMOTE	RNA_KMEANS	8.817951	0.881146
KNN_SMOTE	RF_KMEANS	5.067951	0.454713
KNN_SMOTE	KNN_KMEANS	8.067951	0.976144
KNN_SMOTE	SVM_BORDERLINE	11.56795	0.419444
KNN_SMOTE	RNA_BORDERLINE	6.817951	0.742205
KNN_SMOTE	RF_BORDERLINE	2.567951	0.178417
KNN_SMOTE	KNN_BORDERLINE	9.192951	0.810931
KNN_SMOTE	SVM ADASYN	10.31795	0.611205
KNN_SMOTE	RNA ADASYN	9.192951	0.810931
KNN_SMOTE	RF ADASYN	1.317951	0.100037
KNN_SMOTE	KNN ADASYN	8.567951	0.928518
SVM_KMEANS	RNA_KMEANS	5.317951	0.491595
SVM_KMEANS	RF_KMEANS	1.567951	0.112996
SVM_KMEANS	KNN_KMEANS	4.567951	0.385836
SVM_KMEANS	SVM_BORDERLINE	8.067951	0.976144
SVM_KMEANS	RNA_BORDERLINE	3.317951	0.243523
SVM_KMEANS	RF_BORDERLINE	-0.93205	0.029041
SVM_KMEANS	KNN_BORDERLINE	5.692951	0.549797
SVM_KMEANS	SVM ADASYN	6.817951	0.742205
SVM_KMEANS	RNA ADASYN	5.692951	0.549797
SVM_KMEANS	RF ADASYN	-2.18205	0.013066
SVM_KMEANS	KNN ADASYN	5.067951	0.454713
RNA_KMEANS	RF_KMEANS	4.442951	0.369668
RNA_KMEANS	KNN_KMEANS	7.442951	0.857609

RNA_KMEANS	SVM_BORDERLINE	10.94295	0.510621
RNA_KMEANS	RNA_BORDERLINE	6.192951	0.632329
RNA_KMEANS	RF_BORDERLINE	1.942951	0.134873
RNA_KMEANS	KNN_BORDERLINE	8.567951	0.928518
RNA_KMEANS	SVM ADASYN	9.692951	0.719716
RNA_KMEANS	RNA ADASYN	8.567951	0.928518
RNA_KMEANS	RF ADASYN	0.692951	0.072782
RNA_KMEANS	KNN ADASYN	7.942951	0.95231
RF_KMEANS	KNN_KMEANS	11.19295	0.472956
RF_KMEANS	SVM_BORDERLINE	14.69295	0.119954
RF_KMEANS	RNA_BORDERLINE	9.942951	0.675476
RF_KMEANS	RF_BORDERLINE	5.692951	0.549797
RF_KMEANS	KNN_BORDERLINE	12.31795	0.323738
RF_KMEANS	SVM ADASYN	13.44295	0.20914
RF_KMEANS	RNA ADASYN	12.31795	0.323738
RF_KMEANS	RF ADASYN	4.442951	0.369668
RF_KMEANS	KNN ADASYN	11.69295	0.40243
KNN_KMEANS	SVM_BORDERLINE	11.69295	0.40243
KNN_KMEANS	RNA_BORDERLINE	6.942951	0.764916
KNN_KMEANS	RF_BORDERLINE	2.692951	0.188261
KNN_KMEANS	KNN_BORDERLINE	9.317951	0.787831
KNN_KMEANS	SVM ADASYN	10.44295	0.590399
KNN_KMEANS	RNA ADASYN	9.317951	0.787831
KNN_KMEANS	RF ADASYN	1.442951	0.10636
KNN_KMEANS	KNN ADASYN	8.692951	0.90479
SVM_BORDERLINE	RNA_BORDERLINE	3.442951	0.255822
SVM_BORDERLINE	RF_BORDERLINE	-0.80705	0.031316
SVM_BORDERLINE	KNN_BORDERLINE	5.817951	0.569926
SVM_BORDERLINE	SVM ADASYN	6.942951	0.764916
SVM_BORDERLINE	RNA ADASYN	5.817951	0.569926
SVM_BORDERLINE	RF ADASYN	-2.05705	0.014204
SVM_BORDERLINE	KNN ADASYN	5.192951	0.472956
RNA_BORDERLINE	RF_BORDERLINE	3.942951	0.309292
RNA_BORDERLINE	KNN_BORDERLINE	10.56795	0.569926
RNA_BORDERLINE	SVM ADASYN	11.69295	0.40243
RNA_BORDERLINE	RNA ADASYN	10.56795	0.569926
RNA_BORDERLINE	RF ADASYN	2.692951	0.188261
RNA_BORDERLINE	KNN ADASYN	9.942951	0.675476
RF_BORDERLINE	KNN_BORDERLINE	14.81795	0.112996
RF_BORDERLINE	SVM ADASYN	15.94295	0.063739
RF_BORDERLINE	RNA ADASYN	14.81795	0.112996
RF_BORDERLINE	RF ADASYN	6.942951	0.764916

RF_BORDERLINE	KNN ADASYN	14.19295	0.151186
KNN_BORDERLINE	SVM ADASYN	9.317951	0.787831
KNN_BORDERLINE	RNA ADASYN	8.192951	1
KNN_BORDERLINE	RF ADASYN	0.317951	0.059578
KNN_BORDERLINE	KNN ADASYN	7.567951	0.881146
SVM ADASYN	RNA ADASYN	7.067951	0.787831
SVM ADASYN	RF ADASYN	-0.80705	0.031316
SVM ADASYN	KNN ADASYN	6.442951	0.675476
RNA ADASYN	RF ADASYN	0.317951	0.059578
RNA ADASYN	KNN ADASYN	7.567951	0.881146
RF ADASYN	KNN ADASYN	15.44295	0.08285



What we do in life, echoes in eternity