PREDICCIÓN DE AVISTAMIENTOS DE AVES PARA LA CONSERVACIÓN DE ESPECIES ENDÉMICAS UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Paula Andrea López Arango
Código 8.985.606
María Victoria Escobar Martínez
Código 8.986.266

Proyecto Aplicado para optar al título de
Magíster en Ciencia de Datos

Director(a)
M.Sc. Juan Sebastián Blandón

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, FEBRERO 2 DE 2025

FICHA RESUMEN

POSIBLE TÍTULO: PREDICCIÓN DE AVISTAMIENTOS DE AVES PARA LA CONSERVACIÓN DE ESPECIES ENDÉMICAS UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

1. ÁREA DE TRABAJO: Ciencia de datos aplicada a la biodiversidad.
2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Investigación.
3. ESTUDIANTE(S):

   Paula Andrea López Arango

   María Victoria Escobar Martínez
4. CORREO ELECTRÓNICO:

   pplopez2501@javerianacali.edu.co

   mvescobarm@javerianacali.edu.co
5. DIRECCIÓN Y TELÉFONO:

   +57 310 3582299

   +57 323 3068549
6. DIRECTOR: M.Sc. Juan Sebastián Blandón
7. VINCULACIÓN DEL DIRECTOR: Profesional externo
8. CORREO ELECTRÓNICO DEL DIRECTOR: juan.blandon@utec.edu.uy
9. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Grupo de Investigación en Aplicaciones de Inteligencia Artificial - ARIA, Ingeniería Agroambiental, Departamento de Sostenibilidad Ambiental, Universidad Tecnológica de Uruguay.
10. PALABRAS CLAVE (al menos 5): Ciencia de datos, predicciones espaciotemporales, aves, ecoturismo, Colombia, especies endémicas, biodiversidad.
11. FECHA DE INICIO: 21 de octubre de 2023
12. DURACIÓN ESTIMADA (En meses): 12 meses
13. RESUMEN:

El presente proyecto muestra los resultados sobre predicción de avistamientos de aves para la conservación de especie endémicas mediante la aplicación de algoritmos de Aprendizaje de Automático. La región de América Latina y el Caribe tiene dos características que hacen que el estudio de los efectos del cambio climático sobre la biodiversidad resulte particularmente relevante: i) es una de las regiones más vulnerables frente al cambio climático y ii) es una de las regiones con mayor concentración de biodiversidad del planeta. En Colombia hay

aproximadamente el 20 % de las especies de aves del planeta, convirtiéndose en el país con la mayor diversidad en este ámbito, con un número de especies registradas para el 2020 de 1954, y de las cuales 82 eran endémicas. De esta forma, se desarrolló una metodología de predicción de avistamientos de aves con el fin de aportar insumos para la conservación de especies endémicas a partir de algoritmos de ML. Los resultados de la investigación consistieron en implementar algoritmos en Python/R que aporten a la gestión de datos de avistamientos de aves, además permitiendo tratar datos georreferenciados de variables exógenas, para establecer correlaciones entre estas y datos de avistamientos de aves. El módulo de algoritmos de Modelos de Distribución de Especies permitió la identificación de áreas críticas para la conservación y el desarrollo y/o fortalecimiento del aviturismo para ciertos niveles de amenaza y departamentos específicos. Además, estos resultados llevaron a la generación de conocimiento que sirve de insumo para el desarrollo de planes de conservación y/o planificación del aviturismo en las regiones identificadas

TABLA DE CONTENIDO

ABREVIATURAS 10

INTRODUCCIÓN 12

1. DEFINICIÓN DEL PROBLEMA 13

1.1. PLANTEAMIENTO DEL PROBLEMA 13

1.2. FORMULACIÓN DEL PROBLEMA 14

2. OBJETIVOS DEL PROYECTO 15

2.1. OBJETIVO GENERAL 15

2.2. OBJETIVOS ESPECÍFICOS 15

3. MARCO TEÓRICO Y ANTECEDENTES 16

3.1. MARCO TEÓRICO 16

3.2. ANTECEDENTES 24

4. METODOLOGÍA 28

4.1. DESARROLLAR UNA METODOLOGÍA DE PREPROCESAMIENTO DE DATOS DE AVISTAMIENTOS DE AVES, CON EL FIN DE IDENTIFICAR DEPARTAMENTOS DE ALTA CONCENTRACIÓN DE ESPECIES ENDÉMICAS 28

4.2. DESARROLLAR UNA METODOLOGÍA DE PREPROCESAMIENTO DE DATOS DE VARIABLES EXÓGENAS (PRECIPITACIONES, TEMPERATURAS, ANTROPOGÉNICAS, ENTRE OTRAS) PARA ESTABLECER CORRELACIONES SOBRE LOS AVISTAMIENTOS DE ESPECIES ENDÉMICAS. 39

4.3. IMPLEMENTACIÓN Y VALIDACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE AVISTAMIENTOS DE ESPECIES ENDÉMICAS 51

5. CONCLUSIONES 66

6. REFERENCIAS BIBLIOGRAFICAS 68

7. ANEXO 73

7.1. ANEXOS OBJETIVO 1 73

7.2. ANEXOS OBJETIVO 2 78

7.3. ANEXOS OBJETIVO 3 84

LISTA DE TABLAS

Tabla 1. Categorías de aves según nivel de amenaza 30
Tabla 2. Listado de especies endémicas en Colombia de acuerdo con nivel de amenaza 31
Tabla 3. Listado de variables bioclimáticas de la colección de WorldClim 42
Tabla 4. Variables seleccionadas para matriz de correlación de variables exógenas posterior a diputación VIF.....43
Tabla 5. Resumen del conjunto de datos por categoría. 52
Tabla 6. Resultados de evaluación de los modelos Regresión Logística y Random Forest para la predicción de avistamiento de aves para categoría de peligro VU 56
Tabla 7. Resultado de la importancia de las características del modelo Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas para categoría Vulnerable (VU). 65
Tabla 8. Variables seleccionadas para matriz de correlación de variables exógenas posterior a diputación VIF.....78
Tabla 9. Resultados de evaluación de los modelos Regresión Logística y Random Forest para la predicción de avistamiento de aves para categoría de peligro CR. 84
Tabla 10. Resultados de evaluación de los modelos Regresión Logística y Random Forest para la predicción de avistamiento de aves para categoría de peligro EN. 87
Tabla 10. Resultado de la importancia de las características del modelo Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas para categoría Peligro Critico (CR). 96
Tabla 12. Resultado de la importancia de las características del modelo Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas para categoría En Peligro (EN). 97

LISTA DE FIGURAS

Figura 1. Flujo de trabajo para la realización de modelos de distribución de especies [21]. 19
Figura 2. Distribución de oso perezoso en Suramérica [Elaboración propia].....20
Figura 3. Matriz de confusión, métrica de evaluación de desempeño de modelos de predicción. 23
Figura 4. Diagrama de flujo resumen de metodología. 28
Figura 5. Abundancia de aves endémicas en Colombia 2003 – 2023. SPXX: Código del nombre científico de la especie. Las especies seleccionadas corresponden a: SP08: Henicorhina negreti. SP22: Penelope perspicax. SP41: Hypopyrrhus pyrohypogaster. 33
Figura 6. Abundancia de aves endémicas por departamento (2003-2023) en nivel Vulnerable (VU).....34
Figura 7. Abundancia de aves endémicas para la *Hypopyrrhus pyrohypogaster.* La representación confirma que Risaralda es el departamento que durante cinco años seguidos

presentó la mayor concentración de la especie. 35
Figura 8. Abundancia vs variables de esfuerzo. Aves endémicas VU en Colombia. 2003 – 2023. Las variables de esfuerzo corresponden a: Duración de la observación (Duration of Obs. (h)), Distancia de esfuerzo (Effort Distance), Horas de esfuerzo (Effort Hours) y Velocidad de Esfuerzo (Effort Speed). Los valores para cada variable corresponden a las sumas totales por departamento. 36
Figura 9. Matriz de correlación entre las variables de esfuerzo y aves endémicas categoría CR, EN, VU de Colombia. OC: Observation Count. DO: Duration Observation [hr]. ED: Effort Distance [km]. EH: Total Effort Hours [hr]. ES: Total Effort Speed [km/h]. 37
Figura 10. Aves con mayor abundancia en Vulnerable – VU (Hypopyrrhus pyrohypogaster). Imagen obtenida de Birds of the World. Cornell Lab of Ornithology, Ithaca, NY, EE.UU. 38
Figura 11. Variación de uso del suelo 2014 - 2020 en el departamento de Antioquia, procesamiento de base de datos de MapBiomas de la plataforma Google Earth Engine. 45
Figura 12. Representación espacial de la variable Bioclimática Bio1: Temperatura media anual para el departamento de Antioquia. 47
Figura 13. Representación espacial de la elevación para ambos departamentos. La resolución de esta variable condiciona el detalle con el que se pueda apreciar la tendencia de la variable en el departamento de Antioquia. 48
Figura 14. Distribución espacial y representación hexagonal de registros de *Hypopyrrhus Pyrohypogaster* en Antioquia. Este ejemplar presenta mayor cantidad de avistamientos. Sin embargo, esto también se puede deber a la extensión de territorio que comprende el departamento. 50
Figura 15. Matrices de correlación para el conjunto de características exógenas completas (Full Variables - Izquierda) y el conjunto de datos una vez se aplicó el VIF (Derecha) para el Departamento de Antioquia. 50
Figura 16. Matriz de confusión evaluación inicial de modelos regresión Logistica y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría VU. 56
Figura 17. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica Random over sampling con todas las variables disponibles. Categoría VU. 56
Figura 18. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría VU. 57
Figura 19. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría VU. 57
Figura 20. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica Random over sampling con selección de variables representativas. Categoría VU. 57
Figura 21. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría VU 58
Figura 22. Mapa de presencias registradas en la evaluación del modelo. Categoría VU. 59

Figura 23. Mapa de probabilidad evaluación inicial de modelos Regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría VU. 60
Figura 24. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con todas las variables disponibles. Categoría VU. 61
Figura 25. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría VU. 61
Figura 26. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría VU. 62
Figura 27. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas. Categoría VU. 63
Figura 28. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría VU.....63
Figura 29. Abundancia de aves endémicas por departamento (2003-2023) en nivel de Peligro Crítico (CR).....73
Figura 30. Abundancia de aves endémicas por departamento (2003-2023) en nivel En Peligro (EN).....74
Figura 31. Abundancia de aves endémicas para la *Hernicorhina negreti*. La representación confirma que Risaralda es el departamento que durante cinco años seguidos presentó la mayor concentración de la especie.....75
Figura 32. Abundancia de aves endémicas para la *Penelope perspicax*. La representación confirma que Risaralda es el departamento que durante cinco años seguidos presentó la mayor concentración de la especie.....75
Figura 33. Abundancia vs variables de esfuerzo. Aves endémicas en CR en Colombia. 2003 – 2023. Las variables de esfuerzo corresponden a: Duración de la observación (Duration of Obs. (h)), Distancia de esfuerzo (Effort Distance), Horas de esfuerzo (Effort Hours) y Velocidad de Esfuerzo (Effort Speed). Los valores para cada variable corresponden a las sumas totales por departamento. 76
Figura 34. Abundancia vs variables de esfuerzo. Aves endémicas EN, en Colombia. 2003 – 2023. Las variables de esfuerzo corresponden a: Duración de la observación (Duration of Obs. (h)), Distancia de esfuerzo (Effort Distance), Horas de esfuerzo (Effort Hours) y Velocidad de Esfuerzo (Effort Speed). Los valores para cada variable corresponden a las sumas totales por departamento. 76
Figura 35. Matriz de correlación entre las variables de esfuerzo y aves endémicas categoría CR y EN de Colombia. OC: Observation Count. DO: Duration Observation [hr]. ED: Effort Distance [km]. EH: Total Effort Hours [hr]. ES: Total Effort Speed [km/h]. 77
Figura 36. Aves con mayor abundancia en las categorías Peligro Crítico – CR (Henicorhina negreti) y En Peligro – EN (Penelope perspicax). Imágenes obtenidas de Birds of the World. Cornell Lab of Ornithology, Ithaca, NY, EE.UU. 77

Figura 37. Variación de uso del suelo 2014 - 2020 en el departamento de Risaralda, procesamiento de base de datos de MapBiomas de la plataforma Google Earth Engine. 79
Figura 38. Representación espacial de la variable Bioclimática Bio1: Temperatura media anual para el departamento de Risaralda.....80
Figura 39. Representación espacial de la elevación para ambos departamentos. La resolución de esta variable condiciona el detalle con el que se pueda apreciar la tendencia de la variable en el departamento de Risaralda.....81
Figura 40. Distribución espacial y representación hexagonal de registros de *Henicorhina Negreti* en Risaralda. Los puntos rojos representan la presencia de la especie, mientras que los azules corresponden a las ausencias obtenidas a partir del preprocesamiento de los datos. 82
Figura 41. Distribución espacial y representación hexagonal de registros de *Penelope Perspicax* en Risaralda. Esta especie, a comparación de la del nivel CR, tiene más presencias hacia el sur del departamento, lo que tiene sentido teniendo en cuenta el tipo de nivel de amenaza en el que está categorizada. 83
Figura 42. Matrices de correlación para el conjunto de características exógenas completas (Full Variables - Izquierda) y el conjunto de datos una vez se aplicó el VIF (Derecha) para el Departamento de Risaralda. Se evidencia que, si bien en el conjunto VIF existen algunas relaciones fuertes, la umbralización implicó preservar estas variables. 84
Figura 43. Matriz de confusión evaluación inicial de modelos regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría CR. 85
Figura 44. Matriz de confusión evaluación de modelos Regresión logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con todas las variables disponibles. Categoría CR.....85
Figura 45. Matriz de confusión evaluación de modelos Regresión Logistica y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría CR.....85
Figura 46. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría CR. 86
Figura 47. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas. Categoría CR.....86
Figura 48. Matriz de confusión evaluación de modelos regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría CR.....86
Figura 49. Matriz de confusión evaluación inicial de modelos Regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría EN. 87
Figura 50. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con todas las variables disponibles. Categoría EN.....87
Figura 51. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría EN. 88

Figura 52. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest sin realizar balanceo y con variables representativas. Categoría EN. 88
Figura 53. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas. Categoría EN.....88
Figura 54. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría EN.....89
Figura 55. Mapa de presencias registradas en la evaluación del modelo. Categoría CR. 89
Figura 56. Mapa de probabilidad evaluación inicial de modelos Regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría CR. 90
Figura 57. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con todas las variables disponibles. Categoría CR.....90
Figura 58. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría CR.....91
Figura 59. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría CR.....91
Figura 60. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con selección de variables representativas. Categoría CR.....92
Figura 61. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría CR.....92
Figura 62. Mapa de presencias registradas en la evaluación del modelo. Categoría EN. 93
Figura 63. Mapa de probabilidad evaluación inicial de modelos Regresión logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría EN. 93
Figura 64. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con todas las variables disponibles. Categoría EN.....94
Figura 65. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría EN. 94
Figura 66. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría EN.....95
Figura 67. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con selección de variables representativas. Categoría EN.....95
Figura 68. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría EN.....96

ABREVIATURAS

Abreviatura	Descripción
AI	Inteligencia Artificial
ARIA	Grupo de Investigación en Aplicaciones de Inteligencia Artificial
ASORHUILA	Asociación Ornitológica del Huila
CAPTAIN	Priorización de Áreas de Conservación mediante Inteligencia Artificial
CC	Cambio Climático
CR	Peligro Crítico
CRS	Sistema de Referencia de Coordenadas (Coordinate Reference System)
CS	Ciencia Ciudadana
CSV	Valores Separados por Coma (formato de datos delimitado por comas)
CV	Variabilidad Climática
DS	Ciencia de Datos
EBD	Conjunto de Datos Básicos de eBird
EN	En Peligro
EObs	Duración de Observación
ED	Distancia de Esfuerzo
EH	Horas de Esfuerzo
ES	Velocidad de Esfuerzo
FN	Falso Negativo *(False Negative)*
FP	Falso Positivo *(False Positive)*
GBIF	Sistema Global de Información sobre Biodiversidad
GEE	Google Earth Engine
IPCC	Panel Intergubernamental sobre el Cambio Climático
LC	Preocupación Menor
MADS	Ministerio de Ambiente y Desarrollo Sostenible
ML	Aprendizaje Automático
NT	Casi Amenazada
OC	Conteo de Observaciones
RF	Random Forest *(Bosque Aleatorio)*
RL	Regresión Logística
SED	Datos de Evento de Muestreo
SDM	Modelos de Distribución de Especies
SIG	Sistemas de Información Geográfica

SMOTE	Técnica de Sobremuestreo de Minorías Sintéticas
SVM	Máquinas de Vectores de Soporte
TN	Verdadero Negativo *(True Negative)*
TP	Verdadero Positivo *(True Positive)*
VIF	Factor de Inflación de Varianza
VU	Vulnerable
CAPTAIN	Priorización de Áreas de Conservación mediante Inteligencia Artificial
MaxEnt	Máxima Entropía
SPXX	Código de Especie
CO-ANT	Código de Antioquia
CO-RIS	Código de Risaralda

INTRODUCCIÓN

La conservación de la biodiversidad ha emergido como una prioridad global debido a los efectos adversos del Cambio Climático (*Climate Change*, CC), como lo son el aumento de la temperatura global, modificaciones en los patrones de precipitación y la intensificación de eventos climáticos extremos. Estas alteraciones ambientales representan una amenaza significativa para la supervivencia de los ecosistemas y la existencia de especies endémicas en distintas regiones del planeta [2]. En este contexto, América Latina y el Caribe, que albergan una de las mayores concentraciones de biodiversidad, se enfrentan a retos particulares. Esta región es especial por su diversidad aviar, siendo hogar de la mayor concentración de especies de aves del mundo [4]. Teniendo en cuenta lo anterior, los esfuerzos de conservación de avifauna en esta área son de gran importancia, dado que la pérdida de biodiversidad aviar en América Latina no solo afecta los ecosistemas locales, sino también el equilibrio ecológico global.

A nivel mundial, diversas investigaciones se han dedicado a abordar la protección y conservación de las especies más vulnerables. Iniciativas como eBird han desempeñado un papel fundamental al recopilar, gestionar y proporcionar acceso a datos sobre avistamientos de aves a nivel global [9]. Este esfuerzo ha facilitado la acumulación de grandes cantidades de datos, que mediante técnicas avanzadas de análisis de datos y la aplicación de algoritmos de Aprendizaje de Máquina (*Machine Learning*, ML), han permitido identificar patrones recurrentes y predecir la distribución de especies sometidas a los efectos del CC y la Variabilidad Climática (*Climate Variability*, CV).

La integración de técnicas de ML e Inteligencia Artificial (*Artificial Intelligence*, AI) en el análisis de datos de eBird ha revolucionado el campo de la conservación. Estos métodos permiten no solo predecir la presencia y abundancia de aves en diferentes regiones, sino también identificar áreas críticas para su conservación y formular estrategias basadas en datos. Ejemplos de estos enfoques incluyen la predicción de migraciones, identificación de hábitats vulnerables y el monitoreo continuo de poblaciones de aves [22].

Este trabajo tuvo como objetivo predecir avistamientos de aves endémicas en diversas regiones de Colombia utilizando herramientas de Ciencia de Datos (*Data Science*, DS), específicamente algoritmos de ML. Este enfoque permitió extraer conocimiento de grandes volúmenes de datos, generando insumos preliminares para la formulación de estrategias de conservación y promoción del aviturismo que respete la estabilidad de los ecosistemas. Así, se busca contribuir significativamente a la preservación de la biodiversidad aviar y al fortalecimiento del conocimiento sobre la dinámica de estas especies en el país.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

De acuerdo con el Panel Intergubernamental sobre el Cambio Climático (IPCC), el CC ocasionará aumentos paulatinos en la temperatura promedio de la superficie de la tierra y de los océanos, modificaciones de los patrones de precipitación, cambios de intensidad y frecuencia de los eventos climáticos extremos y un alza en el nivel medio del mar [6]. La región de América Latina y el Caribe tiene dos características que hacen que el estudio de los efectos del cambio climático sobre la biodiversidad resulte particularmente relevante: i) es una de las regiones más vulnerables frente al cambio climático [1] y ii) es una de las regiones con mayor concentración de biodiversidad del planeta [2].

La ONU ha identificado 178 regiones ecológicas en América Latina y el Caribe que albergan entre el 50% y el 80% de toda la biodiversidad del planeta, incluyendo algunos de los países más biodiversos del mundo como lo son Brasil, Colombia, México y Perú [7]. En términos de biodiversidad, las aves son los organismos más conocidos a nivel mundial cuando se trata de investigaciones sobre el clima. El enorme conjunto de datos recolectados por millones de observadores de aves alrededor del mundo ha permitido aproximarse a los efectos del CC sobre sus poblaciones en América Latina y el Caribe, donde se concentra la mayor diversidad de especies [8]. En Colombia hay aproximadamente el 20 % de las especies de aves del planeta, lo que lo hace convertirse en el país con mayor diversidad en este ámbito, con un número de especies registradas para el 2020 de 1954, y de las cuales 82 son endémicas, solo en el territorio colombiano [3].

Ante el auge de las tecnologías de la información se ha logrado compilar con mayor facilidad los registros de observación de especies de aves a lo largo del país y en diferentes temporalidades, en sistemas de bases de datos como lo es eBird [9]. En este contexto, fue de importancia comprender la dinámica poblacional de estas especies a través de un análisis preliminar de datos, que permitió identificar patrones temporales y espaciales con interés particular en las especies endémicas, ya que se han convertido en un pilar fundamental en el ecoturismo, particularmente el aviturismo. De hecho, el país cuenta con 59 áreas protegidas, 23 de ellas abiertas al ecoturismo comunitario y al menos 25 donde se promueve la observación de aves [5].

Con este propósito, se realizó el análisis exploratorio de datos a través de herramientas gráficas y estadísticas mediante Python, identificando especies con el mayor número de abundancia, los períodos de mayor avistamiento y los departamentos que concentraban el mayor número de abundancia y riqueza para distintos estados de amenaza (En Peligro Crítico, CR; En peligro, EN; Vulnerable, VU). Se estableció la base para la aplicación de algoritmos de ML que llevaron a cabo predicciones de forma que los resultados de esta investigación sirvan como apoyo para la formulación de estrategias de conservación de la biodiversidad aviar en el país, contribuyendo al fortalecimiento del aviturismo. A partir de lo anterior se abordó la problemática de gestionar la

biodiversidad en el país, que es un aspecto vital para garantizar la preservación y conservación de esta.

1.2. FORMULACIÓN DEL PROBLEMA

A nivel mundial, diversos estudios han buscado entender la dinámica de las poblaciones aviares, su impacto en el medio ambiente y las variables de conservación de su ecosistema. En la última década, la investigación de este campo ha experimentado un incremento notable, con el aprovechamiento de la DS para analizar grandes volúmenes de información, el empleo de técnicas de modelado y visualización, y la exploración del potencial de los algoritmos de ML para comprender y prever comportamientos aviares [10, 11]. Sin embargo, son escasos los estudios que han contextualizado estos enfoques globales dentro del marco específico de la rica biodiversidad aviar colombiana y sus retos únicos de conservación.

Este proyecto tuvo como finalidad desarrollar una metodología específica que permitió predecir avistamientos de especies endémicas, integrando datos de observaciones con variables exógenas y utilizando algoritmos de ML. Para ello, se establecieron las siguientes preguntas de sistematización que permitieron concretar la metodología: ¿Es posible identificar departamentos de Colombia de alta concentración de especies endémicas de aves a partir de avistamientos? A su vez, ¿es factible desarrollar una metodología de gestión de datos que permita establecer relaciones sobre los avistamientos de especies endémicas en estos departamentos y variables exógenas asociadas (precipitaciones, temperaturas, antropogénicas, entre otras)? De esta forma, ¿es posible mediante algoritmos de ML estimar avistamientos de especies endémicas a partir de variables externas? Y finalmente, ¿se podrían generar estrategias de validación de los resultados de predicción que fortalezcan la interpretabilidad de los resultados obtenidos para revelar comportamientos espaciotemporales de especies endémicas? La resolución de estos interrogantes permitió abordar integralmente a la problemática, abarcando desde el preprocesamiento de datos hasta la implementación de modelos de distribución de especies, contribuyendo así al fortalecimiento del conocimiento de la biodiversidad aviar en Colombia. Así, se planteó la investigación que rigió este trabajo: ¿Es posible desarrollar una metodología de predicción de avistamientos de aves para aportar insumos para la conservación de especies endémicas a partir de algoritmos de ML?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Desarrollar un modelo de aprendizaje automático que permita predecir avistamientos de especies endémicas para identificar patrones temporales y espaciales en Colombia.

2.2. OBJETIVOS ESPECÍFICOS

- Desarrollar una metodología de preprocesamiento de datos de avistamientos de aves, con el fin de identificar departamentos de alta concentración de especies endémicas.
- Desarrollar una metodología de preprocesamiento de datos de variables exógenas (precipitaciones, temperaturas, antropogénicas, entre otras) para establecer correlaciones sobre los avistamientos de especies endémicas.
- Implementar algoritmos de aprendizaje automático para predecir avistamientos de especies endémicas a partir de variables exógenas.
- Implementar estrategias validación de resultados a partir de medidas de desempeño y técnicas de visualización de información para revelar comportamientos espaciales y temporales de especies endémicas.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

3.1.1. Ciencia de Datos y Biodiversidad: La DS ha emergido como una disciplina que aborda desafíos críticos de diversas áreas y proporciona las herramientas necesarias para el análisis, compresión y predicción de fenómenos a partir de datos. En la era digital actual, se erige como un facilitador en la resolución de problemas. Más del 50% de los artículos científicos publicados señalan una colaboración entre la DS y otros campos como la ingeniería, medicina, ciencias sociales, ciencias ambientales y de conservación, entre otros [20].

En el ámbito ambiental, la DS ha demostrado ser una herramienta invaluable para la comprensión y conservación de los ecosistemas. Investigaciones recientes han explorado cómo la aplicación de técnicas de DS puede mejorar significativamente la monitorización y gestión de la biodiversidad. Artículos científicos han destacado desarrollos específicos, como el uso de algoritmos de aprendizaje automático para predecir patrones de migración de especies, la utilización de imágenes satelitales para evaluar cambios en la cobertura vegetal y la implementación de modelos predictivos para identificar áreas críticas para la conservación [11].

En la intersección entre la DS y la biodiversidad, investigaciones han empleado análisis espaciales y temporales de conjuntos de datos masivos para comprender mejor los patrones de distribución de especies y las interacciones entre ellas y su entorno [12]. Estos enfoques permiten una toma de decisiones precisa y basada en evidencia para la gestión de áreas protegidas, la identificación de especies en riesgo y la evaluación del impacto de cambios ambientales. La colaboración entre científicos de datos y ecologistas ha dado lugar a nuevas formas de abordar desafíos urgentes en la conservación de la biodiversidad.

3.1.2. Rol de la Ciencia Ciudadana en la Biodiversidad: En el ámbito de la biodiversidad, la Ciencia Ciudadana (*Citizen Science*, CS) juega un papel crucial al involucrar a personas no especializadas en la recopilación de datos científicos. Un ejemplo destacado de esta participación es eBird, un proyecto global que permite a los observadores de aves contribuir con sus avistamientos a una base de datos masiva que alimenta investigaciones y decisiones de conservación.

eBird se basa en una idea simple pero poderosa: cada observador de aves posee un conocimiento y experiencia únicos que pueden contribuir al entendimiento científico de las aves [9]. A través de eBird, los usuarios registran listas de verificación de las aves que observan, incluyendo cuándo, dónde y cómo se realizaron las observaciones. Estos datos se archivan y se comparten libremente para impulsar nuevos enfoques científicos y de conservación basados en datos.

Con más de 100 millones de avistamientos de aves contribuidos anualmente por usuarios de todo el mundo, eBird es uno de los proyectos de ciencia ciudadana relacionados con la biodiversidad más grande a nivel global. Gestionado por el Cornell Lab of Ornithology, esta plataforma colabora con cientos de organizaciones, expertos regionales y usuarios, lo que asegura un crecimiento constante y una participación amplia.

Los datos de eBird se han utilizado en decisiones de conservación, artículos, proyectos estudiantiles, y siguen informando sobre aves a nivel mundial [9]. Al involucrar a la comunidad en la recolección de datos y enriquecer la base de conocimiento científico, no solo se amplían el alcance y la efectividad de los esfuerzos científicos, sino que también se fomenta una cultura de participación y responsabilidad ambiental.

3.1.3. Riqueza: La riqueza de especies de aves en un ecosistema se refiere al número total de especies diferentes presentes en ese lugar. Este concepto es crucial para medir la biodiversidad, ya que una mayor riqueza de especies suele indicar un ecosistema más saludable y equilibrado. Medir la riqueza de especies ayuda a identificar áreas de alta biodiversidad que requieren conservación, como se observa en los estudios de biodiversidad global que destacan la importancia de los hotspots biológicos [31].

3.1.4. Abundancia: La abundancia de especies de aves se refiere al número de individuos de cada especie en una región determinada. Este parámetro es esencial para entender la estructura y dinámica de las comunidades ecológicas. La abundancia no solo ayuda a evaluar la salud de las poblaciones de aves, sino también a diseñar y ajustar estrategias de conservación para especies amenazadas o en declive [32]**.**

3.1.5. Aprendizaje Automático (Machine Learning, ML): Los modelos predictivos son una alternativa para obtener resultados a mediano plazo usando herramientas estadísticas, informáticas y geográficas sobre la información biológica disponible para elaborar predicciones que permitan estimar la distribución de la diversidad biológica si no hay datos exhaustivos [13]. Los modelos han evolucionado desde su aplicación a especies aisladas hasta análisis de cientos o miles de taxones para combinarlos en el análisis de la biodiversidad y riqueza específica [14].

Para aplicar los modelos predictivos es importante comprender la fundamentación o conceptos del ML, Aprendizaje Supervisado y Aprendizaje No Supervisado. El aprendizaje se refiere a un amplio espectro de situaciones en las cuales el aprendiz incrementa su conocimiento o sus habilidades para cumplir una tarea. En este enfoque, un agente autónomo debe tener la capacidad de realizar una misma tarea de varias maneras, si es posible, y dependiendo de las circunstancias. Debe poder tomar decisiones sobre el curso apropiado para resolver un problema y modificarlas cuando las condiciones lo requieran. Por esto, uno de los objetivos centrales de esta área es construir sistemas que sean capaces de adaptarse dinámicamente y sin un entrenamiento previo a situaciones nuevas y aprender como resultado de resolver el problema (o problemas) que estas situaciones presentan [15]. Teniendo en cuenta lo anterior, se determina

que el aprendizaje automático es la ciencia de enseñar a las computadoras a hacer predicciones basadas en datos y pedirle que realice una predicción [5].

3.1.6. Aprendizaje No Supervisado: Con relación al aprendizaje no supervisado, este consiste en que los modelos descubran por sí mismos características, regularidades, correlaciones o categorías en los datos de entrada y se obtengan de forma codificada en la salida. En algunos casos, la salida representa el grado de similitud entre la información que se le está presentando en la entrada y la que se le ha mostrado en el pasado, es decir, que no necesitan un asesor externo para realizar su aprendizaje [16].

3.1.7. Aprendizaje Supervisado: Una de las subcategorías del ML es el Aprendizaje Supervisado que tiene como objetivo construir, a partir de los datos de entrenamiento un modelo de predicción. Usando este modelo, se puede predecir la variable respuesta de un nuevo conjunto de datos no vistos únicamente conociendo sus variables explicativas [17]. En el Aprendizaje Supervisado existen diferentes tareas, dentro de ellas la tarea de Clasificación y la de Regresión.

3.1.8. Clasificación y Regresión: La tarea de clasificación utiliza un algoritmo para asignar datos de prueba a categorías específicas. Reconoce entidades específicas dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo se deben etiquetar o definir esas entidades. Los algoritmos de clasificación más comunes son los clasificadores lineales, las Máquinas de Vectores de Soporte *(Support Vector Machines*, SVM), los Árboles de Decisión y los Bosques Aleatorios. Por otra parte, la tarea de Regresión se utiliza para comprender la relación entre variables dependientes e independientes. Se usa comúnmente para realizar proyecciones, como ingresos por ventas para una empresa determinada. Algunos de los algoritmos de este tipo de tarea que son ampliamente utilizados son la Regresión lineal, Regresión Logística y Regresión Polinómica [18].

3.1.9. Integración de Datos Geoespaciales: Los datos geoespaciales incluyen información geográfica que puede ser visualizada y analizada mediante Sistemas de Información Geográfica (SIG), lo cual permite mapear y modelar la distribución de especies, identificar hábitats críticos y evaluar el impacto de los cambios ambientales.

Uno de los principales beneficios de la integración de datos geoespaciales es la capacidad de combinar diversas fuentes de información, como imágenes satelitales, datos de sensores remotos y observaciones in situ. Esta combinación permite una comprensión de los patrones ecológicos y los procesos que afectan a las aves. Mediante el uso de imágenes satelitales, los científicos pueden monitorear cambios en la cobertura vegetal, detectar deforestación y degradación de hábitats, y evaluar el impacto de fenómenos climáticos extremos sobre los ecosistemas aviares.

La integración de datos geoespaciales con técnicas de DS y ML potencia el análisis predictivo y la modelación de escenarios futuros. Los algoritmos de ML pueden identificar tendencias y relaciones complejas en los datos, facilitando la predicción de cambios en la distribución de aves

debido al cambio climático, la urbanización y otras presiones antropogénicas. Este enfoque permite a los investigadores y gestores de conservación diseñar estrategias más efectivas y basadas en evidencia para la protección y restauración de las poblaciones de aves [22].

Proyectos como eBird utilizan coordenadas geográficas para mapear avistamientos de aves, generando mapas de distribución en tiempo real y detectando patrones migratorios. Estos datos, integrados con otras fuentes geoespaciales, enriquecen los análisis y mejoran la capacidad de respuesta ante cambios ambientales. La integración de datos geoespaciales es una herramienta poderosa que amplia nuestra capacidad para estudiar y conservar avifauna [22].

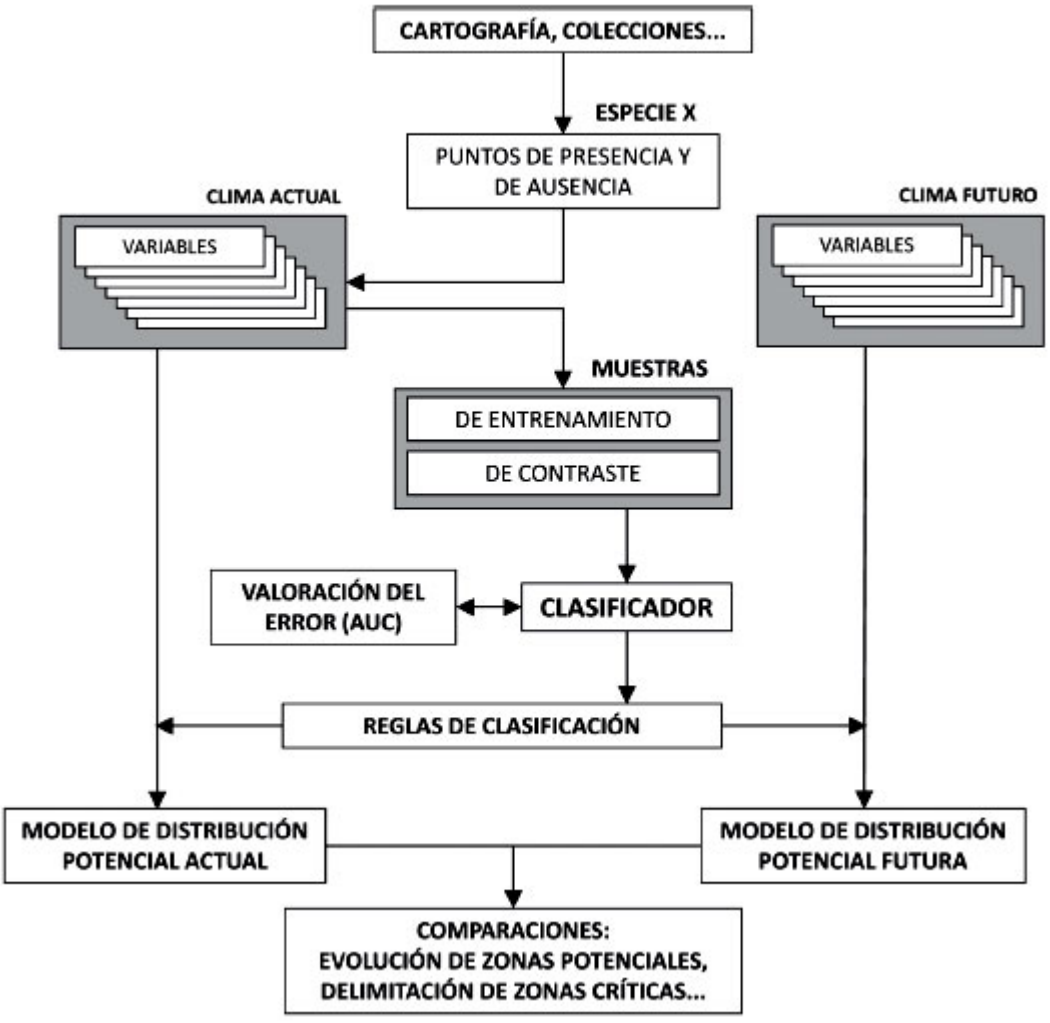


Figura 1. Flujo de trabajo para la realización de modelos de distribución de especies [21].

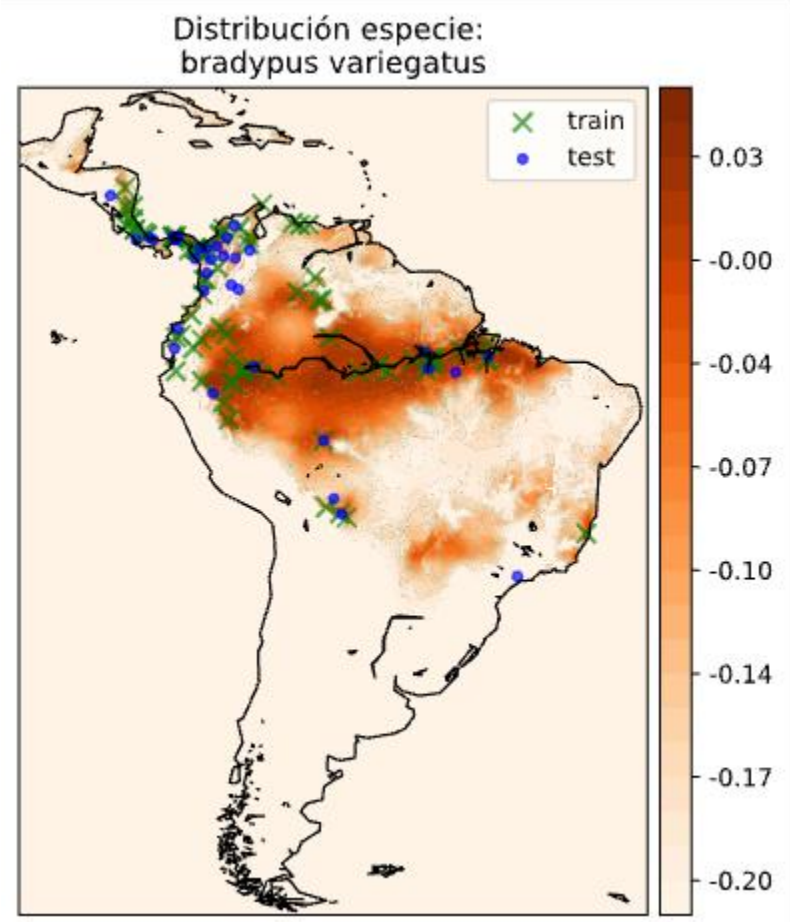


Figura 2. Distribución de oso perezoso en Suramérica [Elaboración propia]

3.1.10. Modelos de Distribución de Especies (*Species Distribution Models*, SDM): Los SDM son representaciones cartográficas de un espacio para la presencia de una especie en función de las variables empleadas para generar dicha representación [19]. Estos muestran información que sintetiza relaciones entre especies y variables ambientales que serían difíciles de interpretar o incluso de apreciar por otros medios. La capacidad de construir modelos más realistas está limitada por la comprensión de los sistemas ecológicos y por los datos disponibles, que son insuficientes. A pesar de estas limitaciones, un elevado número de estudios han demostrado su utilidad en campos en los que pocas técnicas pueden brindar ayuda para entenderlos, como predecir la presencia de especies aún no descritas [20]. Para la construcción de modelos de distribución de especies se realiza en una serie de pasos [21] como se muestra en la Fig. 1, cada uno de los cuales presenta múltiples alternativas de ejecución que influyen en la calidad del resultado final. En un primer paso, los datos conocidos sobre la distribución del organismo se asocian estadísticamente con diferentes variables independientes que describen las condiciones ambientales. De existir esta relación, se extrapola al resto del área de estudio y se obtiene un valor en cada lugar que suele interpretarse como la probabilidad de presencia de la especie en

ese punto. En realidad, solo señalan la similitud ambiental de cada punto del terreno con las zonas de presencia actual de la especie. La "probabilidad de presencia" es una interpretación abusiva de la medida de similitud ambiental que debería interpretarse como valor de idoneidad para el desarrollo de la especie. Puede que el modelo delimite zonas potenciales muy alejadas geográficamente de las actuales; la probabilidad de encontrar la especie en ellas no es a priori alta, aunque las condiciones ambientales fueran favorables [21].
La evaluación del resultado final de un modelo de distribución de especies y la comparación entre los diferentes métodos aplicables al problema se realiza mediante estadísticos que miden el desempeño y la consistencia del modelo en cuanto a su capacidad de discriminar entre los datos de entrada (presencias y ausencias o pseudoausencias) y datos independientes de contraste [21]. A continuación, se presenta la distribución del oso perezoso en Suramérica (Fig. 2) por medio de un modelo de distribución de especie bajo elaboración propia.

3.1.11. Factor de Inflación de Varianza (VIF): Para abordar el problema de multicolinealidad, se decidió aplicar el Factor de Inflación de Varianza (VIF), una técnica estadística utilizada en análisis de regresión lineal para cuantificar la multicolinealidad entre variables independientes [40]. Este índice evalúa cuánto aumenta la varianza de los coeficientes de un modelo debido a la colinealidad, utilizando la fórmula:

$$VIF = \frac{1}{1 - R^2}$$

Donde R^2 es el coeficiente de determinación obtenido de una regresión auxiliar que modela una variable en función de las demás. Un VIF de 1 indica ausencia de multicolinealidad, valores entre 1 y 5 sugieren colinealidad moderada, y valores superiores a 5 o 10 evidencian colinealidad alta, lo que puede afectar la fiabilidad de las estimaciones.

3.1.12. Aprendizaje de Máquina utilizados en el contexto de los Modelos de Distribución de Especies

Algoritmos de Máxima Entropía (MaxEnt): se basa en el principio de maximizar la entropía *H(p)*, ajustando un modelo que describe la probabilidad de ocurrencia de una especie sin asumir distribuciones previas más allá de los datos disponibles. Es ampliamente utilizado por su capacidad para trabajar con datos de ocurrencia y su flexibilidad para incorporar variables ambientales [41].

$$H(p) = -\sum_i p_i \log\left(p_i\right)$$

Donde p_i es la probabilidad estimada de ocurrencia en la celda *i.*

Redes neuronales (Deep Learning): consisten en modelos inspirados en la estructura del cerebro

humano que capturan relaciones complejas en los datos mediante múltiples capas de procesamiento. Aunque son poderosas para tareas de clasificación, su falta de interpretabilidad limita su aplicación en estudios donde se requiere comprender los factores subyacentes [42].

$$y = f(Wx + b),$$

Donde x es el vector de entrada, W es la matriz de pesos, b es el sesgo (bias), f es la función de activación (por ejemplo, ReLU, sigmoide o softmax), e y es la salida (predicción) de la capa.

Modelos basados en árboles (Random Forest): utilizan múltiples árboles de decisión $T_K(x)$ donde cada árbol T_K es construido a partir de un subconjunto aleatorio de los datos. La predicción del modelo es el promedio (para regresión) o el voto mayoritario (para clasificación) de los árboles individuales para realizar predicciones robustas, permitiendo identificar la importancia relativa de las variables explicativas. Su robustez y precisión los hacen populares en plataformas como eBird para generar mapas de distribución de especies [43].

$$y = \frac{1}{K}\sum_{k=1}^{K} T_k(x)$$

Donde *y* es la predicción final, *K* es el número de árboles y $T_K(x)$ es la predicción de k-ésimo árbol.

Regresión Logística: este modelo estadístico se utiliza para problemas de clasificación binaria, modelando la relación entre variables explicativas y la probabilidad de un resultado. Su simplicidad y facilidad de interpretación lo convierten en una herramienta clave en estudios ecológicos. Este modelo es equivalente al de MaxEnt [44].

$$P(y = 1x) = \frac{1}{1 + e^{-(Bo+B1x1+B2x2+\cdots+BpXp)}}$$

Donde *x1, x2, xp* son las variables explicativas del modelo, Bo es el Intercepto y *B1, B2, Bp* son los coeficientes asociados a las variables explicativas.

3.1.13. Métricas de evaluación de desempeño de Clasificación: Para evaluar el desempeño de los modelos utilizados, se emplearon métricas tradicionalmente aceptadas en problemas de clasificación como precisión, exhaustividad, exactitud, F1-Score y la matriz de confusión. Estas herramientas permiten analizar de manera integral la efectividad de las predicciones realizadas. A continuación, se describen brevemente cada una de ellas.

Matriz de Confusión: tabla que muestra la distribución de predicciones en verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN), permitiendo evaluar errores y aciertos en detalle [45].

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

Figura 3. Matriz de confusión, métrica de evaluación de desempeño de modelos de predicción.

Precisión: proporción de predicciones positivas correctas respecto al total de predicciones positivas realizadas. Mide la exactitud de las predicciones positivas [46].

$$\text{Precision} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Sensibilidad (Recall): Proporción de observaciones positivas correctamente identificadas. Indica la capacidad del modelo para detectar casos positivos [47].

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Exactitud (Accuracy): Porcentaje de predicciones correctas sobre el total de observaciones. Refleja el desempeño general del modelo [47].

$$\text{Accuracy} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}$$

F1-Score: Promedio armónico entre precision y recall, útil en conjuntos de datos desbalanceados [48].

$$\text{F1} - \text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

3.1.14. Manejo de desbalance de clases, implementación de técnicas de sobre muestreo: Para abordar el desbalance en los datos y mejorar la sensibilidad hacia la clase minoritaria, se implementaron dos técnicas de sobremuestreo: Random OverSampler y SMOTE (Synthetic Minority Over-sampling Technique).

Random OverSampler es una técnica que aumenta la representación de la clase minoritaria duplicando aleatoriamente las observaciones existentes de dicha clase. Este enfoque tiene como objetivo equilibrar la proporción de clases en el conjunto de entrenamiento, incrementando la cantidad de datos disponibles para entrenar los modelos en la clase menos representada. Sin embargo, al tratarse de una duplicación directa de los datos originales, esta técnica puede introducir redundancias en el conjunto de datos, lo cual puede limitar la capacidad del modelo para generalizar correctamente cuando se enfrenta a datos nuevos.

SMOTE aborda el desbalance de clases mediante la creación de nuevas observaciones sintéticas para la clase minoritaria [49]. A diferencia de Random OverSampler, SMOTE no realiza duplicaciones directas, en su lugar, genera nuevos datos a través de interpolaciones entre observaciones existentes de la clase minoritaria. Este enfoque permite preservar e incluso incrementar la diversidad de las características dentro de los datos balanceados, lo que contribuye a un entrenamiento más efectivo de los modelos. Al introducir variabilidad en las muestras adicionales, SMOTE reduce el riesgo de que los modelos se sobreajuste.

3.2. ANTECEDENTES

Los aportes de la DS y el ML a campos como la biodiversidad y la ecología han tomado fuerza recientemente. Además, uno de los ejes propulsores de la intersección entre estas ramas del conocimiento ha sido el desarrollo de proyectos de CS. De hecho, en el trabajo de [22] se usan datos ecológicos recopilados por el público en general para abordar una amplia gama de investigaciones ecológicas y generar insumos para la planificación de la conservación, gracias al rápido aumento en el alcance y volumen de datos disponibles. Sin embargo, los datos de proyectos a gran escala con observadores voluntarios, como eBird, presentan varios desafíos que pueden obstaculizar inferencias ecológicas sólidas. Estos desafíos incluyen sesgo espacial, variación en el esfuerzo y sesgo en la notificación de especies. Por ello, en este proyecto estimaron dos métricas ampliamente utilizadas de distribuciones de especies: tasa de encuentro y probabilidad de ocupación. Para cada métrica, se evaluó críticamente el impacto de los pasos de procesamiento de datos que degradan o refinan los datos utilizados en los análisis. La densidad de datos de CS varía mucho en todo el mundo, por lo que probaron si las diferencias en el rendimiento del modelo eran robustas según el tamaño de la muestra. En cuanto al aporte del proyecto para la presente investigación se destaca la información suministrada con relación a los desafíos que se pueden encontrar al trabajar con una base de datos como la de eBird, como son los sesgos ya mencionados. Además, se hace referencia a la implementación de algoritmos automáticos para estimar distribuciones de especies y evaluaciones del impacto de los pasos de preprocesamiento de datos en la estimación de distribuciones de especies. Sin embargo, este

trabajo difiere respecto al que aquí se propone respecto al área geográfica en la que se enfocan los proyectos, y el uso de variables exógenas para estimar avistamientos de especies endémicas.

La CS, representada por proyectos como eBird, ha demostrado ser fundamental para superar estos desafíos mediante la participación de la comunidad en la recopilación de datos. El involucramiento de observadores de aves de todo el mundo ha permitido la creación de una base de datos robusta y diversa que no solo facilita la investigación científica, sino que también potencia la toma de decisiones en conservación. En 2014 un estudio utilizó datos de eBird para identificar áreas críticas para la conservación de aves migratorias en América del Norte, lo que resultó en la protección de varios hábitats importante [23]. La colaboración entre ciudadanos y científicos ha generado una red global de monitoreo de aves, lo que ha mejorado significativamente la calidad y cantidad de datos disponibles para el análisis ecológico y la planificación de la conservación.

La integración de datos geoespaciales con técnicas avanzadas de DS y ML ha revolucionado la forma en que entendemos y gestionamos la biodiversidad. Los SIG y el uso de imágenes satelitales permiten mapear y modelar la distribución de especies de manera más precisa y en tiempo real. Al combinar estos datos con algoritmos de ML, es posible predecir cambios en los patrones de distribución de aves y evaluar el impacto de factores ambientales como el cambio climático. Por ejemplo, en 2018, investigadores utilizaron datos geoespaciales y algoritmos de ML para predecir el impacto del cambio climático en la distribución de especies de aves en Europa, ayudando a desarrollar estrategias de conservación basadas en datos [24]. Esta integración no solo mejora la capacidad de respuesta ante cambios ambientales, sino que también proporciona una base fáctica para el diseño de estrategias de conservación basadas en evidencia científica.

Por otra parte, en el estado del arte se evidencia que otro de los factores al tener en cuenta al momento de formular SDMs es la inclusión de variables exógenas que permitan explicar ciertos comportamientos de las especies. Por ejemplo, en el trabajo de [25] se establece que los SDM son esenciales para conservar la biodiversidad, con aplicaciones importantes como priorización espacial de acciones de conservación y elucidación de las relaciones entre los predictores ambientales y las respuestas de las especies. Estos modelos son más útiles para los gestores de la conservación cuando incluyen factores externos como el fuego y sus posibles implicaciones. En este estudio, se recopilo un conjunto de registros de mamíferos de una región propensa a incendios en el sudeste de Australia, donde estas especies han experimentado declives en las últimas décadas. Se utilizaron SDM para (1) determinar la influencia relativa del clima, el fuego, la vegetación y la topografía en las distribuciones de mamíferos terrestres; (2) determinar las respuestas de las especies al tiempo transcurrido desde el último incendio y; (3) proporcionar predicciones espaciales de la idoneidad del hábitat para la planificación de la conservación. Esta investigación aporta información relevante sobre la implementación de SDMs para la conservación de la biodiversidad, e información sobre la importancia de incluir variables dinámicas, en los SDMs. Sin embargo, como ya se ha expuesto, el ámbito geográfico es

determinante, así como las especies de interés en tanto que condicionarán la interpretación de los resultados.

Gracias a la inclusión de disciplinas de AI dentro de la ecología y la biodiversidad, se han generado insumos para el diseño de planes de conservación. Por ejemplo, en [26] se resalta la necesidad urgente de políticas de conservación que maximicen la protección de la biodiversidad para sostener sus diversas contribuciones a la vida de las personas. Se presenta un marco para la priorización espacial de la conservación basado en aprendizaje por refuerzo que supera consistentemente al software disponible mediante el uso de datos simulados y empíricos. La metodología, Priorización de Áreas de Conservación a través de Inteligencia Artificial (CAPTAIN, por sus siglas en inglés), cuantifica el equilibrio entre los costos y beneficios de la protección del área y la biodiversidad, permitiendo la exploración de múltiples métricas de biodiversidad. Como producto de esta investigación, se estableció que la IA resulta promisoria para mejorar la conservación y el uso sostenible de los valores biológicos y ecosistémicos en un mundo que cambia rápidamente y tiene recursos limitados. En el marco de esta propuesta, este proyecto aporta información sobre el potencial del ML para la conservación de especies endémicas en Colombia. No obstante, hay unas diferencias que resaltan como: (1) la escala del área geográfica de estudio, ya que el proyecto de referencia se centra en la predicción de tendencias poblacionales de aves a nivel mundial y (2) el proyecto de referencia utilizó datos de avistamientos de aves a nivel mundial, datos de registros de aves simulados y empíricos y datos de variables ambientales para predecir tendencias poblacionales de aves.

Como se ha notado, el potencial de la DS, el ML y la AI para abordar el problema es trasladable a múltiples escenarios, donde se consideran especies tan diversas como mamíferos hasta anfibios. Por ejemplo, en [27] se busca dar a conocer cómo el aumento de la urbanización en los territorios está amenazando la biodiversidad nativa de las especies endémicas de regiones en Quito, Ecuador, causando fragmentación de su ecosistema y haciendo que pierdan sus hábitats naturales. La investigación se centra en la rana marsupial andina del centro de la región y alrededores, que se ha visto obligada a adaptarse a ciertos ambientes alterados por el desplazamiento realizado por la ampliación y urbanización del territorio, por eso, mediante modelos biogeográficos generados con datos obtenidos en plataformas web de acceso libre, se busca evaluar el impacto de los cambios y alteraciones asociadas a las condiciones climáticas y medioambientales causadas por el desplazamiento y adaptación a nuevas formas de supervivencia. En consecuencia, este trabajo dio un aporte importante para el desarrollo del proyecto actual, ya que expone el problema mundial sobre el desplazamiento de especies endémicas por la expansión de las grandes ciudades y su industrialización. También brindó herramientas y estrategias que ayudaron positivamente para desarrollar el proyecto, como el uso de plataformas de acceso libre para obtener información relevante de las diferentes especies de las regiones. Además, se aplicaron modelos y herramientas de software para realizar predicciones de condiciones medioambientales y de espacios óptimos para la supervivencia de cada especie sobre la cual se realizó propiamente el estudio. Sin embargo, hay diferencias respecto a la

propuesta en tanto que se particulariza a una única especie y en considerar un único factor de impacto en la distribución del ejemplar analizado.

Otro ejemplo de aplicación de la DS y el ML a los SDMs se evidencia en el trabajo de [28]. En este proyecto se menciona en primera instancia que el cambio climático global está causando impactos sin precedentes en la biodiversidad. En este estudio, se demuestra la aplicabilidad de los datos de eBird y del Sistema Global de Información sobre Biodiversidad (*Global Biodiversity Information Facility*, GBIF), para producir pronósticos a escala nacional para examinar los posibles impactos del CC en la avifauna terrestre en India. Utilizando datos recopilados por científicos ciudadanos, se desarrollaron SDMs optimizados y se predijeron 1091 especies de aves terrestres que se distribuirían en India para 2070 en dos superficies climáticas (RCP 4.5 y 8.5), utilizando algoritmos de distribución de especies basados en Máxima Entropía (MaxEnt). Este estudio ha resultado en mapas de alta resolución de la riqueza de especies de aves terrestres en toda la India, así como en predicciones de cambios predominantemente hacia el norte en los rangos de las especies, similares a las predicciones hechas para la avifauna en otras regiones, como Europa y EE. UU. Este estudio es relevante dado que evidencia el desarrollo de un modelo de ML con el fin de predecir cómo impactará el cambio climático a la distribución de especies de aves en India, siendo también relevante la forma como preprocesan los datos de las variables exógenas. Asimismo, el trabajo muestra que los datos de CS como los de eBird y la integración de variables ambientales permiten intuir los posibles impactos del CC. No obstante, este trabajo se diferencia en que abordan únicamente variables asociadas al cambio climático, dejando de lado otro tipo de factores externos que quizás también contribuyen a la alteración de las distribuciones de las especies.

4. METODOLOGÍA

La metodología de este trabajo se puede resumir en el Diagrama de Flujo de la Figura 4. Este esquema sintetiza de manera visual las diferentes etapas de la metodología, desde la descarga de los datos hasta la generación de los resultados finales, facilitando una comprensión clara del flujo de trabajo implementado que se explicara a lo largo de este capítulo.

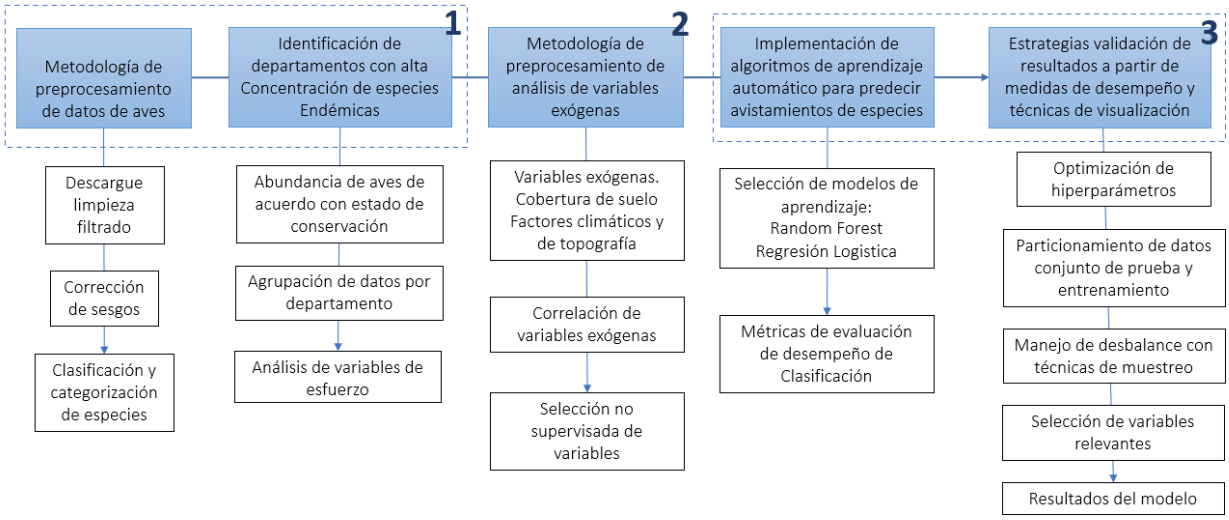


Figura 4. Diagrama de flujo resumen de metodología.

4.1. DESARROLLAR UNA METODOLOGÍA DE PREPROCESAMIENTO DE DATOS DE AVISTAMIENTOS DE AVES, CON EL FIN DE IDENTIFICAR DEPARTAMENTOS DE ALTA CONCENTRACIÓN DE ESPECIES ENDÉMICAS

El análisis de los avistamientos de aves en Colombia se basó en los datos proporcionados por el portal eBird, una plataforma colaborativa que recopila observaciones de aves a nivel mundial [9]. Dada la naturaleza participativa de eBird, los datos presentaron variaciones en calidad y consistencia, lo que requirió la implementación de estrategias de preprocesamiento antes de abordar cada uno de los objetivos planteados. Estas estrategias incluyeron la descarga, el filtrado y el ajuste de los registros, la corrección de inconsistencias geográficas y temporales, así como la clasificación de las especies endémicas según su estado de conservación. Estas acciones se realizaron para garantizar que los datos utilizados fueran confiables y adecuados para los análisis, reflejando de manera precisa las observaciones dentro del área de estudio.

4.1.1. Materiales y métodos

Se creó una cuenta en el portal eBird y se realizó una solicitud formal de los registros geográficos correspondientes a Colombia. Los datos fueron entregados en dos conjuntos principales: el Sampling Event Data (SED), que proporciona listas de verificación de las especies observadas, y el

eBird Basic Dataset (EBD), que contiene registros individuales de observaciones. El portal de eBird ofrece una guía para utilizar sus datos con el fin de extraer información a partir de estos. [29].

4.1.2. Configuración experimental o descripción de la base de datos

Descarga y filtrado inicial de datos

A partir de la importación de datos de eBird, se procedió a aplicar una serie de filtros para asegurar la validez y consistencia de las observaciones. En primer lugar, se filtraron los registros en función de la duración del esfuerzo de observación, limitando esta variable a un máximo de seis horas, lo cual permitió excluir observaciones cuya duración excesiva podría haber introducido sesgos o inconsistencias. Del mismo modo, se aplicó un filtro a la distancia recorrida durante las observaciones, estableciendo un límite máximo de 10 kilómetros. Este criterio se utilizó para garantizar que las observaciones se realizaran en un área concentrada, minimizando la posibilidad de errores derivados de desplazamientos extensos. Además, se restringió el número de observadores a un máximo de 10 personas por evento de observación, con el fin de mantener la homogeneidad del esfuerzo entre los distintos registros.

A su vez, se eliminaron registros cuya ubicación correspondiera a Colombia continental, excluyendo aquellas situadas fuera de los límites definidos o con coordenadas imprecisas. Con estos filtros, se aseguró que los datos utilizados reflejaran un esfuerzo de observación adecuado, tanto en términos de calidad geoespacial como temporal.

Corrección de sesgos en los datos de avistamientos

Los datos obtenidos de eBird presentaron varios tipos de sesgos derivados de la naturaleza colaborativa de su recolección. Para garantizar la validez de los análisis, se implementaron estrategias específicas para mitigar cada uno de los sesgos identificados:

- **Sesgo taxonómico:** los observadores suelen registrar ciertas especies frecuentemente debido a preferencias personales, lo que puede llevar a una representación desbalanceada. Para mitigar este sesgo, se dio prioridad a las listas de verificación completas, que incluyen todas las especies observadas en un evento de muestreo, lo que aseguró un esfuerzo de observación uniforme.

- **Sesgo espacial:** los datos tienden a concentrarse en áreas urbanas o zonas populares para la observación de aves [22]. Este problema se abordó mediante la aplicación de un filtrado espacial que garantizó la inclusión de observaciones exclusivamente dentro de los límites geográficos de Colombia.

- **Sesgo temporal:** los observadores tienden a reportar más datos durante fines de semana o en épocas de migración, lo que afecta la distribución temporal de los registros. Se aplicaron

filtros temporales para normalizar la distribución de las observaciones en el tiempo, equilibrando así los datos según las estacionalidades y evitando que picos de observaciones en periodos específicos distorsionaran el análisis.

- **Desbalance de clases:** las especies raras o difíciles de detectar suelen estar subrepresentadas. Para abordar este desafío, se implementó el método de zero filling, que permite inferir la ausencia de una especie en una lista de verificación, diferenciando entre no detección y ausencia real.

Clasificación y categorización de especies endémicas

Una vez preprocesados los datos de observaciones, se procedió a clasificar las especies endémicas de Colombia en función de su estado de conservación, tomando como referencia la lista de especies silvestres amenazadas de la diversidad biológica continental y marino-costera de Colombia expedida por el Ministerio de Ambiente y Desarrollo Sostenible (MADS), con base en la resolución 0126 del 06 de febrero de 2024 [30]. Las categorías para especies amenazadas presentadas en esta resolución se describen en Tabla 1.

Estado de conservación	Definición
Peligro Crítico (CR)	Aquellas que están enfrentando un riesgo de extinción extremadamente alto en estado de vida silvestre.
En Peligro (EN)	Aquellas que están enfrentando un riesgo de extinción muy alto en estado de vida silvestre.
Vulnerable (VU)	Aquellas que están enfrentando un riesgo de extinción alto en estado de vida silvestre.
Casi Amenazada (NT)	Aquellas que están cerca de cumplir con los criterios para considerarse en alguna de las categorías de mayor riesgo.
Preocupación Menor (LC)	Aquellas que no cumplen con los criterios para considerarse en ninguna de las categorías anteriores y se consideran de menor riesgo.

Tabla 1. Categorías de aves según nivel de amenaza

Por otra parte, se descargó la lista de aves de Colombia 2024 con última actualización a enero de 2024 del Comité Colombiano de Registros Ornitológicos, con el fin de determinar las aves endémicas de Colombia. Este proceso se llevó a cabo con la colaboración de un experto en ornitología perteneciente a la Asociación Ornitológica del Huila-ASORHUI, cuya asistencia fue fundamental para filtrar especies de interés. Una vez obtenidas las listas, se filtraron las aves endémicas en Colombia categorizándolas por nivel de amenaza, este proceso fue esencial para generar conjuntos de datos de las aves relevantes para la investigación. Finalmente se consolidó en un archivo Microsoft Excel la lista de especies endémicas por estado de conservación como se puede evidenciar la Tabla 2.

Estado de conservación	Aves endémicas de Colombia
Peligro Crítico (CR)	*Atlapetes blancae, Chrysuronia lilliae, Coeligena orina, Crax alberti, Eriocnemis isabellae, Eriocnemis mirabilis, Grallaria urraoensis, Hapalopsittaca fuertesi, Henicorhina negreti, Lipaugus weberi, Oxypogon cyanolaemus, Thryophilus nicefori, Troglodytes monticola.*
En Peligro (EN)	*Atlapetes flaviceps, Bangsia aureocincta, Campylopterus phainopeplus, Cistothorus apolinari, Diglossa gloriosissima, Grallaria kaestneri, Leptotila conoveri, Macroagelaius subalaris, Myiotheretes pernix, Odontophorus strophium, Penelope perspicax, Phylloscartes lanyoni, Psarocolius cassini, Pyrrhura viridicata, Rallus semiplumbeus, Ramphomicron dorsale, Saucerottia castaneiventris, Scytalopus canus, Scytalopus rodriguezi.*
Vulnerable (VU)	*Anthocephala berlepschi, Anthocephala floriceps, Bangsia melanochlamys, Bolborhynchus ferrugineifrons, Capito hypoleucus, Chlorochrysa nitidissima, Coeligena prunellei, Dacnis hartlaubi, Grallaria bangsi, Grallaria milleri, Hypopyrrhus pyrohypogaster, Myiothlypis basilica, Oxypogon stuebelii, Pyrrhura calliptera, Synallaxis fuscorufa, Vireo caribaeus.*
Casi Amenazada (NT)	*Arremon basilicus, Atlapetes fuscoolivaceus, Bucco noanamae, Clibanornis rufipectus, Drymophila caudata, Drymophila hellmayri, Habia gutturalis, Myiothlypis conspicillata, Odontophorus hyperythrus.*
Preocupación Menor (LC)	*Anisognathus melanogenys, Atlapetes melanocephalus, Cercomacroides parkeri, Chaetocercus astreans, Chlorostilbon olivaresi, Coeligena phalerata, Cranioleuca hellmayri, Euphonia concinna, Habia cristata, Henicorhina anachoreta, Megascops gilesi, Melanerpes pulcher, Myiarchus apicalis, Myioborus flavivertex, Ortalis columbiana, Ortalis garrula, Oxypogon guerinii, Picumnus granadensis, Saucerottia cyanifrons, Scytalopus alvarezlopezi, Scytalopus latebricola, Scytalopus sanctaemartae, Scytalopus stilesi, Synallaxis subpudica, Vireo approximans.*

Tabla 2. Listado de especies endémicas en Colombia de acuerdo con nivel de amenaza

Exportación de datos

Tras aplicar los filtros y organizar la información por estado de conservación, los datos fueron exportados y almacenados en archivos CSV uno por cada categoría. Para efectos de maniobrabilidad en términos de cantidad de datos e importancia ecológica, en este trabajo se hizo un análisis específico de las especies clasificadas en las tres primeras categorías: Peligro Crítico (CR), En Peligro (EN) y Vulnerable (VU), debido a su alto nivel de amenaza y relevancia para la conservación. Aunque las especies clasificadas como Casi Amenazada (NT) y Preocupación Menor (LC) se incluyen en la tabla para proporcionar una visión completa de las aves endémicas de Colombia, el foco del análisis estará en aquellas en mayor riesgo, específicamente las especies en las categorías de Peligro Crítico (CR), En Peligro (EN) y Vulnerable (VU).

Estas categorías son fundamentales porque representan especies que enfrentan un alto riesgo de extinción si no se toman medidas de conservación. Priorizar el análisis de estas especies y generar predicciones sobre su ubicación y temporalidad mediante el uso de algoritmos de ML es esencial para guiar la toma de decisiones en la conservación, permitiendo a los expertos en biodiversidad focalizar sus esfuerzos en áreas críticas y en momentos clave. De esta manera, se optimizan los

recursos destinados a la protección de estas especies y los ecosistemas que dependen de ellas.

Identificación de departamento de alta concentración de especies endémicas

El análisis se realizó empleando técnicas de procesamiento de datos geoespaciales mediante Python. El objetivo fue generar análisis de estadística descriptiva y gráficas sobre la abundancia de especies endémicas en diferentes regiones del país, facilitando la identificación de las áreas con mayor riqueza de especies y brindando una base para los análisis posteriores. El procedimiento descrito a continuación fue aplicado de manera uniforme a las primeras tres categorías de conservación Peligro Crítico (CR), En Peligro (EN) y Vulnerable (VU), siguiendo un enfoque estandarizado para garantizar la coherencia en el tratamiento de los datos.

Agrupación de datos por departamento

Se agruparon los datos por departamento, con el fin de identificar las áreas geográficas con la mayor concentración de especies endémicas. Se realizó la intersección de las coordenadas de los avistamientos con los límites geográficos de los departamentos de Colombia, lo que permitió asignar cada registro a su departamento correspondiente. Este proceso fue repetido para las tres categorías de conservación (CR, EN, VU), utilizando los mismos procedimientos en cada caso.

La agrupación incluyó la suma de las observaciones para cada especie en cada departamento. Este enfoque permitió generar un panorama preliminar la distribución espacial de las especies endémicas, facilitando la identificación de patrones de concentración y áreas de alta densidad de observaciones. Se procedió a la creación de mapas de calor para visualizar la distribución por departamento de las especies endémicas en Colombia. Los mapas de calor permitieron integrar capas geográficas con los datos de abundancia obtenidos en el paso anterior.

Análisis exploratorio de variables de esfuerzo

Además de la visualización de la abundancia de especies, se llevó a cabo un análisis exploratorio para examinar cómo las variables relacionadas con el esfuerzo de observación podían influir en los resultados. Este análisis incluyó variables como la duración de las observaciones, la distancia recorrida, el número de observadores, y el esfuerzo total en términos de kilómetros y horas de observación. Este análisis es importante desarrollarlo ya que determinará si se consideran o no las variables de esfuerzo para la etapa de entrenamiento de los modelos de ML.

4.1.3. Resultados y discusión

Una vez cargados y verificados los datos, se procedió identificar la especie de ave más abundante por categoría, para esto se elaboró un gráfico de barras que representara la abundancia de aves endémicas en Colombia durante el período 2003-2023. En la Figura 5, el eje horizontal muestra los nombres científicos de las especies endémicas, mientras que el eje vertical indica la cantidad

de observaciones registradas. Los colores utilizados en el gráfico representan los diferentes niveles de amenaza para la conservación de las especies. Los mapas de calor también permitieron resaltar diferencias entre los departamentos, ayudando a identificar las áreas prioritarias para estudios detallados o esfuerzos de conservación.

Del análisis de la gráfica, se observó que las categorías con mayor nivel de amenaza En Peligro Crítico (CR) y En Peligro (EN) presentan un menor número de observaciones en comparación con las categorías de menor riesgo. Asimismo, se identificó la especie con mayor abundancia en cada una de las tres categorías principales evaluadas en este estudio. Para la categoría CR, el ave con mayor número de registros es *Henicorhina negreti*; en la categoría EN, fue *Penelope perspicax*; y en la categoría VU la especie predominante es *Hypopyrrhus pyrohypogaster*.

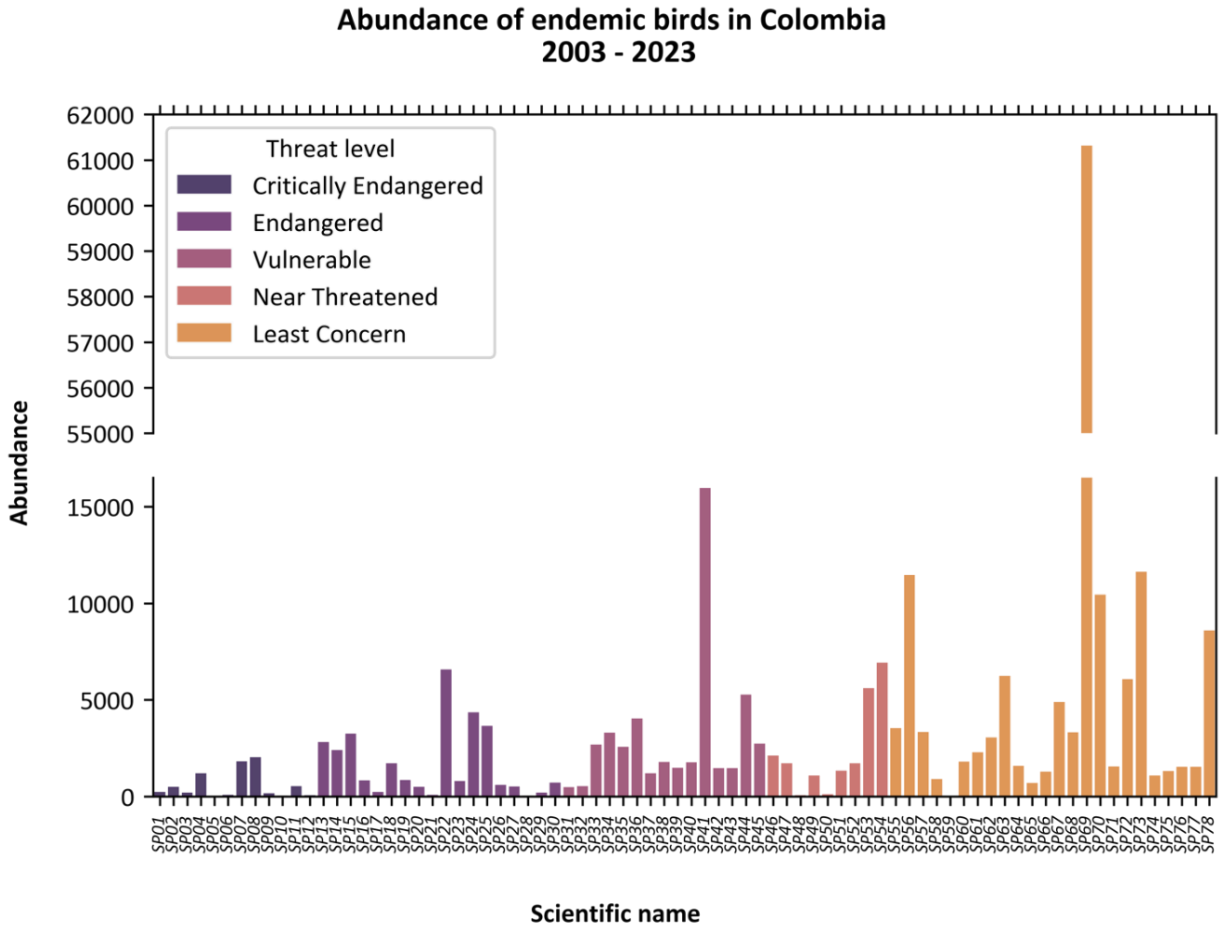


Figura 5. Abundancia de aves endémicas en Colombia 2003 – 2023. SPXX: Código del nombre científico de la especie. Las especies seleccionadas corresponden a: SP08: *Henicorhina negreti*. SP22: *Penelope perspicax*. SP41: *Hypopyrrhus pyrohypogaster*.

En la Figura 6 se presenta un mapa de calor que permite visualizar la distribución de avistamiento

de aves endémicas para en Colombia por departamento durante el período 2003-2023 para la categoría de Vulnerable (VU). En el eje horizontal se ubican los nombres científicos de las especies de la categoría y en el eje vertical los departamentos. Se observa que la especie *Hypopyrrhus pyrohypogaster* tiene una mayor abundancia en el departamento de Antioquia. Esta especie cuenta con un volumen de datos significativo para los fines de este trabajo, lo que permite utilizar sus registros en la aplicación de algoritmos de ML en fases posteriores del trabajo.

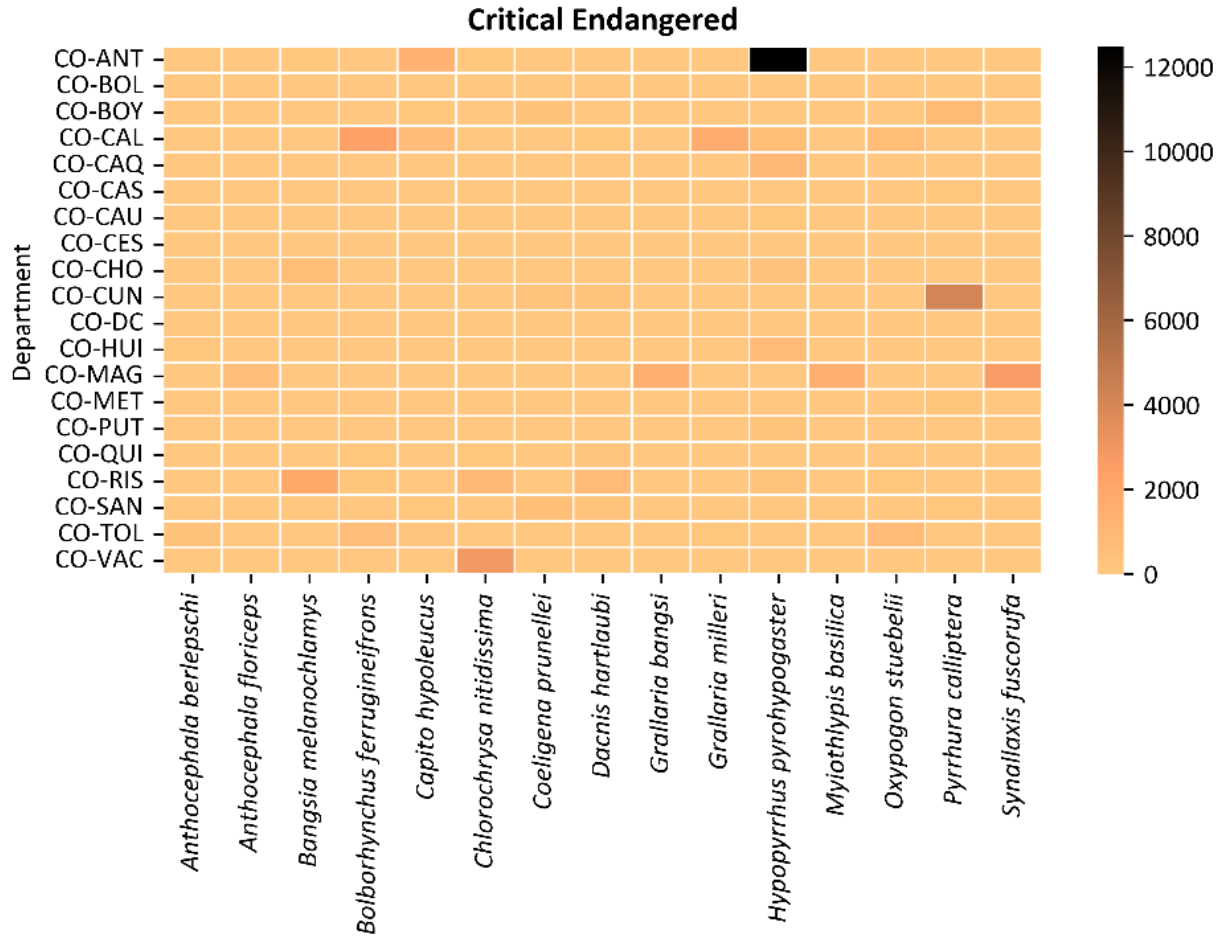


Figura 6. Abundancia de aves endémicas por departamento (2003-2023) en nivel Vulnerable (VU).

Adicionalmente, en la Figura 7 se grafica el número de avistamientos de la especie en el departamento de Antioquia a lo largo de los años que presentaban disponibilidad de datos. La codificación del eje vertical corresponde a la que la plataforma de eBird fija para los departamentos, siendo intuitiva en términos de su interpretación. Esta visualización confirma nuevamente que Antioquia destaca para *Hypopyrrhus pyrohypogaster.* De esta manera, se garantiza que la selección de especies corresponda a aquellas que cuenten con un mayor número de registros por departamento.

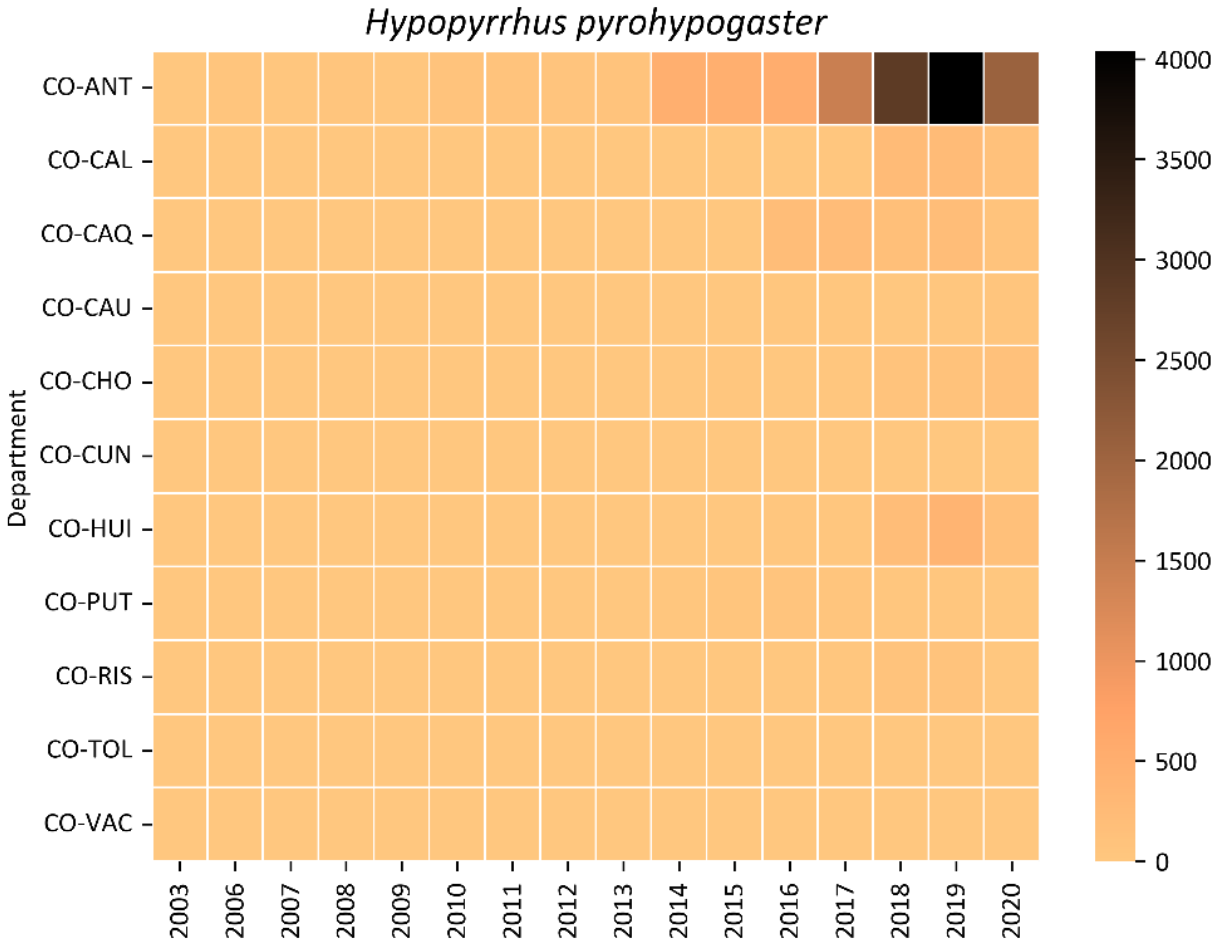


Figura 7. Abundancia de aves endémicas para la *Hypopyrrhus pyrohypogaster.* La representación confirma que Risaralda es el departamento que durante cinco años seguidos presentó la mayor concentración de la especie.

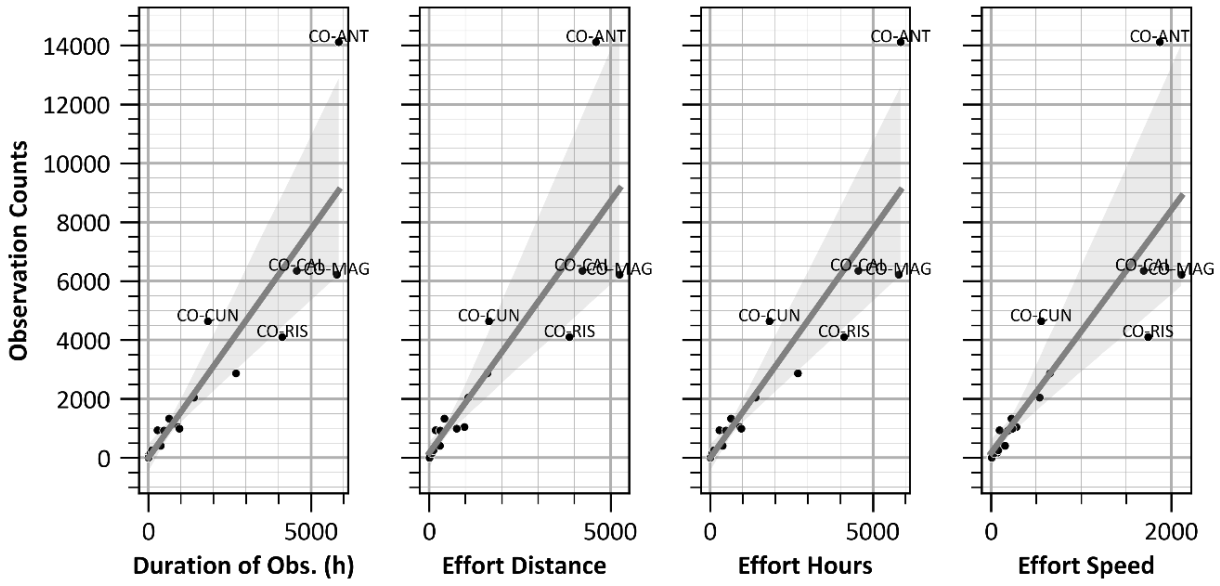
Además, se utilizaron gráficos de dispersión para relacionar estas variables de esfuerzo con la abundancia de especies en el departamento, con el objetivo de detectar tendencias y patrones. Este análisis ayudó a comprender mejor las condiciones bajo las cuales se realizaban las observaciones y su relación con la cantidad de avistamientos reportados.

La figura 8 presenta el gráfico de dispersión para las variables de esfuerzo (*duration of observation*, *effort distance [km]*, *total effort hours* y *total effort speed [km/h]*). Este muestra una tendencia clara en la que un mayor esfuerzo, ya sea en términos de duración de observación, distancia recorrida, horas de esfuerzo o velocidad de esfuerzo, se asocia consistentemente con un mayor número de avistamientos. Esto sugiere que el número de registros está estrechamente relacionado con la intensidad de los esfuerzos de observación, un punto que subraya la

importancia de la continuidad y extensión del trabajo de campo para la obtención de datos precisos sobre estudio de aves. Asimismo, se resalta la necesidad de integrar estas variables asociadas a la tarea de avistamiento al modelado de las distribuciones de las especies, ya que de alguna manera se tiene en cuenta el sesgo que cada avistador puede introducir al registrar un avistamiento.

El departamento de Antioquia sobresale en el gráfico con mayores abundancias de especies, lo que refleja la concentración de esfuerzo en estas zonas. Este patrón podría interpretarse como una indicación de que los esfuerzos de observación en estos departamentos son particularmente exitosos o continuos.

Asimismo, en las líneas de tendencia con intervalos de confianza muestran que la relación entre esfuerzo y abundancia es predecible, con una correlación positiva fuerte en la mayoría de los casos. Esta intensidad en la relación entre variables de esfuerzo y avistamientos puede ser una herramienta que en trabajos futuros permita fortalecer las estrategias de monitoreo de especies en regiones submuestreadas o incluso sin muestreos aparentes que conduzca a conseguir más registros.



.

Figura 8. Abundancia vs variables de esfuerzo. Aves endémicas VU en Colombia. 2003 – 2023. Las variables de esfuerzo corresponden a: Duración de la observación (Duration of Obs. (h)), Distancia de esfuerzo (Effort Distance), Horas de esfuerzo (Effort Hours) y Velocidad de Esfuerzo (Effort Speed). Los valores para cada variable corresponden a las sumas totales por departamento.

De manera complementaria, la Figura 9 presenta la matriz de correlación entre las variables de esfuerzo y la abundancia de la categoría VU, evidenciando una fuerte correlación entre todos

estos atributos. Los valores, comprendidos entre 0.97 y 1.00, confirman una alta correlación positiva, lo que ratifica que existe una relación aparentemente alta entre las variables de esfuerzo y los avistamientos de la especie. Además, también puede indicar que quizás exista multicolinealidad entre dichas variables de esfuerzo, factor que no se debe dejar de lado al momento de entrenar los algoritmos de ML en etapas posteriores [49].

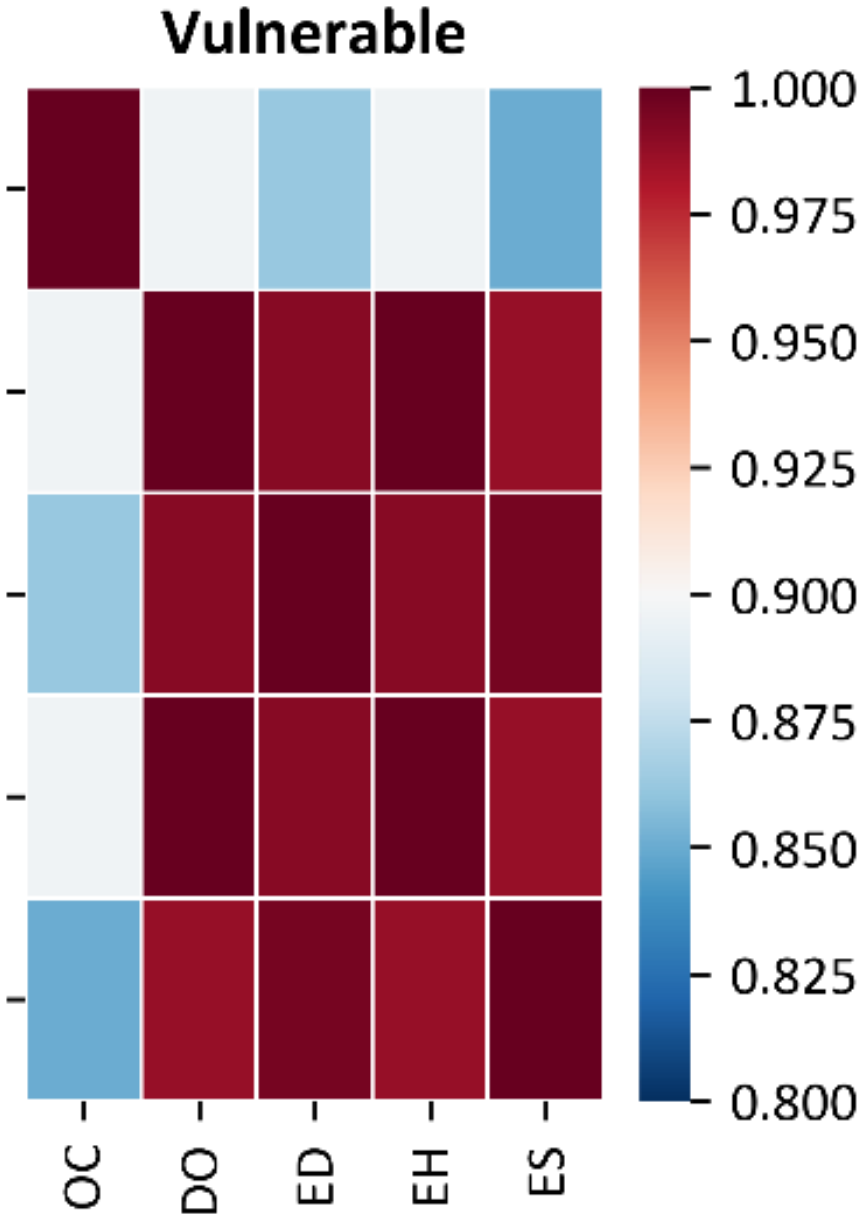


Figura 9. Matriz de correlación entre las variables de esfuerzo y aves endémicas categoría VU de Colombia. OC: Observation Count. DO: Duration Observation [hr]. ED: Effort Distance [km]. EH: Total Effort Hours [hr]. ES: Total Effort Speed [km/h].

Finalmente, se presentan en la figura 10 la imagen del ave con mayor abundancia en la categoría Vulnerable – VU (Hypopyrrhus pyrohypogaster).

Figura 10. Aves con mayor abundancia en Vulnerable – VU (Hypopyrrhus pyrohypogaster). Imagen obtenida de Birds of the World. Cornell Lab of Ornithology, Ithaca, NY, EE.UU.

4.1.4. Sumario

El proceso de preprocesamiento de los datos provenientes de eBird permitió depurar y construir base de datos para analizar los avistamientos de aves endémicas en Colombia. La aplicación de filtros relacionados con duración, distancia, número de observadores y calidad geoespacial garantizó que los registros fueran consistentes y representativos del esfuerzo de observación.

La identificación de sesgos contribuyó a mejorar la representatividad de los datos. La clasificación de las especies endémicas por nivel de amenaza y su agrupación por departamento facilitaron la identificación de regiones con alta concentración de especies de interés, destacándose departamentos como Antioquia.

Las matrices de correlación confirmaron una fuerte asociación positiva entre estas variables, evidenciando la importancia de considerarlas en etapas posteriores de modelado para reducir sesgos y optimizar las predicciones.

4.2. DESARROLLAR UNA METODOLOGÍA DE PREPROCESAMIENTO DE DATOS DE VARIABLES EXÓGENAS (PRECIPITACIONES, TEMPERATURAS, ANTROPOGÉNICAS, ENTRE OTRAS) PARA ESTABLECER CORRELACIONES SOBRE LOS AVISTAMIENTOS DE ESPECIES ENDÉMICAS.

En esta sección se describe el flujo de trabajo diseñado para integrar y correlacionar variables exógenas junto con covariables de esfuerzo de avistamiento, con el fin de evaluar las interacciones entre estas características y la presencia de especies endémicas.

La metodología implementada combina datos provenientes de plataformas de acceso libre, como MapBiomas y WorldClim, procesados en entornos como Google Earth Engine (GEE). Además, se aplican técnicas de análisis geoespacial mediante la generación de mallas hexagonales y el uso de datos vectoriales para delimitar áreas de datos.

Se empleó el cálculo iterativo del Factor de Inflación de Varianza (VIF) para reducir problemas de multicolinealidad entre variables independientes. Todo este procedimiento se replica de forma consistente para las categorías de conservación, asegurando la replicabilidad de la metodología, sin embargo, solo se detallan los resultados de la categoría VU en Antioquia, los demás resultados se encuentran en la sección de Anexos.

4.2.1. Materiales y métodos

Para el análisis de la variable exógena suelo, se emplearon datos provenientes de la plataforma MapBiomas, disponible en Google Earth Engine (GEE) [33]. Esta herramienta proporciona mapas anuales de cobertura y uso del suelo para América Latina, con una resolución mínima de 30 metros x 30 metros, siendo óptima para estudios de ecología y conservación. Esta variable permite visualizar cambios en la cobertura del suelo a lo largo del tiempo, lo que es importante para identificar posibles influencias de las dinámicas del paisaje sobre la distribución de las especies endémicas.

Por su parte, el portal GEE proporciona un repositorio con una guía práctica para el uso de sus datos [34]. Esta guía fue utilizada como base para el procesamiento en esta sección, ajustándola a las necesidades del trabajo y los datos específicos a utilizar. Para integrar esta información con los datos de avistamientos previamente recopilados, el archivo de observaciones se transformó para poder cruzar información entre los registros de las aves y la variable exógena en cuestión.

La plataforma WorldClim [36], proporciona una base de datos global que ofrece estimaciones promedio de variables climáticas derivadas de datos de temperatura y precipitación entre los

años 1960 y 1990. Estas variables bioclimáticas, están disponibles en una resolución espacial de 2.5 arcmin. El término arcmin hace referencia a "minutos de arco", una unidad de medida angular que equivale a 1/60 de un grado. En términos geoespaciales, 2.5 arcmin corresponden aproximadamente a una resolución de 4 a 5 kilómetros por píxel en latitud y longitud, lo que permite un análisis de las condiciones climáticas aproximadas en el área de interés.

Además, emplearon datos vectoriales espaciales (shapefiles) de los departamentos de interés para definir la región de estudio. A partir de esta región, se generó una malla hexagonal que segmentó el área de estudio en celdas regulares definidas de acuerdo con el radio desde el punto medio de los polígonos hasta sus aristas. El tamaño del radio se define de acuerdo con la movilidad esperada de las especies de estudio, por lo que este parámetro no debe ser ajustado. Posteriormente, los registros de observaciones de especies fueron interceptados con las celdas hexagonales, asignando cada registro a la celda correspondiente. Este paso permitió integrar espacialmente las observaciones con las características del área, facilitando el análisis geoespacial.

Finalmente, para tratar el problema de multicolinealidad, se estableció un umbral de VIF igual a 5 para determinar qué variables debían eliminarse, un valor comúnmente aceptado en estudios predictivos. Este paso fue crítico, ya que la reducción de la multicolinealidad no solo mejora la estabilidad de los modelos, sino que también favorece la interpretabilidad de las variables asociadas al fenómeno.

4.2.2. Configuración experimental o descripción de la base de datos

Delimitación del área de interés

En esta sección se tomó como referencia el repositorio de [35], que comprende el tratamiento de los datos de MapBiomas, para poder ser incorporado en trabajos de SDMs.

Para analizar los cambios en la cobertura y uso del suelo en los departamentos con mayor concentración de especies endémicas en cada categoría de conservación, se delimitó el área de interés utilizando un archivo shapefile que contenía las divisiones administrativas a nivel departamental de Colombia. Este archivo fue procesado de manera que permitió manipular las geometrías espaciales y ajustar el sistema de referencia de coordenadas (*Coordinate Reference System*, CRS) a EPSG:4326 (coordenadas de latitud y longitud). La correcta asignación del CRS aseguró la interoperabilidad de los datos geoespaciales provenientes de diversas fuentes de información. Una vez definida la región de interés, se procedió a construir un polígono regular que cubriera dicha región (Bounding Box), que abarcaba los límites espaciales mínimos y máximos de los departamentos en estudio.

Variables exógenas: Cobertura y Uso de Suelo de MapBiomas

Del conjunto de datos MapBiomas, disponible en la plataforma GEE, se descargaron imágenes correspondientes a los años entre 2014 y 2020, específicamente para los departamentos de interés con mayor concentración de especies endémicas en cada categoría de conservación elegida. Estas imágenes reflejaban diferentes clases de cobertura y uso del suelo, como bosques, agricultura, áreas urbanas y cuerpos de agua, entre otras. Para garantizar la consistencia y facilitar el análisis, cada imagen fue recortada utilizando los límites espaciales de la bounding box definida previamente para cada área de estudio.

Una vez recortadas las imágenes se exportaron en formato TIFF, codificación convencional en el contexto de imágenes geoespaciales por su capacidad para almacenar múltiples capas de información, preservando una alta calidad. Esta exportación permitió optimizar el flujo de trabajo, de manera que se pudiesen procesar las imágenes fuera del entorno de GEE. Cada imagen contenía varias capas de información sobre el uso del suelo, lo que permitió analizar las variaciones en las clases de cobertura dentro del área de estudio.

Variables Exógenas: factores climáticos y de topografía

Se descargaron y procesaron variables climáticas y topográficas utilizando los datos de la plataforma WorldClim, en total, se descargaron 19 variables, cada una de ellas diseñada para describir distintos aspectos climatológicos que pueden influir en los patrones de distribución de las especies (Ver Tabla 3). Estas variables ofrecen una descripción integral del entorno climático, abarcando factores como temperatura, precipitación y estacionalidad, todos ellos de gran relevancia para entender las condiciones ambientales [4].

Posteriormente, las variables bioclimáticas y elevación fueron recortadas utilizando los límites geográficos de Antioquia.

Variable	Descripción
Bio1	Temperatura media anual.
Bio2	Rango de temperatura diurno medio.
Bio3	Isotermalidad.
Bio4	Estacionalidad de la temperatura (desviación estándar).
Bio5	Temperatura máxima del mes más cálido.
Bio6	Temperatura mínima del mes más frío.
Bio7	Rango de temperatura anual.
Bio8	Temperatura media del trimestre más húmedo.
Bio9	Temperatura media del trimestre más seco.
Bio10	Temperatura media del trimestre más cálido.
Bio11	Temperatura media del trimestre más frío.
Bio12	Precipitación anual.
Bio13	Precipitación del mes más húmedo.
Bio14	Precipitación del mes más seco.
Bio15	Estacionalidad de la precipitación.
Bio16	Precipitación del trimestre más húmedo.

Bio17	Precipitación del trimestre más seco.
Bio18	Precipitación del trimestre más cálido.
Bio19	Precipitación del trimestre más frío

Tabla 3. Listado de variables bioclimáticas de la colección de WorldClim

Preparación y estructuración de datos

A partir de esta región de estudió definida a través de los shapefiles, se generó una malla hexagonal que segmentó el área de estudio en celdas regulares definidas de acuerdo con el radio desde el punto medio de los polígonos hasta sus aristas. El tamaño del radio se define de acuerdo con la movilidad esperada de las especies de estudio, por lo que este parámetro no debe ser ajustado. Posteriormente, los registros de observaciones de especies fueron interceptados con las celdas hexagonales, asignando cada registro a la celda correspondiente. Este paso permitió integrar espacialmente las observaciones con las características del área, facilitando el análisis geoespacial.

Concatenación de variables exógenas y covariables de esfuerzo de avistamiento

Una vez se prepararon espacialmente las regiones de estudio (Risaralda y Antioquia), se acumularon las variables de esfuerzo (velocidad, horas de observación, distancia recorrida y número de observadores) para cada hexágono. Se estimó la abundancia de observaciones en cada celda y se asignaron etiquetas binarias para indicar la presencia o ausencia de especies, proporcionando un marco de referencia para los modelos de predicción.

Posteriormente, se integraron las variables exógenas como elevación y las características bioclimáticas de WorldClim, así como los datos de cobertura y uso del suelo de MapBiomas. Este proceso implicó alinear espacialmente los datos raster con la malla hexagonal, imputando las características climáticas, topográficas y de uso del suelo en cada celda. Para ello, se emplearon funciones personalizadas que asignaron los valores de los píxeles raster a los centroides de las celdas hexagonales, garantizando la representación de las condiciones ambientales dentro de cada unidad espacial. Finalmente se combinaron las variables imputadas provenientes de las fuentes de WorldClim y MapBiomas en una única representación de datos. Este facilitó el cálculo de la posterior correlación entre estas variables y la presencia de especies endémicas.

El conjunto de datos resultante, que contenía todas las variables integradas y las variables de esfuerzo fue exportado con el objetivo de conservar una versión completa para análisis posteriores.

Selección no supervisada de variables

El proceso de cálculo del VIF se implementó de manera iterativa. En cada iteración, se identificó la variable con el valor de VIF más alto y se eliminó del conjunto de datos. Luego, se recalculó el VIF para las variables restantes. Este procedimiento se repitió hasta que todas las variables

seleccionadas presentaron valores de VIF dentro del umbral aceptado. Este enfoque garantizó que las variables finales seleccionadas fueran informativamente únicas y relevantes para los objetivos de los modelos. Las variables seleccionadas después de filtrar por VIF se encuentran en Tabla 4.

En el proceso de depuración de variables se contaba con un total de 109 variables predictoras iniciales, tras aplicar el cálculo del Factor de Inflación de Varianza (VIF), se eliminaron 74 variables que presentaban multicolinealidad en el departamento de Antioquía, quedando 35 variables finales que cumplen con los umbrales de independencia aceptables.

Departamento Antioquia		
N°	**Variable**	**VIF**
1	effort_speed_kmph	4.250840
2	number_observers	4.068717
3	elevation	2.445737
4	p_driest_m	3.682484
5	lulc_2014_formacion_natural_no_forestal_inundable	3.032914
6	lulc_2014_manglar	2.866497
7	lulc_2014_mineria	2.046198
8	lulc_2014_mosaico_de_agricultura_pasto	2.453833
9	lulc_2014_no_observado	2.515635
10	lulc_2014_otra_area_sin_vegetacion	1.002053
11	lulc_2014_otra_formacion_natural_no_forestal	2.394529
12	lulc_2014_playas_dunas_bancos_de_arena	2.780682
13	lulc_2014_silvicultura	2.976640
14	lulc_2015_mosaico_de_agricultura_pasto	3.263584
15	lulc_2016_otra_formacion_natural_no_forestal	2.534086
16	lulc_2017_bosque_inundable	1.096440
17	lulc_2017_playas_dunas_bancos_de_arena	1.727587
18	lulc_2017_silvicultura	2.903880
19	lulc_2018_infraestructura_urbana	4.397281
20	lulc_2018_otra_formacion_natural_no_forestal	2.767572
21	lulc_2018_silvicultura	4.894443
22	lulc_2019_mosaico_de_agricultura_pasto	2.273339
23	lulc_2019_no_observado	2.023573
24	lulc_2019_otra_formacion_natural_no_forestal	4.448505
25	lulc_2019_palma_aceitera	2.957906
26	lulc_2019_silvicultura	3.324598
27	lulc_2020_bosque	1.818514
28	lulc_2020_formacion_natural_no_forestal_inundable	2.517011
29	lulc_2020_infraestructura_urbana	1.909788
30	lulc_2020_manglar	2.799804
31	lulc_2020_mineria	3.182958
32	lulc_2020_no_observado	1.121591
33	lulc_2020_otra_area_sin_vegetacion	1.001579
34	lulc_2020_palma_aceitera	2.168568
35	lulc_2020_silvicultura	2.356434

Tabla 4. Variables seleccionadas para matriz de correlación de variables exógenas posterior a diputación VIF.

4.2.3. Resultados y discusión

Se generaron mapas interactivos dentro del entorno de GEE, lo que permitió una visualización dinámica de las transformaciones en la cobertura del suelo. Estos mapas ofrecieron una representación gráfica de los cambios en el uso del suelo a lo largo del período de interés, proporcionando una herramienta visual efectiva para comprender las transformaciones paisajísticas en el departamento de Antioquía.

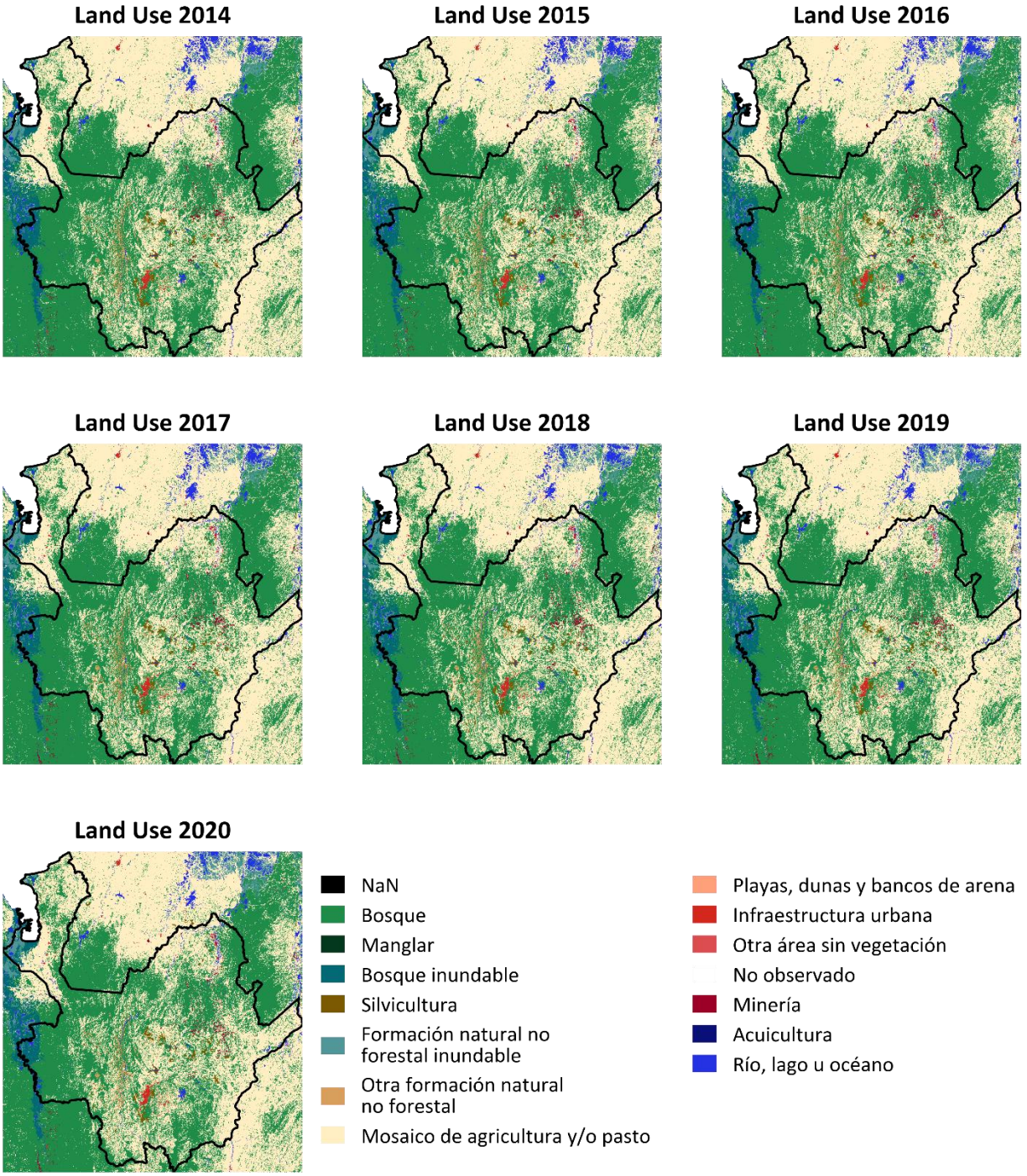


Figura 11. Variación de uso del suelo 2014 - 2020 en el departamento de Antioquia, procesamiento de base de datos de MapBiomas de la plataforma Google Earth Engine.

En Antioquia, el análisis del uso del suelo entre 2014 y 2020, ilustrado en la Figura 11, reveló que las áreas boscosas se mantuvieron constantes, mientras que las actividades agrícolas y de pastoreo se expandieron ligeramente hacia el sur. El crecimiento de la infraestructura urbana fue

gradual, especialmente en las zonas centrales del departamento, indicando un proceso de urbanización en curso. Las áreas dedicadas a la silvicultura también aumentaron, reflejando una mayor actividad en el manejo forestal. En conjunto, estos cambios evidenciaron una evolución moderada del paisaje, con un equilibrio entre la estabilidad de las coberturas naturales y la expansión controlada de las áreas agrícolas, urbanas y forestales.

Los cambios en el uso del suelo permiten identificar patrones que influyen en la presencia de especies endémicas en zonas específicas, como las áreas boscosas estables o las áreas urbanas en expansión. Esto no solo proporciona información para mejorar los algoritmos de predicción, sino que también ayuda a priorizar áreas de conservación, así como a intuir posibles incidencias del factor humano sobre la modificación de los nichos ecológicos de las especies. Estos resultados refuerzan la importancia de integrar variables exógenas en los modelos predictivos. Al combinar estos análisis con modelos de ML, se pueden generar predicciones más precisas sobre las áreas que son propensas a albergar especies en peligro.

Posteriormente, se graficó en la Figura 12 la representación espacial de la primera variable bioclimática del conjunto de datos de variables exógenas descargado, Bio1: Temperatura media anual, en el departamento de Antioquia. Las imágenes permiten observar cómo la temperatura varía en diferentes zonas. Los colores utilizados en la escala cromática indican un gradiente térmico, donde los tonos cálidos representan áreas con temperaturas promedio más altas, mientras que los tonos fríos reflejan zonas con temperaturas más bajas. Se debe tener en cuenta que este tipo de variables son usualmente incorporadas en SDMs, ya que capturan la tendencia de la variable a la que probablemente las especies de estudio se hayan adaptado.

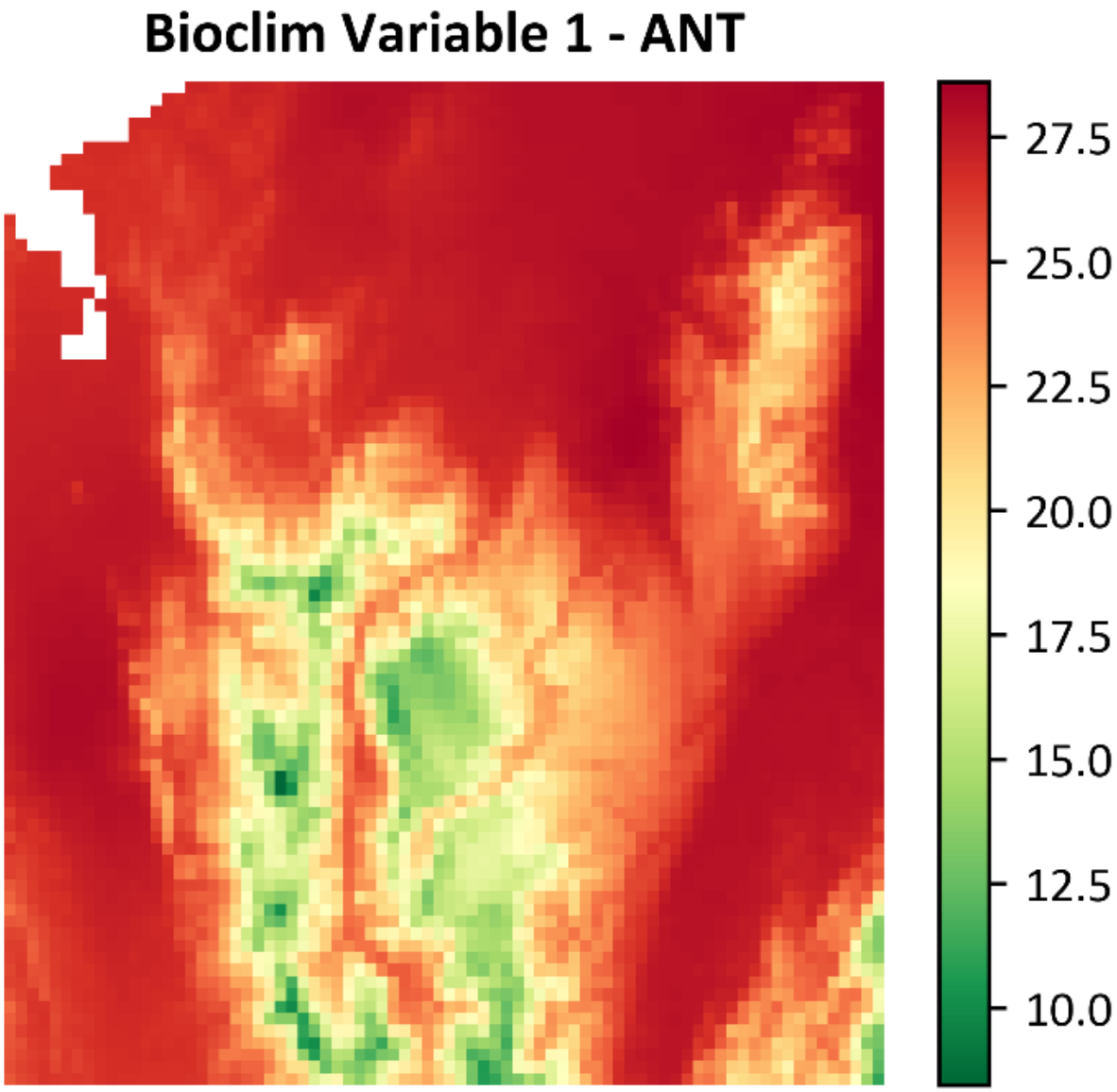


Figura 12. Representación espacial de la variable Bioclimática Bio1: Temperatura media anual para el departamento de Antioquia.

Por su parte, la Figura 13 representa la distribución de la variable de elevación en Antioquia, donde también se observan variaciones altitudinales significativas. Las zonas más bajas, en tonos verdes, se encuentran en áreas periféricas, mientras que las regiones más elevadas, en tonos rojos, se concentran en el centro y suroeste del departamento, alcanzando más de 3000 metros. Esta distribución revela un contraste marcado en las características topográficas de Antioquia, lo que puede influir directamente en la presencia de especies endémicas que requieren hábitats específicos según su altitud.

La topografía desempeña un papel fundamental en la disponibilidad de recursos y en las

condiciones necesarias para la supervivencia de las aves endémicas, ya que crea microclimas y delimita zonas ecológicas. Comprender la distribución altitudinal es esencial para identificar áreas prioritarias de conservación, especialmente en regiones con variaciones significativas en la elevación, donde las especies suelen estar adaptadas a nichos específicos. Por esta razón, la inclusión de esta variable es importante en la correlación con los datos de aves endémicas [38].

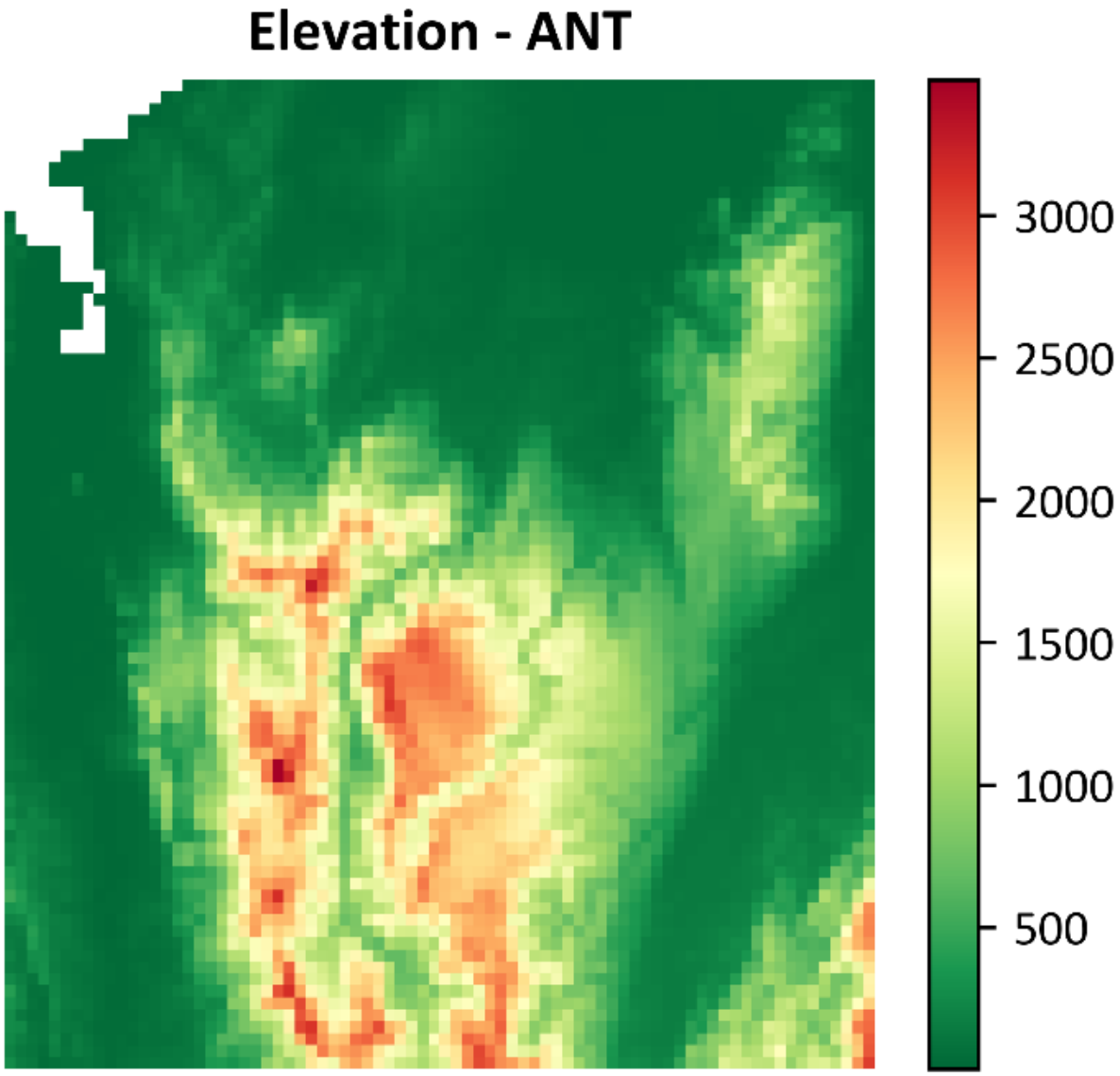


Figura 13. Representación espacial de la elevación para ambos departamentos. La resolución de esta variable condiciona el detalle con el que se pueda apreciar la tendencia de la variable en el departamento de Antioquia.

Cuando se obtuvieron las variables integradas y las variables de esfuerzo en una base de datos completa para análisis se realizó la distribución espacial de los registros de avistamiento del ave

Hypopyrrhus Pyrohypogaster en Antioquia observando una amplia dispersión de los registros a lo largo del departamento. En la Figura 14 se representan estos registros en una malla hexagonal superpuesta sobre el área de interés. Al estructurar los datos de este tipo de representación espacial, el análisis geográfico se facilita al agrupar los registros en celdas regulares, lo que permite identificar patrones de distribución más claros y homogéneos, esenciales para correlacionar estas observaciones con variables exógenas y priorizar áreas de conservación [39].

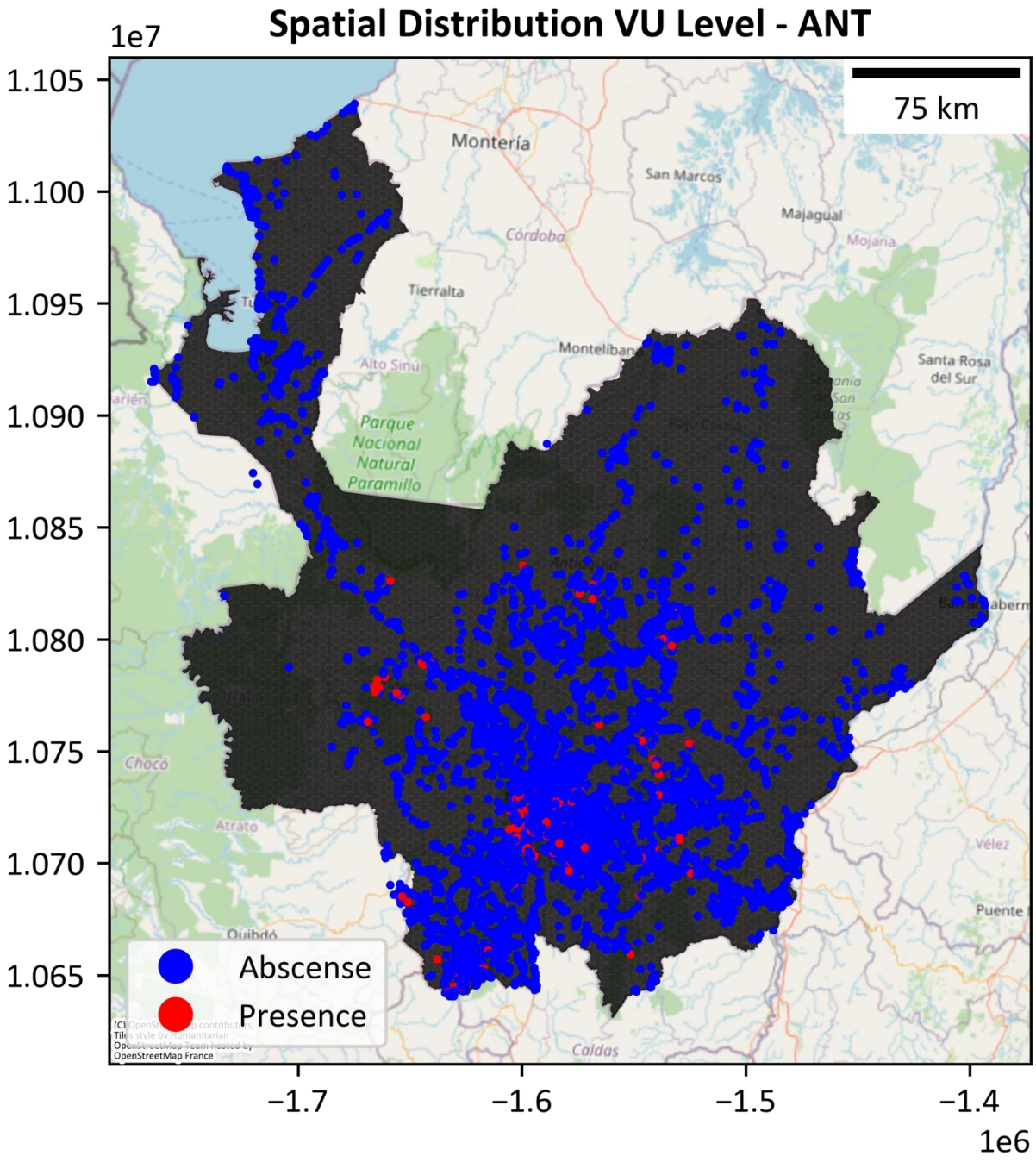


Figura 14. Distribución espacial y representación hexagonal de registros de *Hypopyrrhus Pyrohypogaster* en Antioquia. Este ejemplar presenta mayor cantidad de avistamientos. Sin embargo, esto también se puede deber a la extensión de territorio que comprende el departamento.

Tras la depuración de variables basada en el VIF, se generó una matriz de correlación final, utilizando únicamente las variables seleccionadas. En la Figura 15 (Derecha) se pueden ver dichos resultados. Al comparar estas matrices con las iniciales, se observó una reducción significativa de las correlaciones fuertes, lo que podría indicar que el problema de multicolinealidad quizás fue mitigado. Sin embargo, para evaluar el efecto de lo anterior, se desarrollaron pruebas de clasificación que se verán posteriormente. Las variables restantes mostraron relaciones más claras y diferenciadas lo que facilitó la interpretación. Este enfoque redujo la complejidad inicial y optimizó la preparación de los datos asegurando que los modelos predictivos desarrollados posteriormente se basaran en variables significativas y no redundantes.

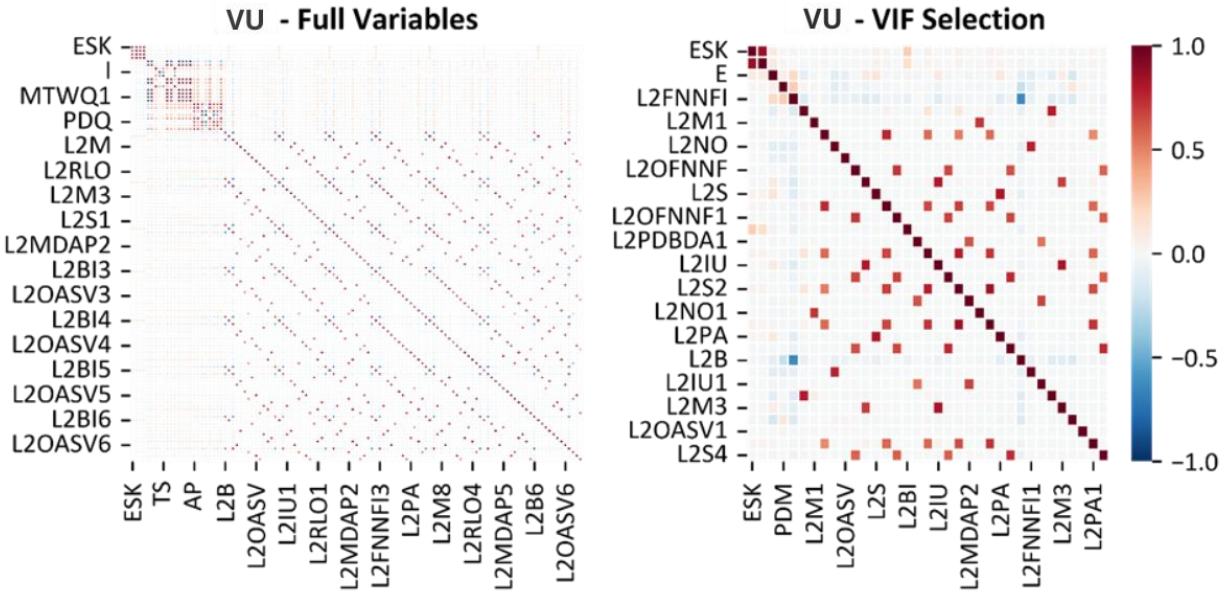


Figura 15. Matrices de correlación para el conjunto de características exógenas completas (Full Variables - Izquierda) y el conjunto de datos una vez se aplicó el VIF (Derecha) para el Departamento de Antioquia.

4.2.4. Sumario

La integración de variables exógenas (cobertura, uso del suelo, factores climáticos y topográficos) y covariables de esfuerzo permitió consolidar una base de datos geoespacial alineada con la región de estudio. El uso de herramientas como MapBiomas, WorldClim y GEE facilitó la obtención y procesamiento de información. La generación de la malla hexagonal y la aplicación del VIF ayudaron a depurar variables redundantes, optimizando la calidad del conjunto de datos para análisis posteriores. Los resultados reflejan patrones de distribución que respaldan la importancia

de considerar estas variables para mejorar los modelos predictivos de avistamientos de especies endémicas.

4.3. IMPLEMENTACIÓN Y VALIDACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO PARA PREDICCIÓN DE AVISTAMIENTOS DE ESPECIES ENDÉMICAS

En esta sección se aborda el tercer y cuarto objetivo de esta investigación: implementar algoritmos de aprendizaje automático para predecir avistamientos de especies endémicas a partir de variables exógenas; e implementar estrategias validación de resultados a partir de medidas de desempeño y técnicas de visualización de información para revelar comportamientos espaciales y temporales de especies endémicas. Para ello se consideraron diversos modelos utilizados regularmente en el contexto de la predicción de distribuciones de especies, destacando sus características generales y las razones para la selección específica de los dos modelos empleados en este trabajo.

Además de las métricas de desempeño, se generaron mapas de probabilidad que permitieron visualizar la distribución espacial de las predicciones realizadas por los modelos de RL y RF. Estos mapas muestran las estimaciones de probabilidad de presencia de la especie en cada celda de la malla hexagonal, proporcionando una perspectiva geoespacial de los resultados y evidenciando el impacto de las diferentes configuraciones evaluadas.

4.3.1. Materiales y métodos

Se seleccionaron la Regresión Logística y Random Forest como los modelos principales debido a sus ventajas específicas en términos de interpretabilidad y aplicabilidad en el contexto de los modelos de distribución de especies. La Regresión Logística fue elegida por su capacidad para proporcionar una interpretación clara de las relaciones entre variables exógenas y la probabilidad de avistamientos, así como por su simplicidad matemática, lo que facilita la comprensión de los resultados. Por otro lado, Random Forest se seleccionó debido a su capacidad para abstraer reglas binarias para realizar predicciones. Asimismo, este modelo permite identificar las variables más relevantes y su adopción en plataformas reconocidas como eBird, donde ha demostrado ser una herramienta robusta y confiable para generar mapas de distribución de especies.

En contraste, se descartaron los modelos de Máxima Entropía (MaxEnt) y basados en Redes Neuronales. En el caso de MaxEnt, si bien es ampliamente utilizado, se basa en fundamentos matemáticos equivalentes al modelo de Regresión Logística, lo que hace redundante su inclusión. Por otro lado, se descartaron los modelos basados en Redes Neuronales, ya que se quisieron priorizar modelos cuya capacidad de interpretación fuese alta, ya que se pretende que esta metodología permita comprender factores subyacentes que afecten los avistamientos.

Como paso complementario al entrenamiento de los modelos Random Forest, se realizó un análisis de importancia de características con el objetivo de identificar las variables que más contribuyeron a la predicción de avistamientos para cada categoría de conservación. Para ello, se utilizó la métrica MDI (Mean Decrease in Impurity), estimando también la desviación estándar asociada [50]. Este análisis permitió contrastar los resultados obtenidos con la etapa previa de preprocesamiento, en la cual se aplicó un análisis de colinealidad mediante el cálculo del Factor de Inflación de Varianza (VIF), con el propósito de preservar únicamente aquellas variables que aportaran información no redundante al modelo.

$$MDI(j) = \sum_{T:v(t)=j} p(t) \cdot \Delta i(t)$$

Donde T representa el conjunto de nodos internos del árbol, v(t) es la variable utilizada para la división en el nodo t, p(t) es la proporción de muestras que llegan al nodo t, y Δi(t) es la disminución de impureza producida por la división.

4.3.2. Configuración experimental o descripción de la base de datos

Resumen del conjunto de datos

En la tabla 5 se presentan la cantidad de muestras para las categorías de conservación Peligro Crítico (CR), En Peligro (EN) y Vulnerable (VU). Adicionalmente se observan el número de variables y el número de presencias y ausencias, evidenciando el desbalance en las clases.

Categoría	Cantidad de muestras	Número de variables	Presencias	Ausencias
CR	6485	90	21	6464
EN	6485	90	65	6420
VU	7122	122	196	6926

Tabla 5. Resumen del conjunto de datos por categoría.

En los siguientes numerales se detalla el flujo de trabajo desarrollado para entrenar y validar los modelos de aprendizaje automático seleccionados: Regresión Logística (RL) y Random Forest (RF). Ambos modelos fueron implementados con el objetivo de predecir la probabilidad de presencia de especies endémicas en función de variables exógenas y covariables de esfuerzo.

Diseño de pipelines y configuración de Hiperparámetros

El primer paso para implementar los modelos consistió en diseñar pipelines que estructuraran el flujo de trabajo por etapas. Estos pipelines permitieron integrar las tareas de preprocesamiento, selección de características y ajuste de hiperparámetros de forma sistemática.

En el caso de la RL, se configuraron los parámetros más relevantes para optimizar el desempeño del modelo. Entre estos, se incluyeron:

- *Regularización:* se exploraron las técnicas L1 (Lasso) y L2 (Ridge) para controlar el sobreajuste, penalizando los coeficientes asociados a las variables.
- *Constante de regularización (C):* se probaron diferentes valores (1.0, 0.5 y 0.1), que definieron la intensidad de la penalización.
- *Solver:* se utilizó liblinear, eficiente para conjuntos de datos de tamaño moderado.

Por su parte, RF fue configurado para garantizar su capacidad de modelar relaciones complejas en los datos. Se ajustaron los siguientes parámetros:

- *Profundidad máxima de los árboles*, controlando la complejidad del modelo.
- *Número mínimo de muestras en las hojas terminales,* regulando la granularidad de las predicciones.
- *Número mínimo de muestras necesarias para dividir un nodo,* que definió las condiciones para seguir ramificando los datos.

El ajuste de los hiperparámetros de ambos modelos se llevó a cabo mediante una búsqueda aleatoria (RandomizedSearchCV), que permitió evaluar múltiples combinaciones de parámetros optimizando la métrica de exactitud.

Estrategias de validación, particionamiento de datos en conjuntos de prueba y entrenamiento

El particionamiento de datos y la validación cruzada garantizaron la evaluación objetiva y la capacidad de generalización de los modelos. En primer lugar, los datos se dividieron en dos subconjuntos principales: entrenamiento (80%) y prueba (20%). Debido a la naturaleza desbalanceada de los datos, donde la clase de ausencia era considerablemente más frecuente que la de presencia, este particionamiento se realizó utilizando un enfoque estratificado.

El desbalance en los datos presentaba un riesgo importante: sin un manejo adecuado, los modelos podrían inclinarse hacia la clase mayoritaria comprometiendo su capacidad para identificar correctamente las observaciones de la clase minoritaria. El enfoque estratificado permitió mantener la proporción de clases (presencia y ausencia) constante en ambos subconjuntos, asegurando que tanto el conjunto de entrenamiento como el de prueba fueran representativos de la distribución real de los datos [49].

El conjunto de entrenamiento fue utilizado para ajustar los parámetros de los modelos, mientras que el conjunto de prueba quedó reservado para evaluar el desempeño final de los modelos, simulando condiciones reales de uso.

Posteriormente, solo para el conjunto de entrenamiento se implementó una validación cruzada con tres particiones. Este método consistió en dividir los datos de entrenamiento en tres subconjuntos adicionales: en cada iteración, uno de ellos fue reservado para validación, mientras que los otros dos fueron utilizados para entrenar el modelo. Este enfoque permitió evaluar el desempeño del modelo durante el ajuste de parámetros y garantizar que el conjunto de prueba quedara completamente independiente, permitiendo medir la capacidad de generalización del modelo una vez finalizado el entrenamiento.

Manejo de desbalance de clases, implementación de técnicas de sobre muestreo

Se aplicaron las técnicas de SMOTE y Random OverSampler por separado al conjunto de entrenamiento, generando dos nuevos conjuntos de datos balanceados. Estos conjuntos fueron posteriormente utilizados para entrenar nuevamente los modelos de RL y RF. Este enfoque no solo permitió evaluar el impacto de las diferentes estrategias de balanceo sobre el desempeño de los modelos, sino que también permitió identificar cuál de estas técnicas era más adecuada para el problema específico de predicción de presencia de especies.

Selección de variable relevantes

Después de aplicar las medidas para abordar el desbalance de clases, se evaluó cómo la selección de variables a través del VIF impactó en los resultados de predicción de los modelos. Con este propósito, se ingresaron las variables seleccionadas a los modelos de ML, con el fin de evaluar el impacto de la selección de atributos sobre el desempeño de los algoritmos. El procedimiento incluyo la transformación y codificación de datos categóricos, así como la evaluación de las variables identificadas [40]. Entre las variables seleccionadas se incluyeron aspectos como el uso y cobertura del suelo, variables climáticas y covariables de esfuerzo, representando los factores más significativos para la predicción de avistamientos. En el caso del departamento Antioquia se identificaron 35, como se detalla en la Tabla 4.

Resultados de evaluación de los modelos Regresión Logística y Random Forest

El proceso de evaluación de los modelos se estructuró en un flujo de tratamiento de datos para probar la capacidad de la generalización de los modelos de RL y RF bajo diferentes configuraciones y técnicas. Este procedimiento se aplicó a las categorías de conservación Peligro Crítico (CR), En Peligro (EN) y Vulnerable (VU). Los escenarios evaluados incluyeron las siguientes configuraciones:

I. **Evaluación inicial sin balanceo:** los modelos fueron evaluados con todas las variables disponibles, sin aplicar técnicas de balanceo de clases, para obtener una línea base del desempeño.
II. **Aplicación de técnicas de balanceo:** se implementaron dos técnicas de balanceo de clases,

Random OverSampler y SMOTE, para mitigar el desbalance y mejorar la capacidad de los modelos para identificar la clase minoritaria.

III. **Evaluación posterior al balanceo:** los modelos fueron reevaluados después de aplicar las técnicas de balanceo, manteniendo todas las variables disponibles.
IV. **Selección de variables representativas:** se redujo el conjunto de predictores a través de un proceso de selección de variables, con el objetivo de eliminar redundancias y optimizar el modelo.
V. **Evaluación con variables representativas sin balanceo:** los modelos fueron evaluados utilizando únicamente las variables seleccionadas, sin aplicar balanceo de clases.
VI. **Aplicación de balanceo a datos con variables representativas:** las técnicas de balanceo Random OverSampler y SMOTE se aplicaron al conjunto reducido de variables representativas.
VII. **Evaluación final con balanceo y variables representativas:** se realizó una evaluación final de los modelos después de aplicar balanceo al conjunto reducido de variables, analizando su desempeño bajo esta configuración optimizada.

En la sección de resultados y discusión se detalla el análisis y los resultados obtenidos en cada uno de estos escenarios, destacando cómo las configuraciones y técnicas implementadas impactaron el desempeño de los modelos.

4.3.3. Resultados y discusión

En la Tabla 6 se presenta un resumen detallado del desempeño de los modelos de Regresión Logística y Random Forest en la predicción de avistamientos de aves para la categoría de conservación Vulnerable (VU). Las métricas evaluadas incluyen precisión (precision), sensibilidad (recall), puntaje F1 (f1-score) y exactitud (accuracy), las cuales permiten analizar la capacidad de los modelos para identificar correctamente las observaciones de la clase minoritaria (presencia) y su equilibrio con la clase mayoritaria (ausencia). Los resultados están organizados según las distintas configuraciones evaluadas, incluyendo la aplicación de balanceo, el uso de todas las variables y la selección de variables representativas.

En las figuras 16 a 21 se presentan las matrices de confusión en los que se basa los cálculos de las métricas de evaluación de desempeño.

Modelo	precision	recall	f1-score	accuracy
Regresión Logistica (variables completas)	98%	98%	97%	98%
Random Forest (variables completas)	97%	98%	97%	98%
Regresión Logistica + Random oversampling (variables completas)	97%	86%	90%	86%
Random Forest + Random oversampling (variables completas)	98%	93%	95%	93%
Regresión Logistica + SMOTE (variables completas)	97%	88%	92%	88%
Random Forest + SMOTE (variables completas)	97%	95%	96%	95%
Regresión Logistica (solo variables representativas)	98%	97%	96%	97%
Random Forest (solo variables representativas)	97%	98%	97%	98%

Regresión Logistica + Random oversampling (solo variables representativas)	96%	82%	88%	82%
Random Forest + Random oversampling (solo variables representativas)	98%	92%	94%	92%
Regresión Logistica + SMOTE (solo variables representativas)	96%	83%	88%	83%
Random Forest + SMOTE (solo variables representativas)	97%	94%	95%	94%

Tabla 6. Resultados de evaluación de los modelos Regresión Logística y Random Forest para la predicción de avistamiento de aves para categoría de peligro VU

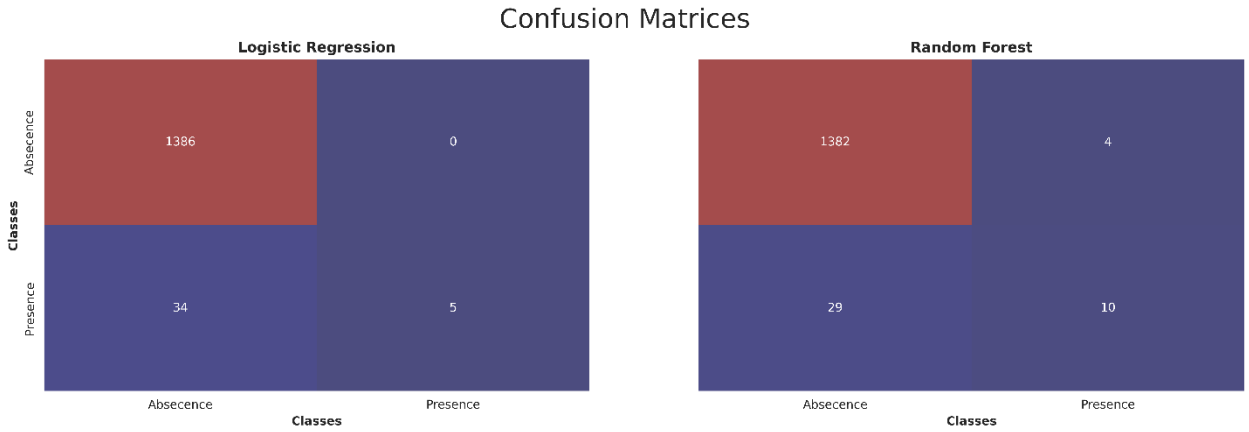


Figura 16. Matriz de confusión evaluación inicial de modelos regresión Logistica y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría VU.

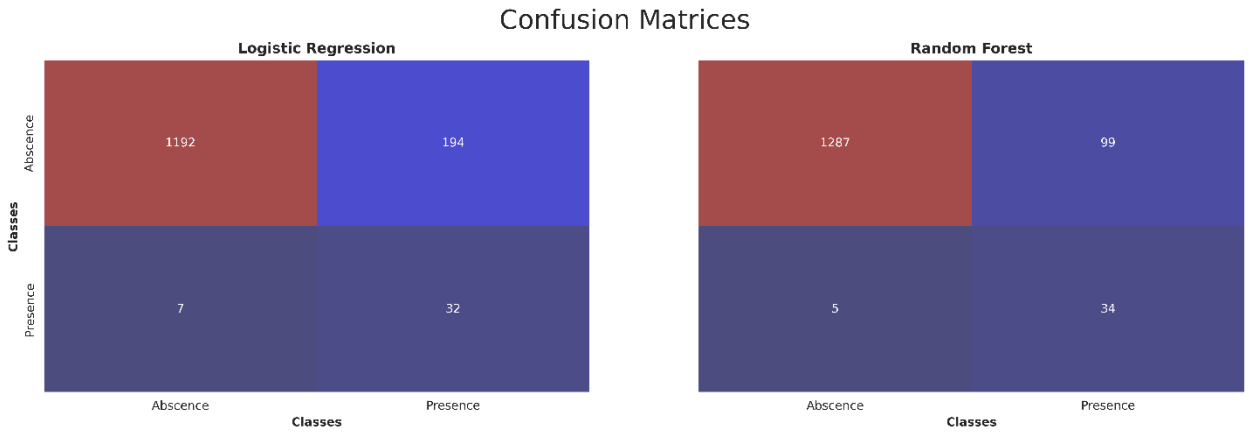


Figura 17. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica Random over sampling con todas las variables disponibles. Categoría VU.

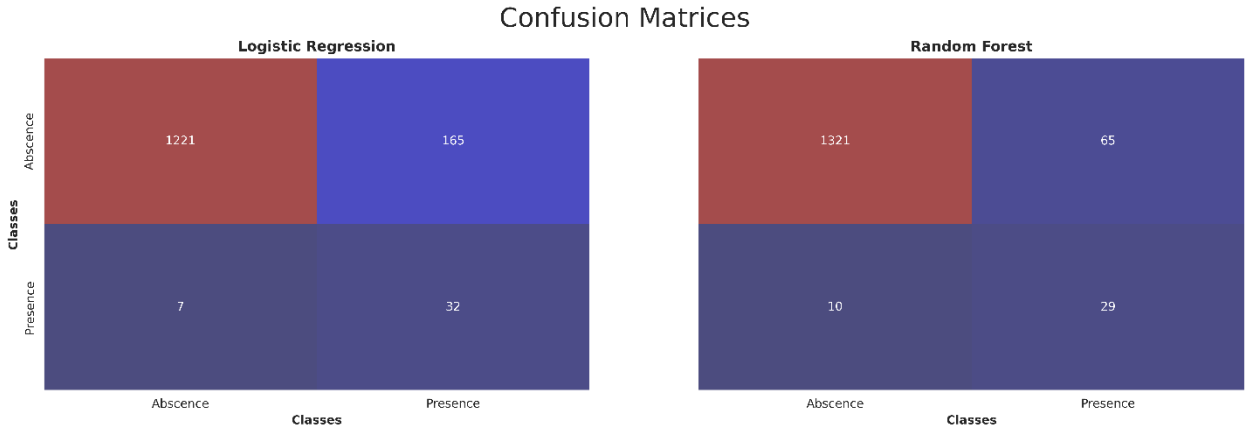


Figura 18. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría VU.

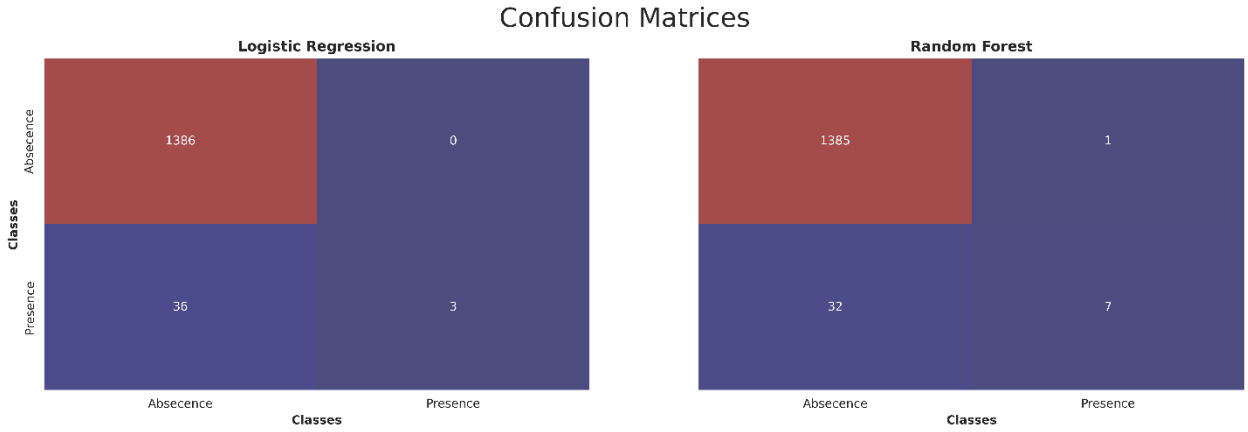


Figura 19. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría VU.

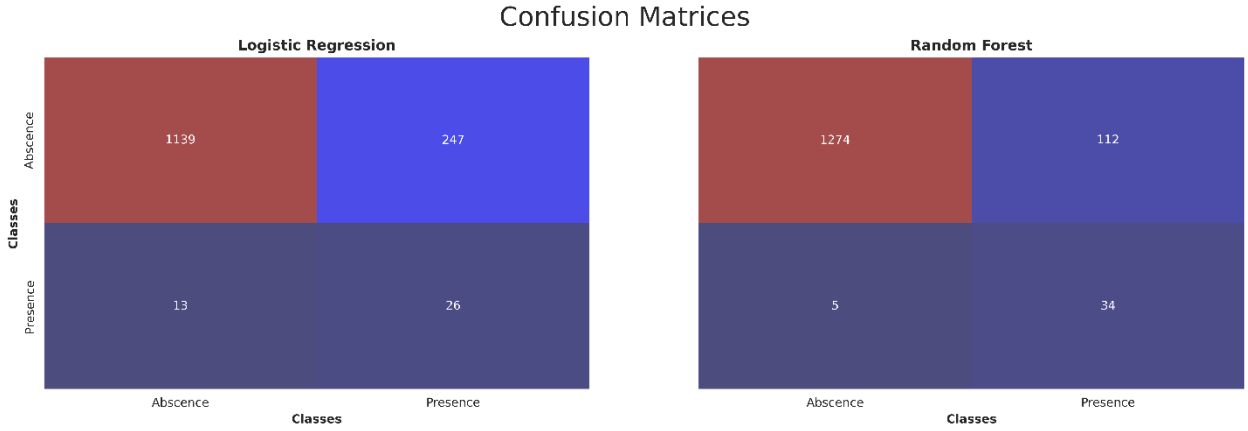


Figura 20. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica Random over sampling con selección de variables representativas. Categoría VU.

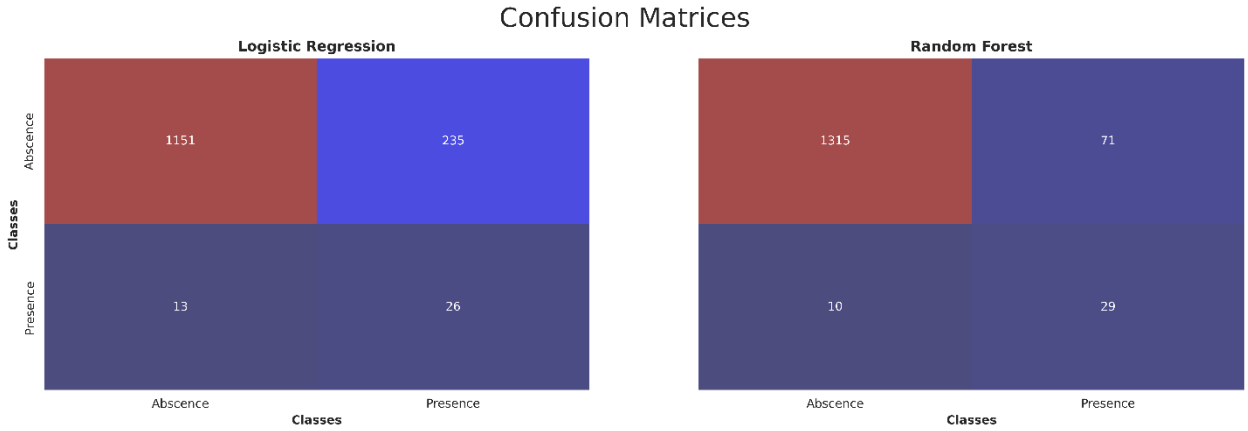


Figura 21. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría VU.

Análisis de resultados de implementación de modelos. Categoría Vulnerable (VU):

Evaluación inicial sin balanceo:

- Ambos modelos tienen dificultades para identificar correctamente la clase minoritaria (Presencia), en el caso de RL identifico menos presencias correctamente.
- RF tiene una ligera ventaja en precisión, aunque los resultados siguen siendo insuficientes para la clase minoritaria.

Aplicación de técnicas de balanceo:

- Al aplicar Random Oversampler y SMOTE, ambos modelos mejoraron significativamente su capacidad para clasificar la clase minoritaria.
- Las matrices de confusión muestran un incremento en las predicciones correctas para la clase minoritaria en ambos modelos, en especial para RL.

Evaluación con variables representativas sin balanceo:

- La selección de variables representativas redujo la complejidad de los modelos, sin embargo, al no tener balanceo de las clases no presentaron una buena clasificación de la clase minoritaria.

Aplicación de balanceo con variables representativas:

- Los dos modelos mejoraron después de aplicar técnicas de balanceo al conjunto reducido de variables.
- RF presentó un mejor desempeño, lo que sugiere que es más adecuado para manejar datasets balanceados con variables representativas.
- Las matrices de confusión muestran una mejor distribución de las predicciones correctas, con una reducción significativa de falsos negativos en comparación con la RL.

Posteriormente se elaboraron los mapas de probabilidad evaluación de los modelos Regresión

Logística y Random Forest para categoría Vulnerable (VU).

En la Figura 22 se evidencian las 39 presencias que se encontraban en los datos utilizados para la prueba que corresponden al 20% del total de datos del conjunto de test. Los mapas de probabilidad generados para la categoría Vulnerable (VU) muestran cómo las diferentes configuraciones y técnicas aplicadas impactaron las predicciones espaciales realizadas por los modelos de RL y RF. Estas visualizaciones reflejan la distribución geográfica estimada de la probabilidad de presencia de especies en esta categoría.

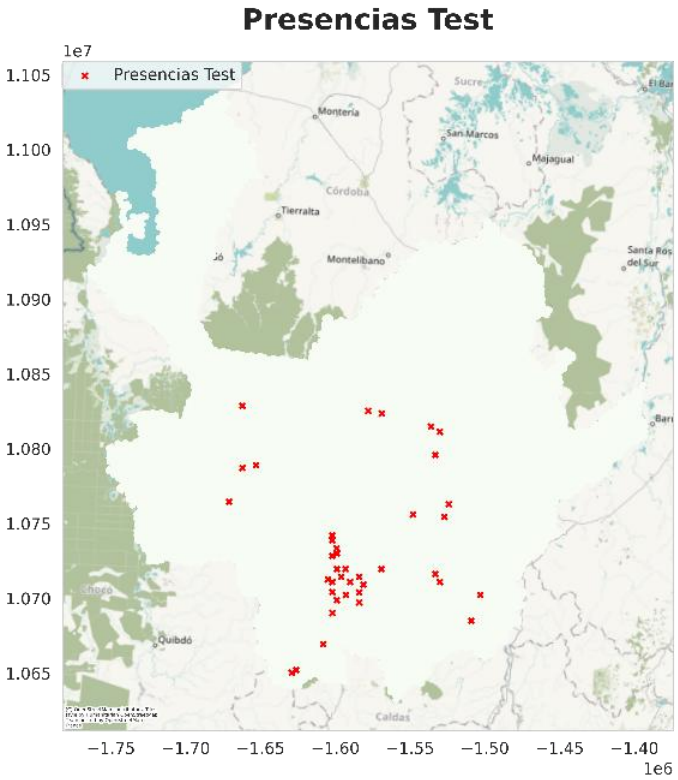


Figura 22. Mapa de presencias registradas en la evaluación del modelo. Categoría VU.

Análisis de mapas sin balanceo de clases

En la configuración inicial, donde no se aplicaron técnicas de balanceo y se incluyeron todas las variables, los mapas Figura 23 evidenciaron predicciones limitadas en las áreas de mayor presencia. Aunque ambos modelos mostraron áreas con probabilidades ligeramente más altas, la Regresión Logística mostró menor dispersión en las estimaciones, mientras que Random Forest identificó con más detalle ciertas regiones clave. Sin embargo, estas predicciones fueron marcadamente afectadas por el desbalance de clases, lo que restringió la detección de patrones relevantes en las áreas de interés.

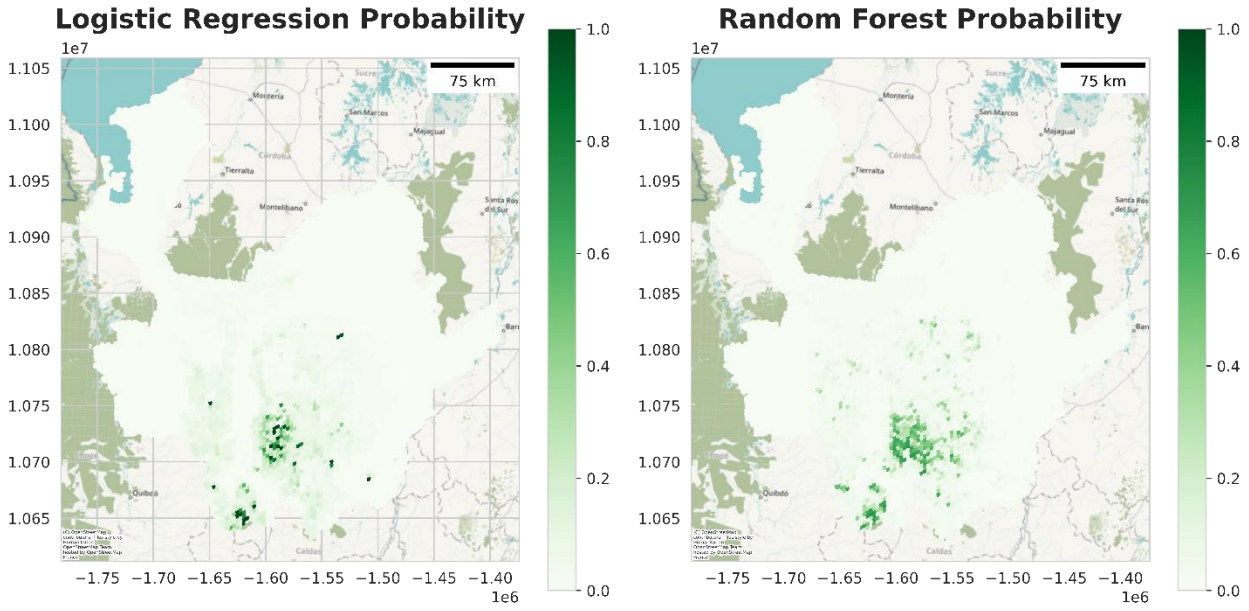


Figura 23. Mapa de probabilidad evaluación inicial de modelos Regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría VU.

Impacto del balanceo en los mapas

Con la implementación de Random Over Sampling y SMOTE, los modelos mostraron una mejora significativa en la representación de probabilidades en las áreas de mayor relevancia para la categoría VU Figuras 24 y 25. La técnica Random Over Sampling permitió aumentar la cobertura espacial en las predicciones. La RL presentó una dispersión más uniforme de las probabilidades, destacando varias áreas adicionales con valores medios. Por otro lado, RF mostró un enfoque más preciso en áreas específicas con probabilidades altas, indicando una mejor capacidad para identificar regiones prioritarias. Los mapas generados tras aplicar SMOTE mostraron una distribución más equilibrada en ambos modelos. La RL logró capturar patrones consistentes en áreas previamente subestimadas, mientras que Random Forest mantuvo un enfoque robusto en las áreas con alta probabilidad, reflejando una mayor precisión en la detección de estas zonas.

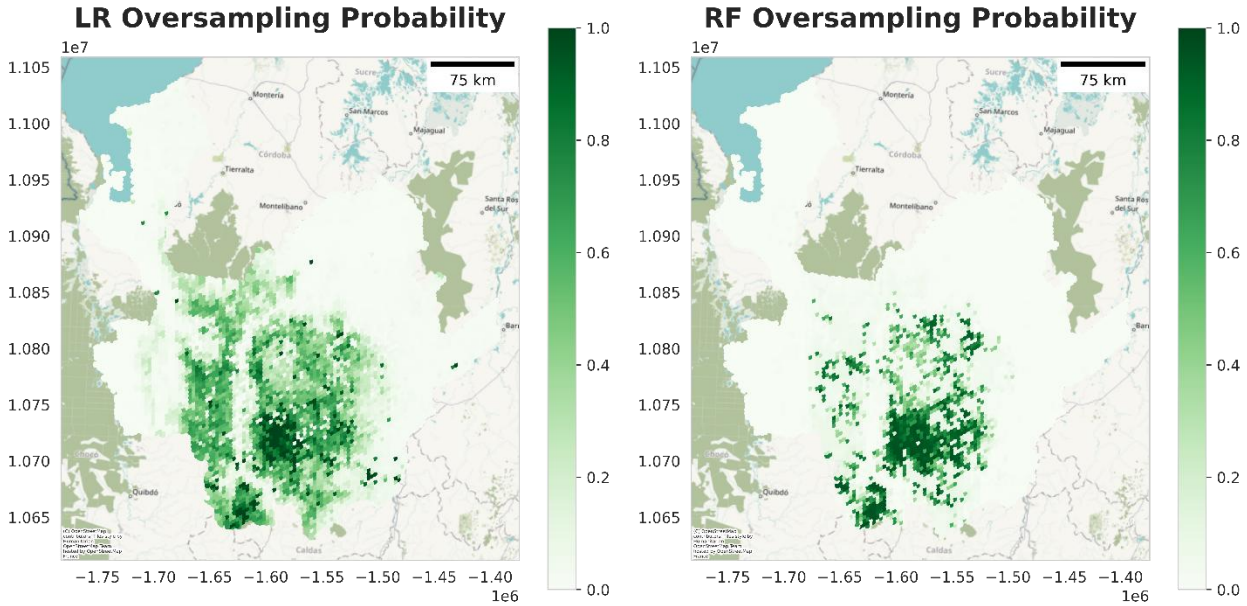


Figura 24. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con todas las variables disponibles. Categoría VU.

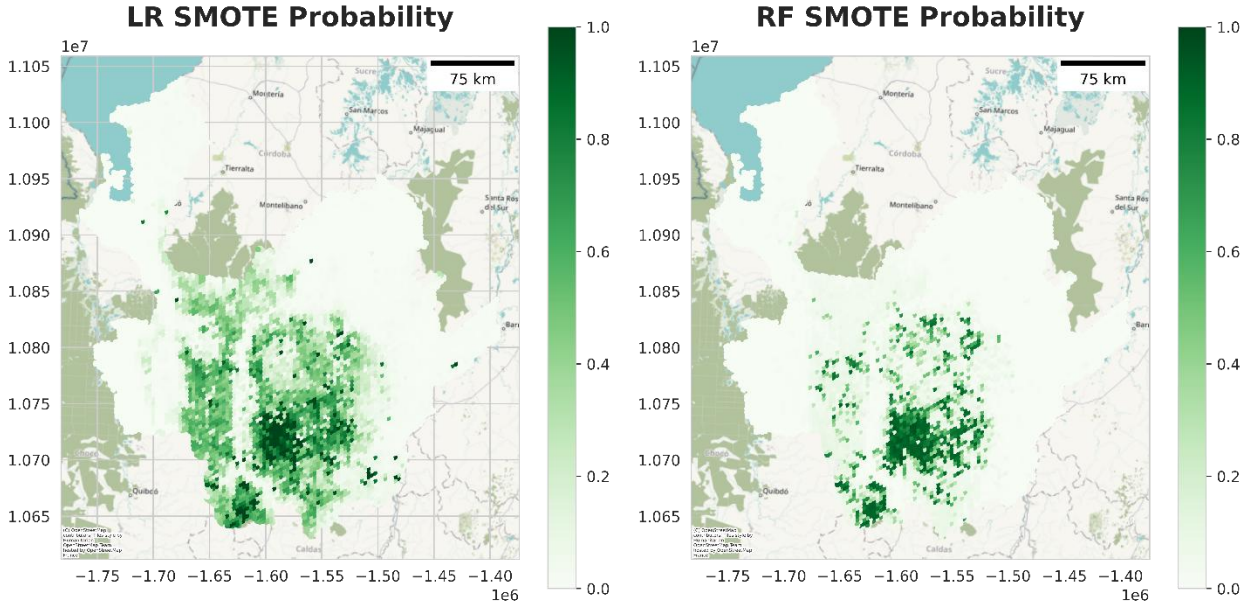


Figura 25. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría VU.

Selección de variables representativas

Después de la selección de variables representativas, los mapas Figura 26 reflejaron una disminución en la dispersión de las probabilidades, ya que los modelos priorizaron un conjunto

reducido de predictores. En esta configuración, ambos modelos mostraron áreas más focalizadas, aunque con menor cobertura general en comparación con las configuraciones previas. RF continuó destacando en la identificación de zonas con alta probabilidad, mientras que la RL mostró una pérdida notable de detalle espacial.

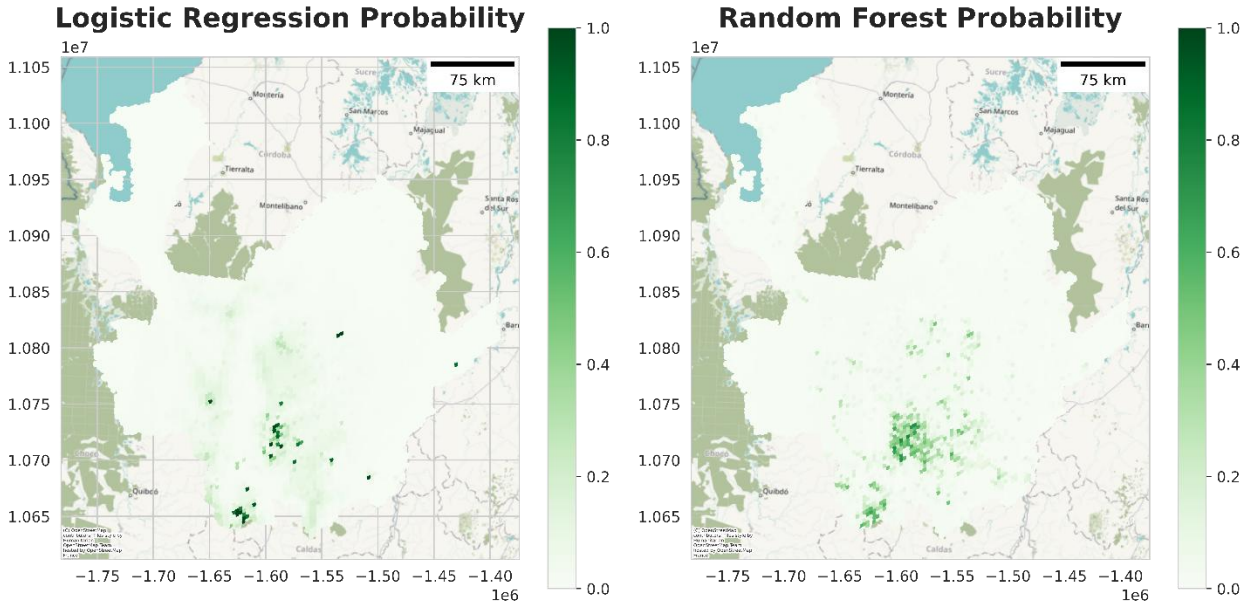


Figura 26. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría VU.

Selección de variables representativas y balanceo

Con la combinación de variables representativas y técnicas de balanceo, los modelos lograron una mejora significativa en las predicciones espaciales Figuras 27 y 28. Con la técnica Random Over Sampling los mapas mostraron una notable expansión en la cobertura de probabilidades altas, especialmente con RF. La RL presentó una mayor uniformidad en la probabilidad estimada. La técnica de SMOTE resaltó áreas clave con alta probabilidad, logrando un balance entre cobertura y precisión en ambos modelos. RF por su parte mostró un mejor desempeño al representar probabilidades más “duras” a lo largo de las regiones del departamento para la categoría VU.

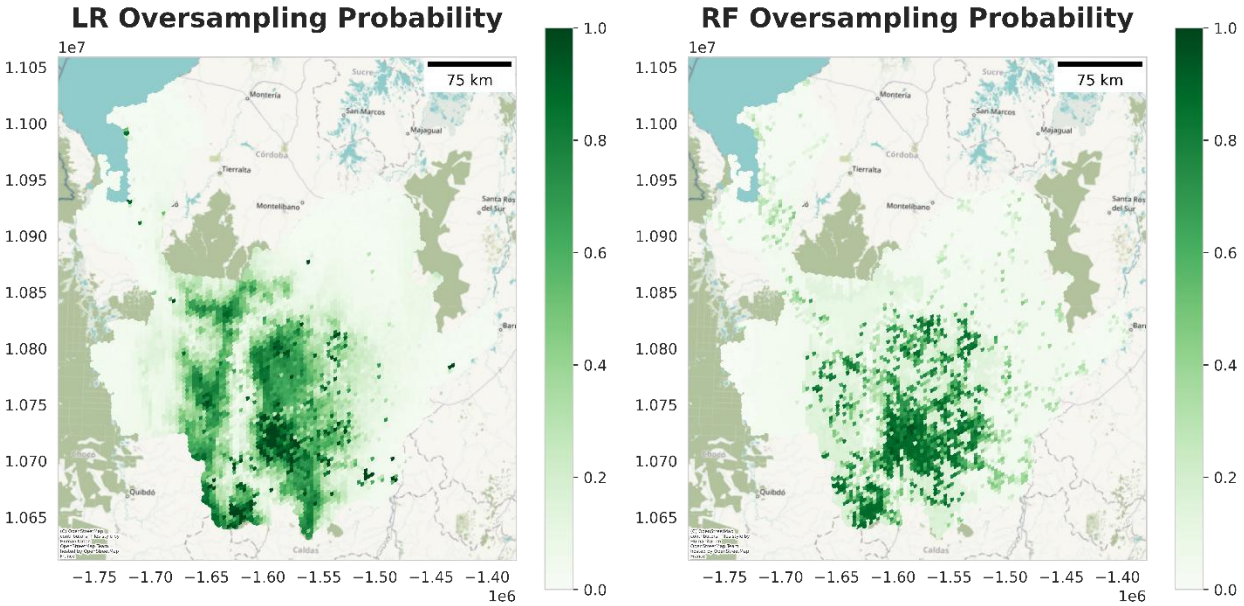


Figura 27. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas. Categoría VU.

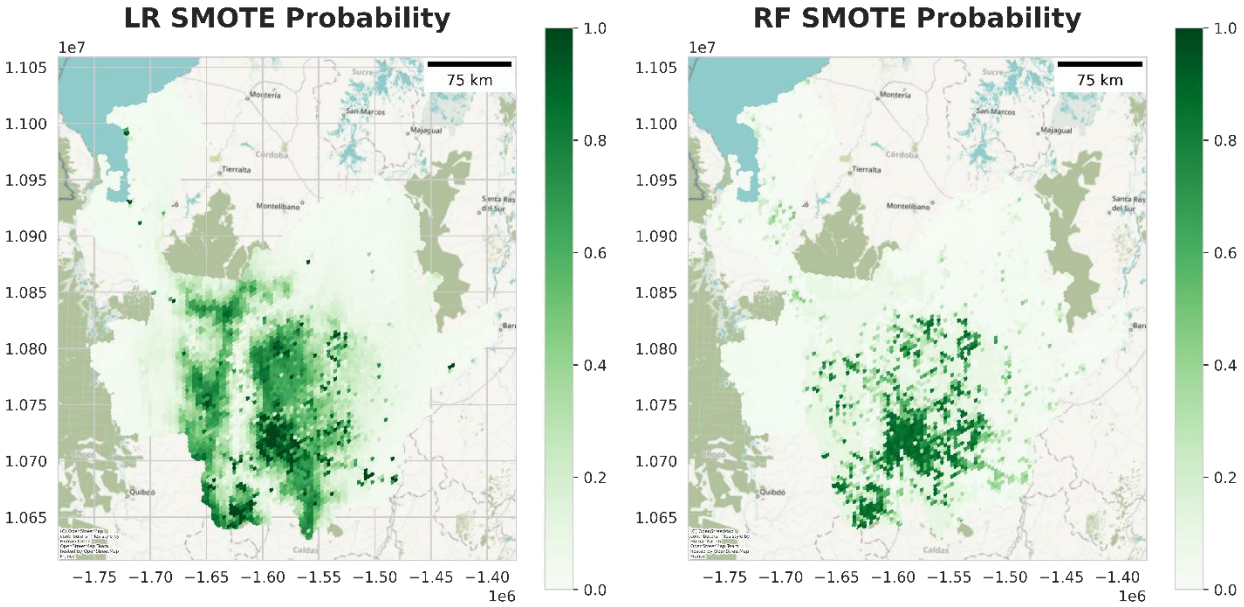


Figura 28. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría VU.

Adicionalmente, se realizó un análisis de importancia de características con el objetivo de identificar las variables que más contribuyeron a la predicción de avistamientos para cada categoría de conservación.

En el caso de la categoría Vulnerable (VU), tabla 7, el modelo entrenado con balanceo mediante SMOTE también identificó como más relevante la variable number_observers (MDI = 0.234797), seguida por effort_speed_kmph y elevation. Estas dos últimas variables, asociadas al esfuerzo de muestreo y características topográficas, respectivamente, fueron también seleccionadas por su bajo VIF en la etapa de preprocesamiento. Adicionalmente, se destacaron variables climáticas como p_driest_m y coberturas como lulc_2020_bosque, lo que indica que una combinación de factores de esfuerzo, ambientales y de cobertura fue determinante para predecir la presencia de especies en esta categoría.

Variable	MDI
number_observers	**0.234797**
effort_speed_kmph	**0.180260**
elevation	**0.088679**
lulc_2020_bosque	**0.023111**
p_driest_m	**0.021123**
lulc_2014_formacion_natural_no_forestal_inundable	0.016922
lulc_2019_palma_aceitera	0.004934
lulc_2014_silvicultura	0.002300
lulc_2017_bosque_inundable	0.001869
lulc_2020_palma_aceitera	0.000977
lulc_2014_playas_dunas_bancos_de_arena	0.000940
lulc_2019_otra_formacion_natural_no_forestal	0.000748
lulc_2018_silvicultura	0.000521
lulc_2020_no_observado	0.000481
lulc_2017_silvicultura	0.000472
lulc_2020_mineria	0.000470
lulc_2020_manglar	0.000407
lulc_2018_infraestructura_urbana	0.000401
lulc_2014_mosaico_de_agricultura_pasto	0.000369
lulc_2014_manglar	0.000213
lulc_2020_silvicultura	0.000075
lulc_2016_otra_formacion_natural_no_forestal	0.000067
lulc_2020_formacion_natural_no_forestal_inundable	0.000048
lulc_2015_mosaico_de_agricultura_pasto	0.000047
lulc_2019_silvicultura	0.000014
lulc_2018_otra_formacion_natural_no_forestal	0.000013
lulc_2014_otra_formacion_natural_no_forestal	0.000007
lulc_2019_no_observado	0.000000
lulc_2014_otra_area_sin_vegetacion	0.000000
lulc_2014_no_observado	0.000000
lulc_2017_playas_dunas_bancos_de_arena	0.000000
lulc_2019_mosaico_de_agricultura_pasto	0.000000

lulc_2020_infraestructura_urbana	0.000000
lulc_2014_mineria	0.000000
lulc_2020_otra_area_sin_vegetacion	0.000000

Tabla 7. Resultado de la importancia de las características del modelo Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas para categoría Vulnerable (VU).

4.3.4. Sumario

En general, RF se destacó en todas las configuraciones, mostrando mapas más precisos y detallados que capturaron áreas críticas para la categoría VU. Las técnicas de balanceo, posterior a la selección de variables representativas, particularmente SMOTE, fueron esenciales para mejorar la representación de probabilidades en zonas con alta relevancia biológica. La RL, aunque consistente, mostró limitaciones en la detección de detalles espaciales y fue más dependiente del balanceo para mejorar sus predicciones.

Los resultados obtenidos confirman la pertinencia de la selección de variables realizada mediante el análisis de colinealidad, ya que las variables preservadas no solo eran estadísticamente independientes, sino que también resultaron ser las más relevantes desde el punto de vista predictivo. Este hallazgo respalda la estrategia metodológica adoptada para el desarrollo de los modelos.

5. CONCLUSIONES

El presente trabajo logró cumplir con los objetivos planteados, integrando metodologías de preprocesamiento, análisis y modelado predictivo para apoyar la conservación de especies endémicas en Colombia. A través de la identificación de áreas prioritarias, el análisis de factores exógenos, la implementación de algoritmos de aprendizaje automático y la validación de resultados, se desarrollaron herramientas para comprender y predecir la distribución de estas especies. A continuación, se presentan las principales conclusiones del estudio.

Desarrollo de metodología de preprocesamiento de datos de avistamientos de aves, con el fin de identificar departamentos de alta concentración de especies endémicas: se implementó una metodología para el preprocesamiento de datos de avistamientos de aves, abordando problemas como la variabilidad en la calidad de los datos y corrigiendo sesgos espaciales y temporales inherentes a la recolección colaborativa. Paralelamente, se procesaron variables exógenas, como clima, suelo y elevación, que permitieron explorar la relación entre factores ambientales y la distribución de estas especies. Los análisis realizados determinaron que los departamentos de Risaralda y Antioquia presentan una mayor concentración de especies endémicas en peligro, como *Henicorhina negreti* (Peligro Crítico), *Penelope perspicax* (en Peligro) y *Hypopyrrhus pyrohypogaster* (Vulnerable). El uso de mapas de calor permitió visualizar claramente la distribución espacial de estas especies, facilitando la identificación de patrones de abundancia y concentración en diferentes regiones.

Desarrollo de metodología de preprocesamiento de datos de variables exógenas (precipitaciones, temperaturas, antropogénicas, entre otras) para establecer correlaciones sobre los avistamientos de especies endémicas: el análisis de las variables exógenas, como la temperatura media anual, precipitación anual, elevación y uso de suelo, permitió establecer correlaciones significativas con la distribución de especies endémicas. Estas especies tienden a concentrarse en áreas con características climáticas y topográficas favorables, destacando la influencia de estas variables en su presencia. Este enfoque ha proporcionado una base para comprender los factores que afectan la biodiversidad y servirá como marco de referencia para integrar en el futuro otro tipo de modelos.

Implementación y validación de algoritmos de aprendizaje automático para predicción de avistamientos de especies endémicas: en resumen, los resultados de los modelos evaluados para las tres categorías de conservación (Peligro Crítico, En Peligro y Vulnerable) indican que Random Forest es consistentemente el modelo más efectivo para clasificar correctamente la clase minoritaria, especialmente cuando se aplican técnicas de balanceo como Random Over Sampling en las categorías CR y EN, y SMOTE en la categoría VU con variables representativas. Las métricas de validación, como la sensibilidad y el F1-Score, confirmaron la capacidad de los modelos para

identificar correctamente especies en las categorías más vulnerables. Asimismo, el análisis de relevancia de variables mediante la métrica MDI permitió identificar consistentemente a *number_observers* como la variable de mayor influencia en las tres categorías, en conjunto con coberturas del suelo y variables climáticas o topográficas seleccionadas previamente por su baja colinealidad. Finalmente, los mapas de probabilidad generados permitieron visualizar patrones espaciales y temporales, ofreciendo herramientas claras para comunicar hallazgos complejos. Estas visualizaciones fortalecen su aplicación en la planificación de estrategias de conservación a nivel regional.

Este trabajo logró integrar metodologías de preprocesamiento, análisis y predicción para apoyar la conservación de especies endémicas en Colombia. Sin embargo, se reconocen limitaciones como la disponibilidad de datos actualizados y la validación en otros contextos geográficos.

6. REFERENCIAS BIBLIOGRAFICAS

[1] A. Bárcena, A. Prado, J. Samaniego y S. Malchik, La Economía del Cambio Climático en América Latina y el Caribe. División de Desarrollo Sostenible y Asentamientos Humanos, CEPAL, Publicación de las Naciones Unidas, 2010.

[2] E. Uribe Botero, Estudios del cambio climático en América Latina: El cambio climático y sus efectos en la biodiversidad en América Latina. Naciones Unidas, Comisión Económica para América Latina y el Caribe (CEPAL), p. 13, 2015.

[3] Asociación Colombiana de Ornitología, "Lista de referencia de especies de aves de Colombia 2020," [En línea]. Disponible: https://ipt.biodiversidad.co/sib/resource?r=aco_listaavescolombia2017#anchorcitation. [Accedido: enero 2020].

[4] C. M. Fernández Barrios, "Incidencia del cambio climático sobre la distribución espacial de tres de las especies de aves con mayor grado de amenaza en Colombia," Tesis de grado, Universidad Antonio Nariño, Bogotá D.C., Colombia, 2023.

[5] A. T. Norman, Aprendizaje automático en acción: Un libro para el lego, guía paso a paso para los novatos. Tektime, 2019.

[6] M. Parry, O. Canziani, J. Palutikof, P. van der Linden y C. Hanson, Cambio Climático 2007: Impacto, Adaptación y Vulnerabilidad. IPCC, 2007.

[7] A. Székely, Latinoamérica y la biodiversidad. México, 2009.

[8] Ç. Sekercioglu, J. Wormworth y R. Primack, "The effects of climate change on tropical birds," Biological Conservation, pp. 1–18, 2011.

[9] eBird, "Base de datos de avistamiento de aves a nivel mundial," [En línea]. Disponible: https://ebird.org/home. [Consultado: 13 de diciembre de 2023].

[10] J. H. Maldonado, R. P. Moreno-Sánchez, S. Espinoza, A. Bruner, N. Garzón y J. Myers, "Peace is much more than doves: The economic benefits of bird-based tourism as a result of the peace treaty in Colombia," World Development, vol. 106, pp. 78–86, 2018.

[11] A. M. Santos y J. M. Silva, "Machine Learning Models for Bird Species Distribution Prediction: A Comprehensive Review," Ecological Informatics, vol. 60, p. 101138, 2020.

[12] R. Gupta y S. Chawla, "Predicting Bird Species Distribution using Machine Learning: A Case

Study in the Western Ghats, India," International Journal of Computer Applications, vol. 181, no. 3, pp. 18–24, 2018.

[13] J. Hortal y J. Lobo, "Modelos predictivos: Un atajo para describir la distribución de la diversidad biológica," enero 2003. [En línea]. Disponible: https://www.researchgate.net/publication/26495204_Modelos_predictivos_Un_atajo_para_describir_la_distribucion_de_la_diversidad_biologica.

[14] R. G. Mateo, "Modelos de distribución de especies: Una revisión sintética," Revista Chilena de Historia Natural, vol. 84, no. 2, pp. 145–164, 2011. [En línea]. Disponible: https://www.scielo.cl/pdf/rchnat/v84n2/art08.pdf.

[15] A. Moreno et al., Aprendizaje automático. 1994. *(Falta editorial o universidad, complétalo si aplica)*

[16] L. E. M. Aplicadas, "Aprendizaje no supervisado y el algoritmo wake-sleep en redes neuronales," Tesis doctoral, Universidad Tecnológica de la Mixteca, 2012.

[17] G. Valenzuela González, "Aprendizaje Supervisado: Métodos, Propiedades y Aplicaciones," 2022.

[18] IBM, "Aprendizaje supervisado," [En línea]. Disponible: https://www.ibm.com/es-es/topics/supervised-learning. [Consultado: 22 de noviembre de 2023].

[19] R. G. Mateo, Á. M. Felicísimo y J. Muñoz, "Modelos de distribución de especies: Una revisión sintética," Revista Chilena de Historia Natural, vol. 84, no. 2, pp. 145–164, 2011.

[20] C. J. Raxworthy, "Predicting distributions of known and unknown reptile species in Madagascar," Nature, vol. 426, pp. 837–841, 2003.

[21] R. G. Mateo, Á. M. Felicísimo y J. Muñoz, "Modelos de distribución de especies: Una revisión sintética," Revista Chilena de Historia Natural, vol. 84, no. 2, pp. 145–164, 2011.

[22] A. Johnston et al., "Analytical guidelines to increase the value of community science data: An example using eBird data to estimate species distribution," Cornell Lab of Ornithology, University, Ithaca, NY, USA, 2021.

[23] A. Johnston et al., "Abundance models improve spatial and temporal prioritization of conservation resources," Ecological Applications, vol. 25, no. 7, pp. 1749–1762, 2015.

[24] R. K. Heikkinen, M. Marmion y M. Luoto, "Does the interpolation accuracy of species distribution models come at the expense of transferability?" Ecological Modelling, vol. 386, pp.

1–7, 2018.

[25] M. Swan, M. Le Pla, J. Di Stefano, J. Pascoe y T. D. Penman, “Species distribution models for conservation planning in fire-prone landscapes,” School of Ecosystem and Forest Sciences, University of Melbourne, 2021.

[26] D. Silvestro et al., “Improving Biodiversity Protection Through Artificial Intelligence,” Department of Biology, University of Fribourg, Fribourg, 2022.

[27] E. Chalco, “Variaciones Espaciotemporales de Distribución de la Rana Marsupial Andina (Gastrotheca Riobambae) y su Relación con la Expansión Urbana de Quito,” Tesis de pregrado, Facultad de Ciencias del Medio Ambiente, Univ. Tecnológica Indoamérica, Quito, Ecuador, 2022.

[28] A. Deomurari, A. Sharma, D. Ghose y R. Singh, “Projected Shifts in Bird Distribution in India under Climate Change,” Academic Editor: Michael Wink, 2023.

[29] M. Strimas-Mackey et al., “Best Practices for Using eBird Data,” version 2.0, Cornell Lab of Ornithology, Ithaca, NY, 2023. [En línea]. Disponible: https://ebird.github.io/ebird-best-practices/. doi: 10.5281/zenodo.3620739.

[30] Ministerio de Ambiente y Desarrollo Sostenible (Colombia), “Resolución No. 0126 del 06 de febrero de 2024 por la cual se establece el listado oficial de las especies silvestres amenazadas de la biodiversidad biológica colombiana continental y marino costera,” 2024.

[31] N. Myers, R. A. Mittermeier, C. G. Mittermeier, G. A. B. Da Fonseca y J. Kent, “Biodiversity hotspots for conservation priorities,” Nature, vol. 403, no. 6772, pp. 853–858, 2000.

[32] C. J. Krebs, Ecological Methodology. Benjamin/Cummings, 1999.

[33] MapBiomas Colombia, “MapBiomas: Cobertura y uso del suelo en Colombia,” [En línea]. Disponible: https://colombia.mapbiomas.org/. [Accedido: agosto 2024].

[34] Google Colab, “Repositorio de Google Colab GEE,” [En línea]. Disponible: https://colab.research.google.com/drive/1vNg_SnKJoZbERVRX2LqEkdvV0pzLF07w. [Accedido: agosto 2024].

[35] J. S. Blandón, “MapBiomas GEE Python Repository,” GitHub, [En línea]. Disponible: https://github.com/jsblandon/mapbiomas_gee_py. [Accedido: agosto 2024].

[36] WorldClim, “WorldClim: Global Climate Data,” [En línea]. Disponible: https://worldclim.org/data/index.html. [Accedido: 12 de octubre de 2024].

[37] J. C. Ortíz-Yusty et al., “Distribución de aves endémicas y migratorias en el departamento de Risaralda: Patrones y prioridades de conservación,” Revista Ornitológica Colombiana, vol. 12, no. 1, pp. 45–58, 2020. [En línea]. Disponible: https://revistas.humboldt.org.co.

[38] R. Maggini, S. Jenni, W. F. La Sorte y H. P. L. Schmid, “Elevational shifts in alpine bird species: A response to climate and topography interactions,” Nature Climate Change, vol. 6, no. 6, pp. 449–455, Jun. 2016.

[39] T. T. McGee, “Alternative Tessellations for the Identification of Urban Employment Subcenters: A Comparison of Triangles, Squares, and Hexagons,” Urban Informatics, vol. 9, no. 1, pp. 1–25, 2024. doi: 10.1007/s41651-024-00200-5.

[40] J. O’Brien, “A Caution Regarding Rules of Thumb for Variance Inflation Factors,” Quality & Quantity, vol. 41, no. 5, pp. 673–690, 2007. doi: 10.1007/s11135-006-9018-6.

[41] J. E. Jaynes, “Information Theory and Statistical Mechanics,” Physical Review, vol. 106, no. 4, pp. 620–630, May 1957. doi: 10.1103/PhysRev.106.620.

[42] W. S. McCulloch y W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” The Bulletin of Mathematical Biophysics, vol. 5, pp. 115–133, 1943. doi: 10.1007/BF02478259.

[43] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[44] D. R. Cox, “The regression analysis of binary sequences,” Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215–232, 1958.

[45] J. H. Fawcett, “An Introduction to ROC Analysis,” Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.

[46] I. H. Witten y E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. San Francisco, CA: Morgan Kaufmann, 2005.

[47] D. M. W. Powers, “Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation,” Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011. [En línea]. Disponible: https://www.researchgate.net/publication/281278629.

[48] C. J. van Rijsbergen, Information Retrieval, 2nd ed. London: Butterworth-Heinemann, 1979.

[49] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 3rd ed. Sebastopol, CA: O'Reilly Media, Inc., 2022.

[50] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

7. ANEXO

Repositorio de GitHub con el código correspondiente a la metodología para la predicción de avistamientos de aves y la conservación de especies endémicas mediante algoritmos de aprendizaje automático:

https://github.com/mvescobarm/Prediccion-de-avistamientos-de-aves-utilizando-algoritmos-de-aprendizaje-automatico.git

7.1. ANEXOS OBJETIVO 1

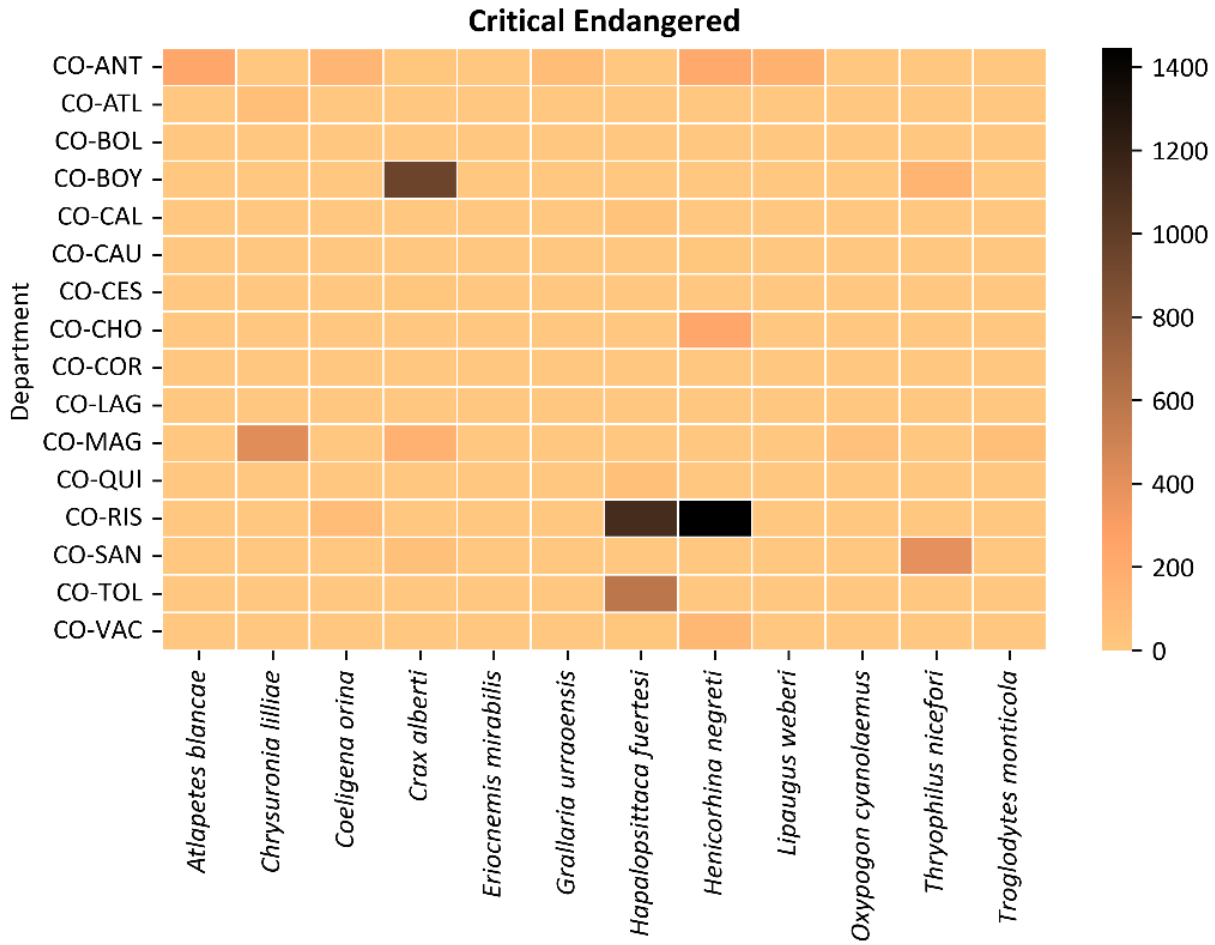


Figura 29. Abundancia de aves endémicas por departamento (2003-2023) en nivel de Peligro Crítico (CR).

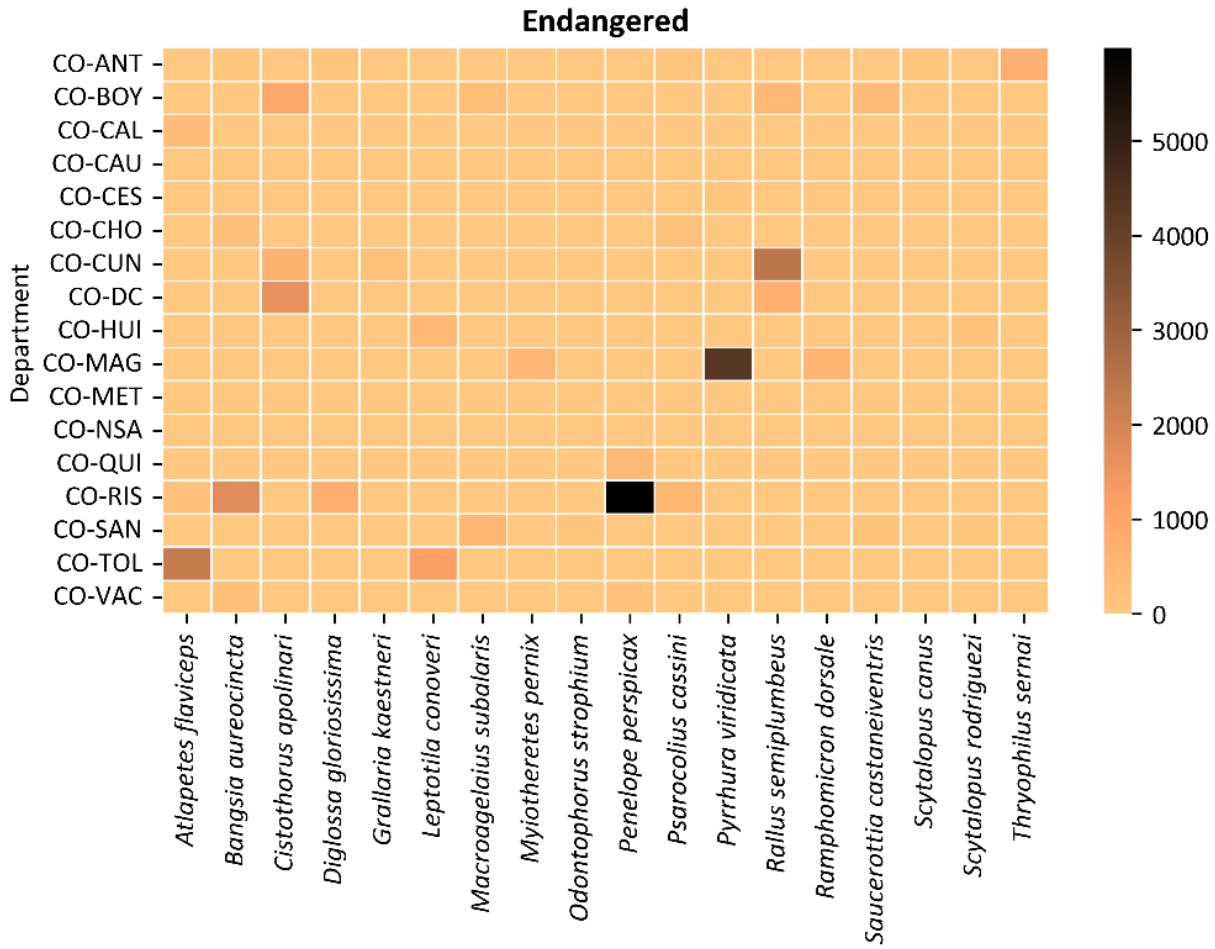


Figura 30. Abundancia de aves endémicas por departamento (2003-2023) en nivel En Peligro (EN).

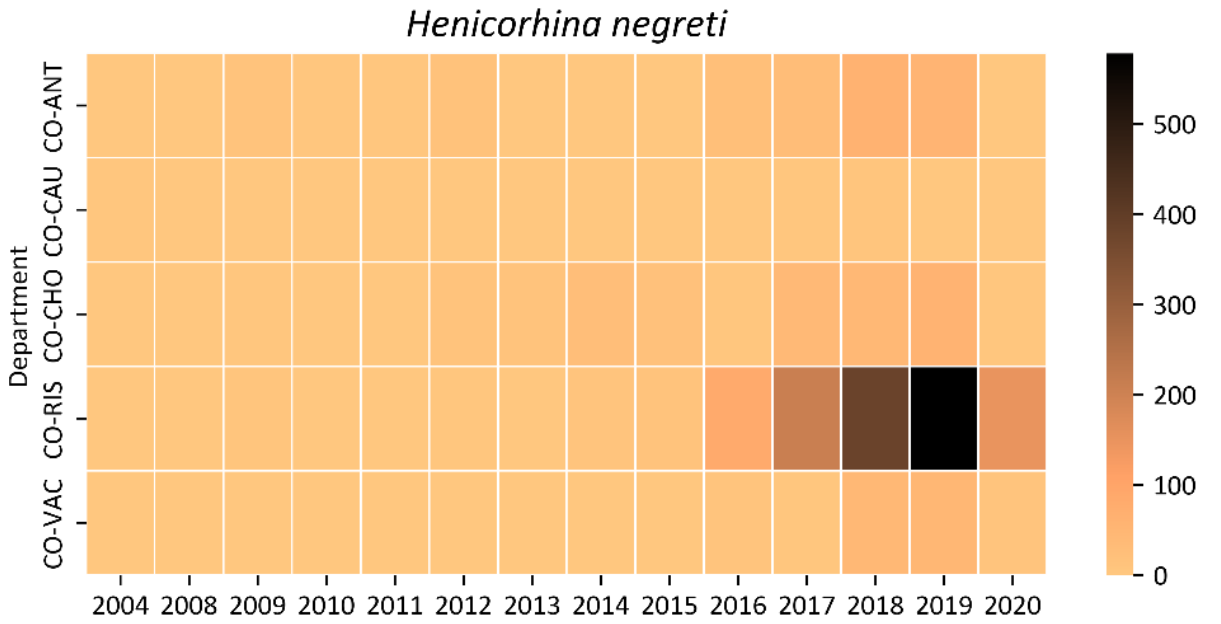


Figura 31. Abundancia de aves endémicas para la *Hernicorhina negreti*. La representación confirma que Risaralda es el departamento que durante cinco años seguidos presentó la mayor concentración de la especie.

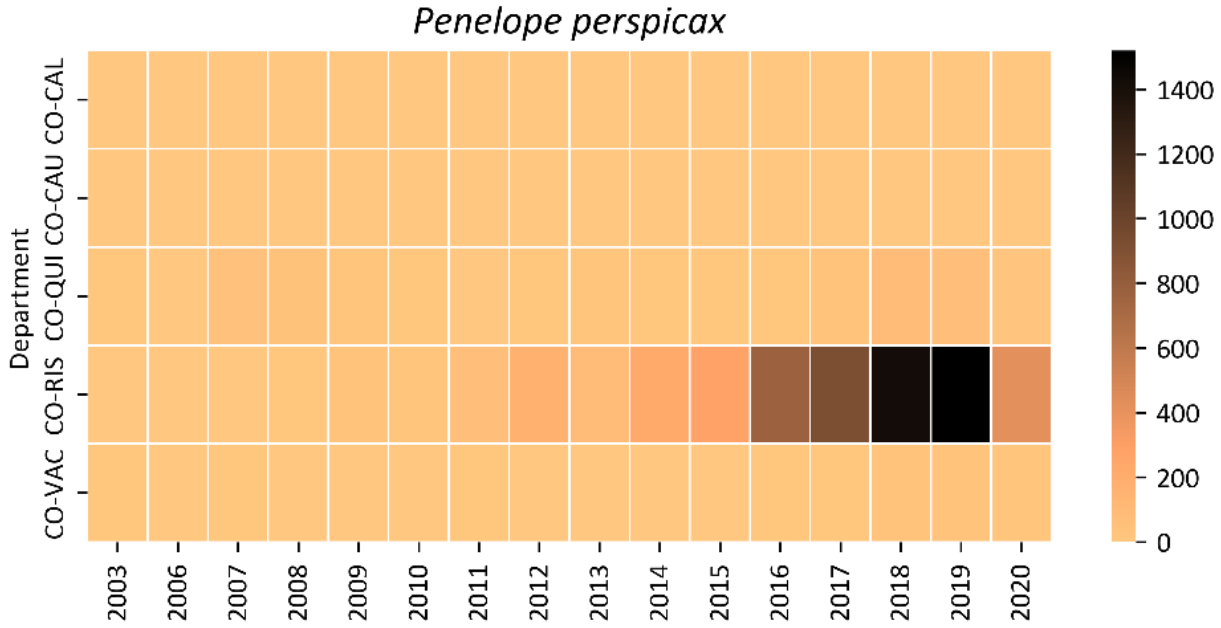


Figura 32. Abundancia de aves endémicas para la *Penelope perspicax*. La representación confirma que Risaralda es el departamento que durante cinco años seguidos presentó la mayor concentración de la especie.

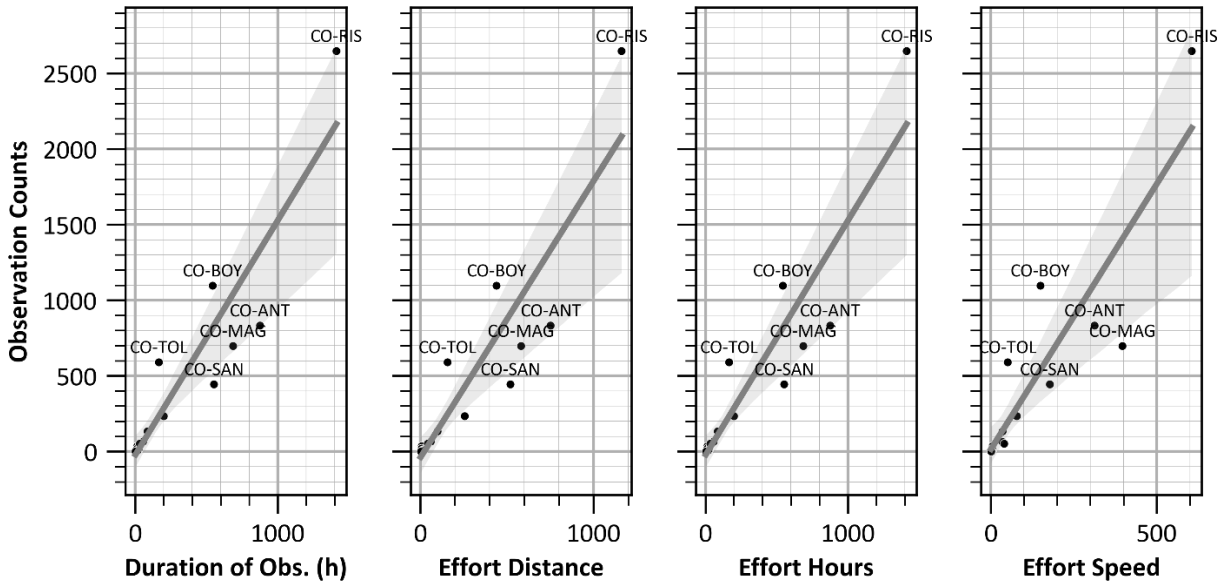


Figura 33. Abundancia vs variables de esfuerzo. Aves endémicas en CR en Colombia. 2003 – 2023. Las variables de esfuerzo corresponden a: Duración de la observación (Duration of Obs. (h)), Distancia de esfuerzo (Effort Distance), Horas de esfuerzo (Effort Hours) y Velocidad de Esfuerzo (Effort Speed). Los valores para cada variable corresponden a las sumas totales por departamento.

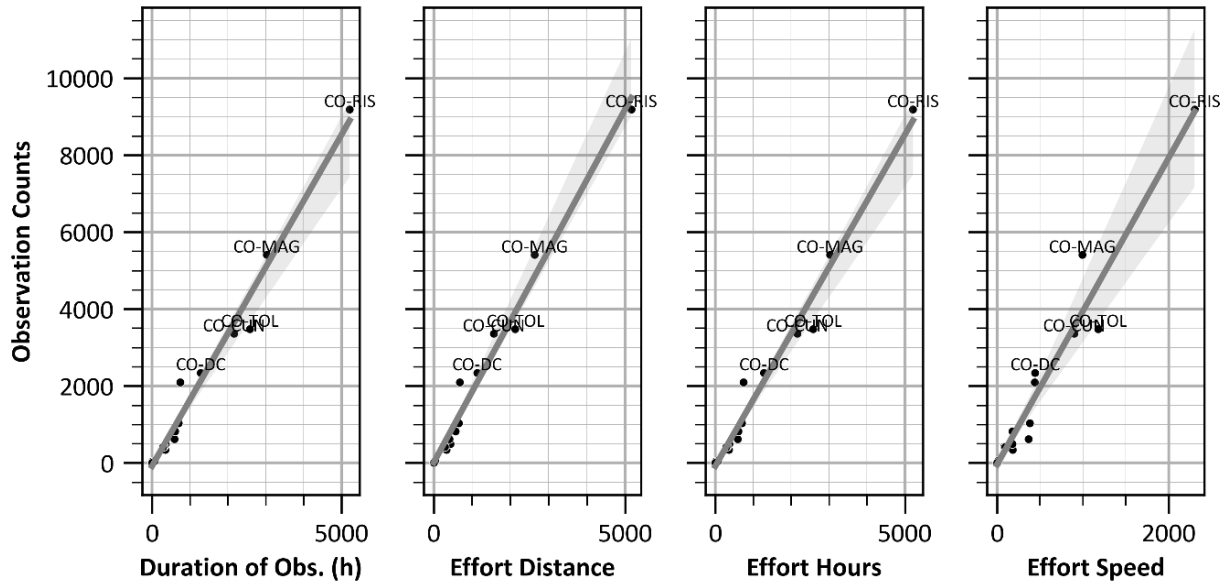


Figura 34. Abundancia vs variables de esfuerzo. Aves endémicas EN, en Colombia. 2003 – 2023. Las variables de esfuerzo corresponden a: Duración de la observación (Duration of Obs. (h)), Distancia de esfuerzo (Effort Distance), Horas de esfuerzo (Effort Hours) y Velocidad de Esfuerzo (Effort Speed). Los valores para cada variable corresponden a las sumas totales por departamento.

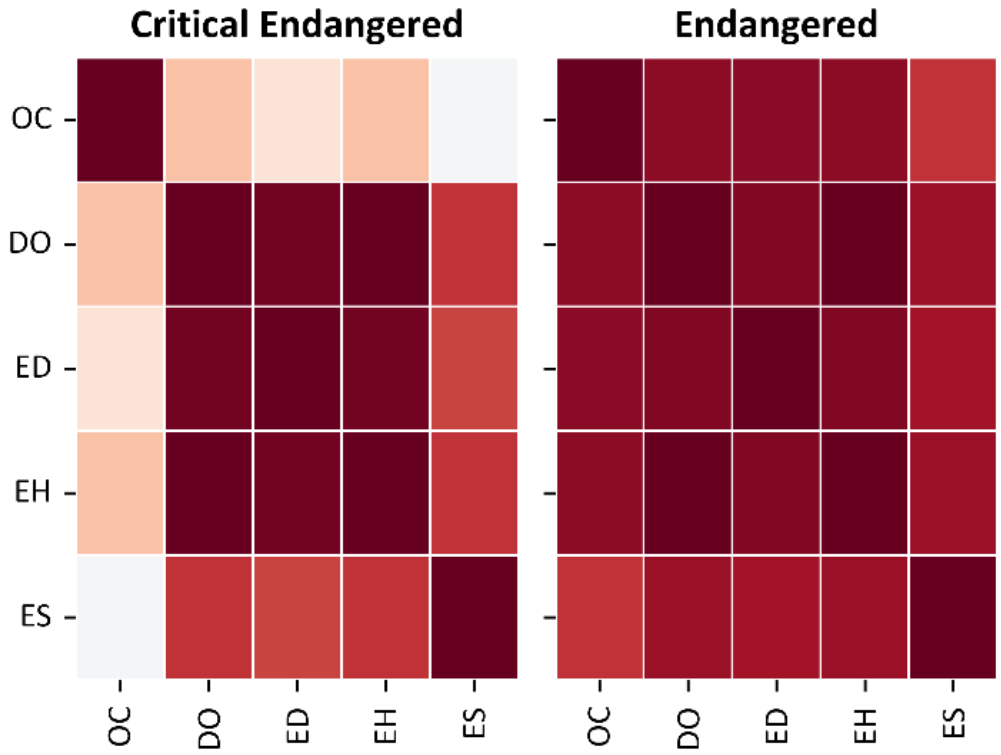


Figura 35. Matriz de correlación entre las variables de esfuerzo y aves endémicas categoría CR y EN de Colombia. OC: Observation Count. DO: Duration Observation [hr]. ED: Effort Distance [km]. EH: Total Effort Hours [hr]. ES: Total Effort Speed [km/h].

Figura 36. Aves con mayor abundancia en las categorías Peligro Crítico – CR (Henicorhina negreti) y En Peligro – EN (Penelope perspicax). Imágenes obtenidas de Birds of the World. Cornell Lab of Ornithology, Ithaca, NY, EE.UU.

7.2. ANEXOS OBJETIVO 2

Departamento Risaralda		
N°	**Variable**	**VIF**
1	number_observers	1.007276
2	lulc_2014_bosque	2.735594
3	lulc_2014_mineria	1.013548
4	lulc_2014_otra_area_sin_vegetacion	1.771380
5	lulc_2014_rio_lago_oceano	1.002478
6	lulc_2016_bosque	2.408034
7	lulc_2017_bosque	1.022207
8	lulc_2017_mosaico_de_agricultura_pasto	3.179864
9	lulc_2018_mosaico_de_agricultura_pasto	2.710037
10	lulc_2018_otra_formacion_natural_no_forestal	1.275549
11	lulc_2018_rio_lago_oceano	1.000016
12	lulc_2019_otra_area_sin_vegetacion	4.596334
13	lulc_2020_afloramiento_rocoso	2.755494
14	lulc_2020_otra_area_sin_vegetacion	3.594014
15	lulc_2020_otra_formacion_natural_no_forestal	1.011007
16	lulc_2020_rio_lago_oceano	1.000000
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-
-	-	-

Tabla 8. Variables seleccionadas para matriz de correlación de variables exógenas posterior a diputación VIF.

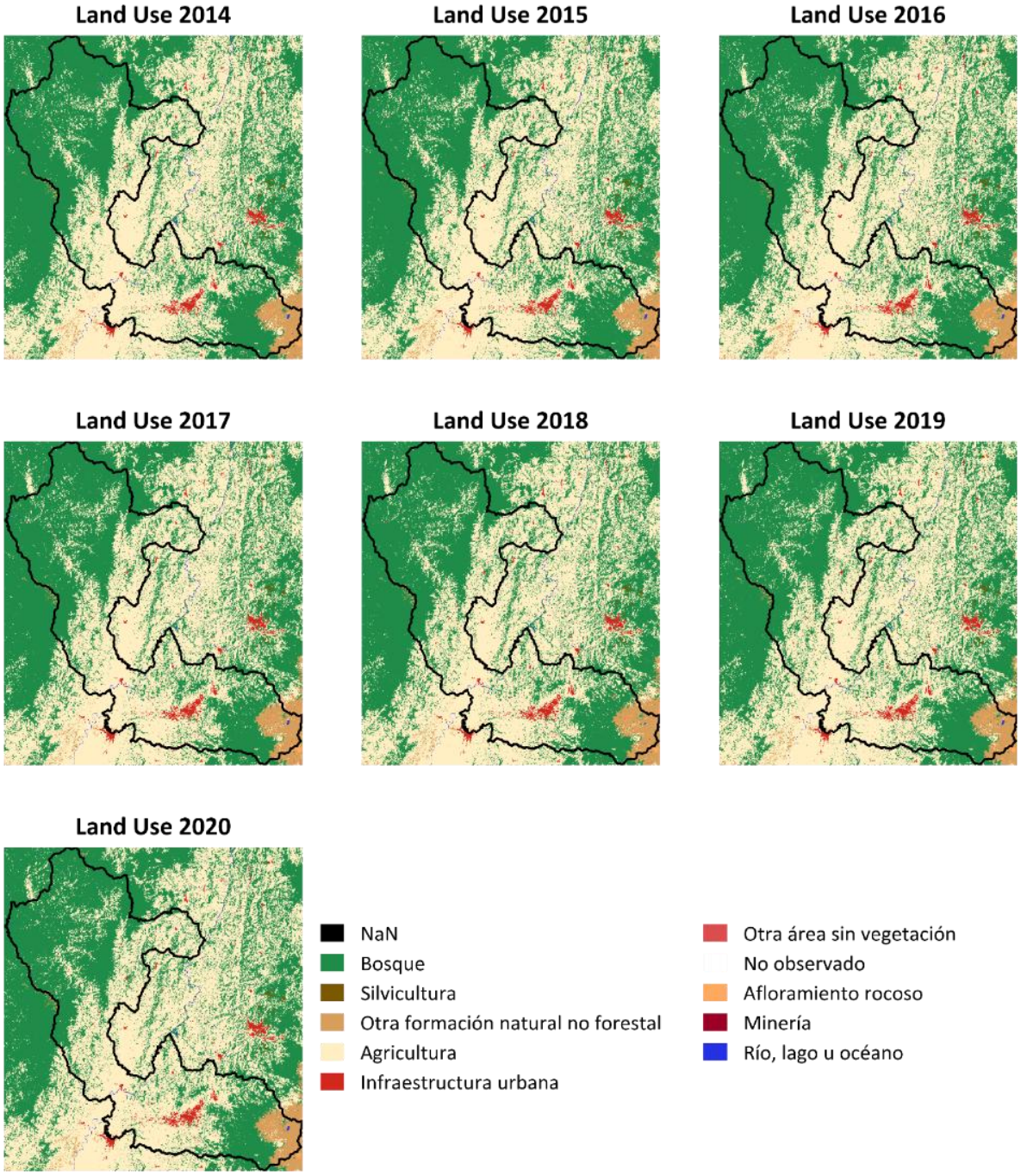


Figura 37. Variación de uso del suelo 2014 - 2020 en el departamento de Risaralda, procesamiento de base de datos de MapBiomas de la plataforma Google Earth Engine.

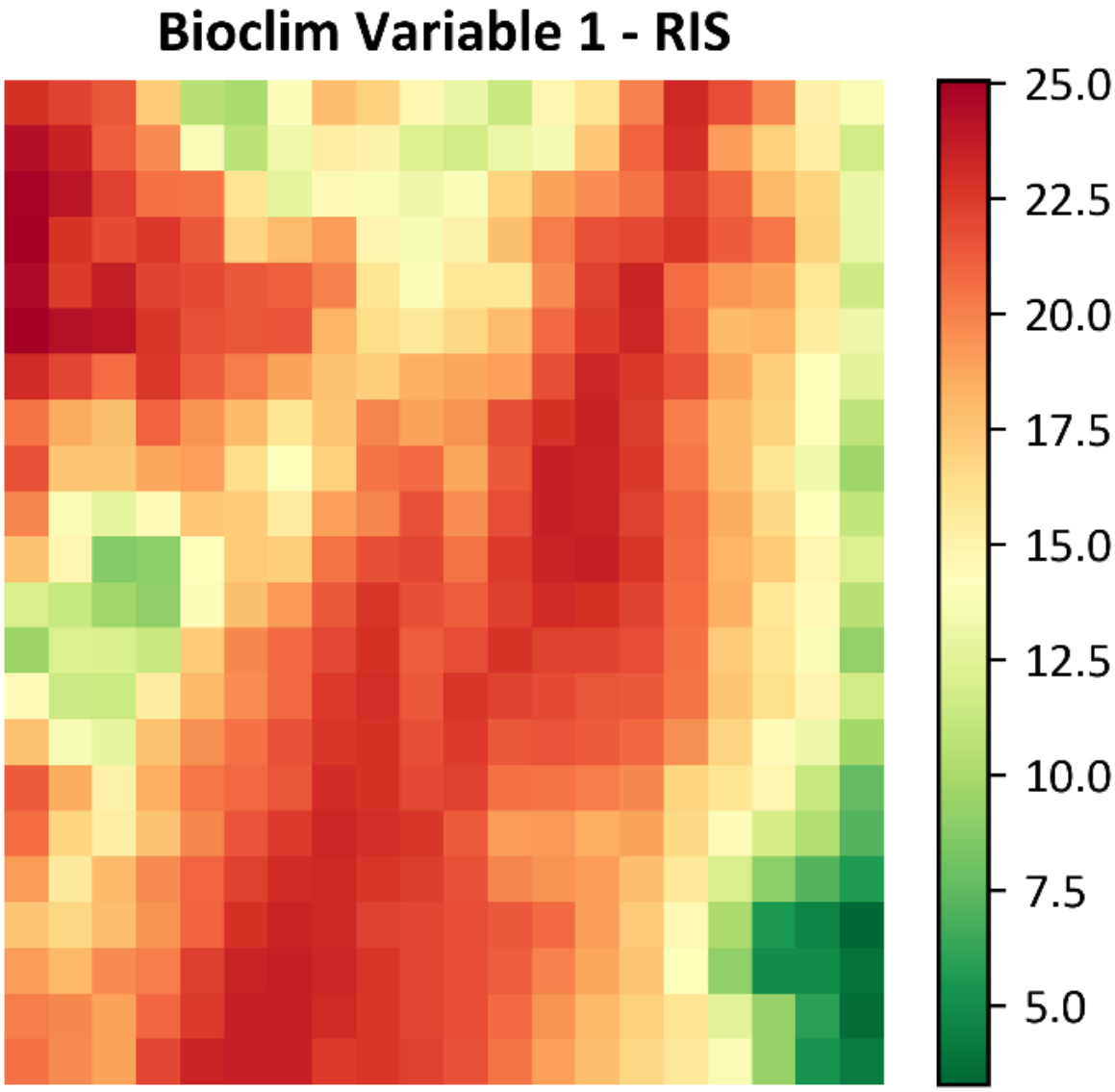


Figura 38. Representación espacial de la variable Bioclimática Bio1: Temperatura media anual para el departamento de Risaralda.

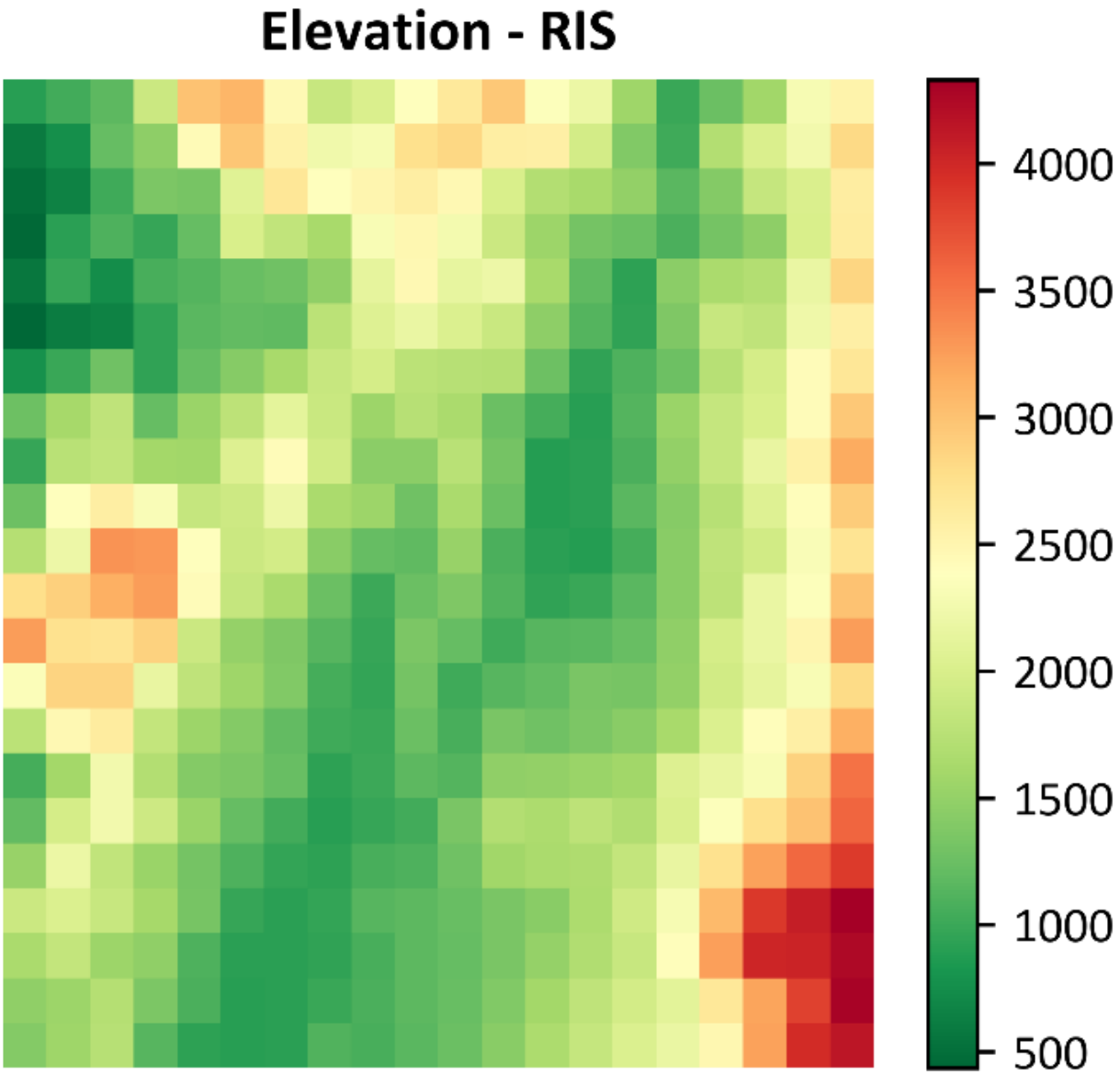


Figura 39. Representación espacial de la elevación para ambos departamentos. La resolución de esta variable condiciona el detalle con el que se pueda apreciar la tendencia de la variable en el departamento de Risaralda.

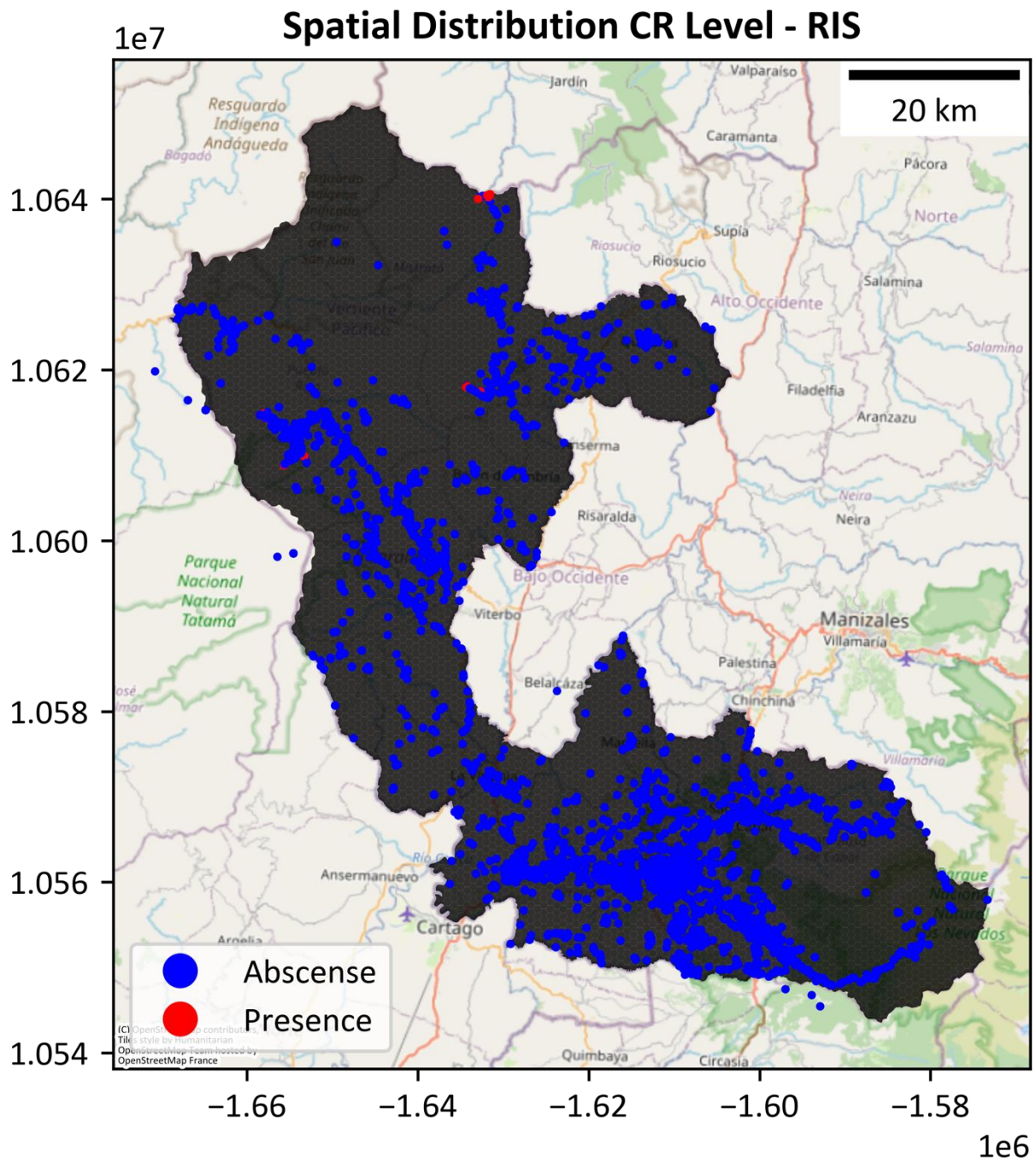


Figura 40. Distribución espacial y representación hexagonal de registros de *Henicorhina Negreti* en Risaralda. Los puntos rojos representan la presencia de la especie, mientras que los azules corresponden a las ausencias obtenidas a partir del preprocesamiento de los datos.

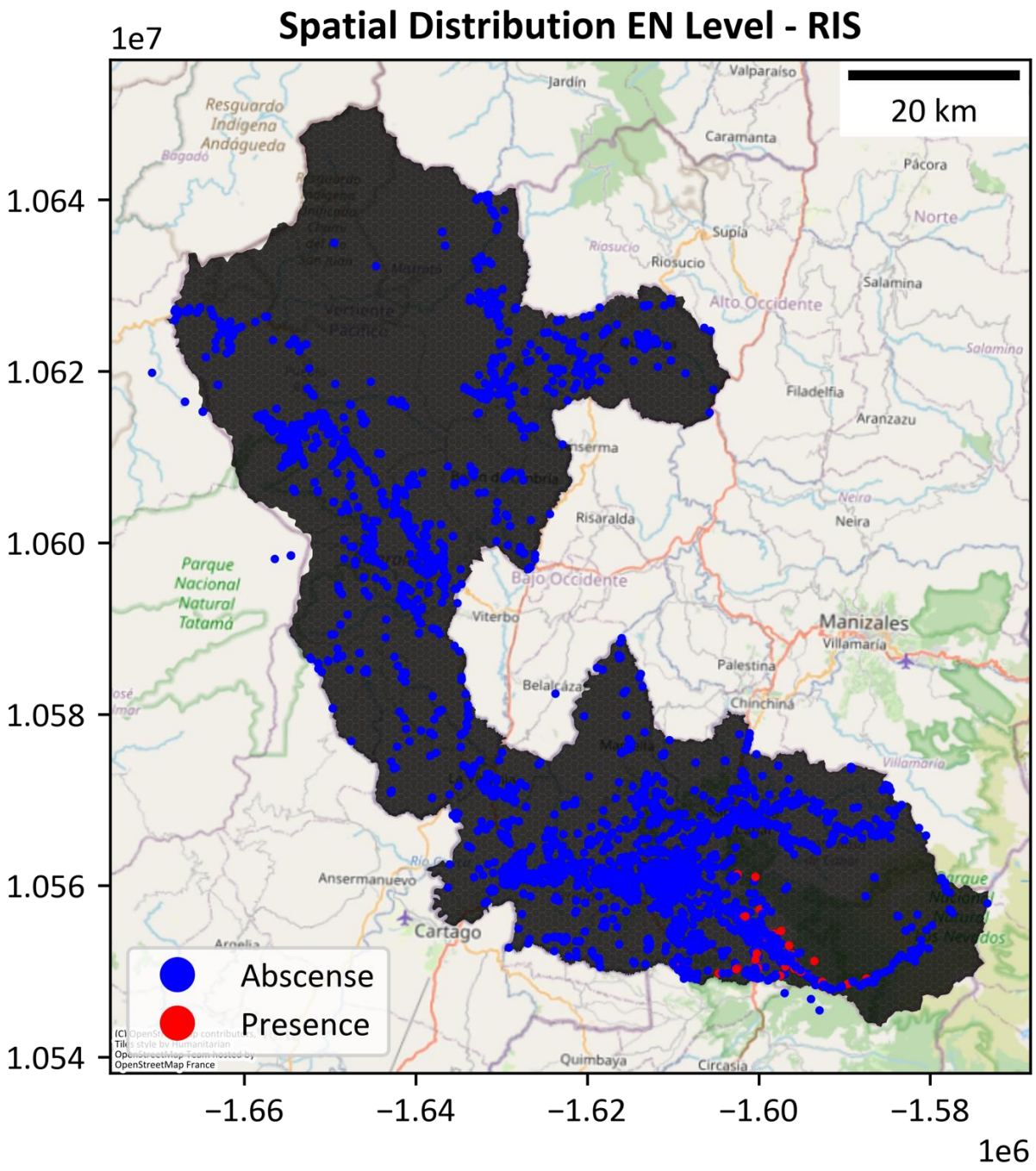


Figura 41. Distribución espacial y representación hexagonal de registros de *Penelope Perspicax* en Risaralda. Esta especie, a comparación de la del nivel CR, tiene más presencias hacia el sur del departamento, lo que tiene sentido teniendo en cuenta el tipo de nivel de amenaza en el que está categorizada.

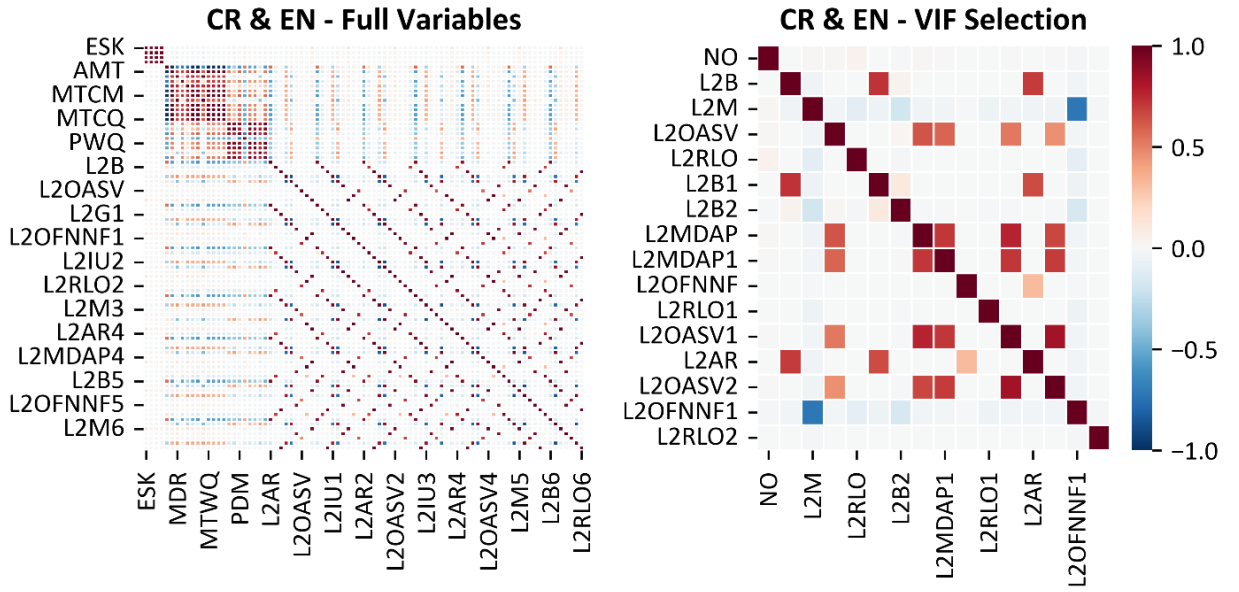


Figura 42. Matrices de correlación para el conjunto de características exógenas completas (Full Variables - Izquierda) y el conjunto de datos una vez se aplicó el VIF (Derecha) para el Departamento de Risaralda. Se evidencia que, si bien en el conjunto VIF existen algunas relaciones fuertes, la umbralización implicó preservar estas variables.

7.3. ANEXOS OBJETIVO 3

Modelo	precision	recall	f1-score	accuracy
Regresión Logistica (variables completas)	100%	100%	100%	100%
Random Forest (variables completas)	99%	100%	100%	100%
Regresión Logistica + Random oversampling (variables completas)	100%	97%	98%	97%
Random Forest + Random oversampling (variables completas)	100%	99%	99%	99%
Regresión Logistica + SMOTE (variables completas)	100%	97%	98%	97%
Random Forest + SMOTE (variables completas)	100%	99%	99%	99%
Regresión Logistica (solo variables representativas)	100%	100%	100%	100%
Random Forest (solo variables representativas)	99%	100%	100%	100%
Regresión Logistica + Random oversampling (solo variables representativas)	100%	98%	99%	98%
Random Forest + Random oversampling (solo variables representativas)	100%	96%	98%	96%
Regresión Logistica + SMOTE (solo variables representativas)	100%	97%	98%	97%
Random Forest + SMOTE (solo variables representativas)	100%	97%	98%	97%

Tabla 9. Resultados de evaluación de los modelos Regresión Logística y Random Forest para la predicción de avistamiento de aves para categoría de peligro CR.

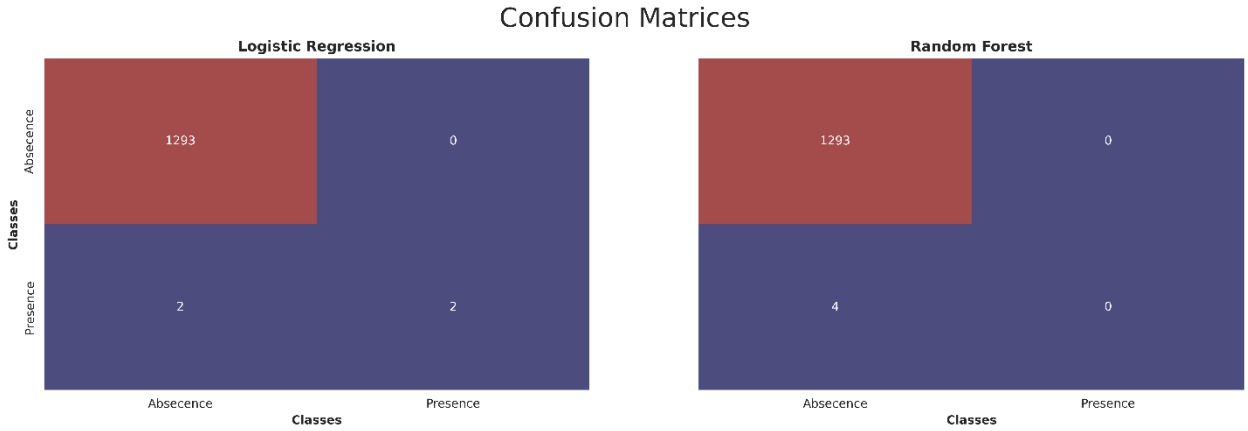


Figura 43. Matriz de confusión evaluación inicial de modelos regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría CR.

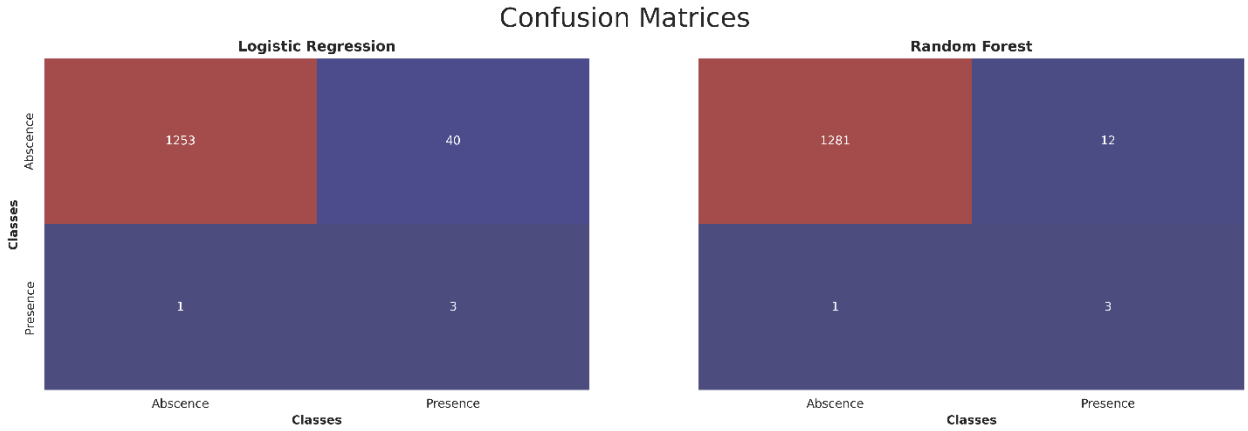


Figura 44. Matriz de confusión evaluación de modelos Regresión logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con todas las variables disponibles. Categoría CR.

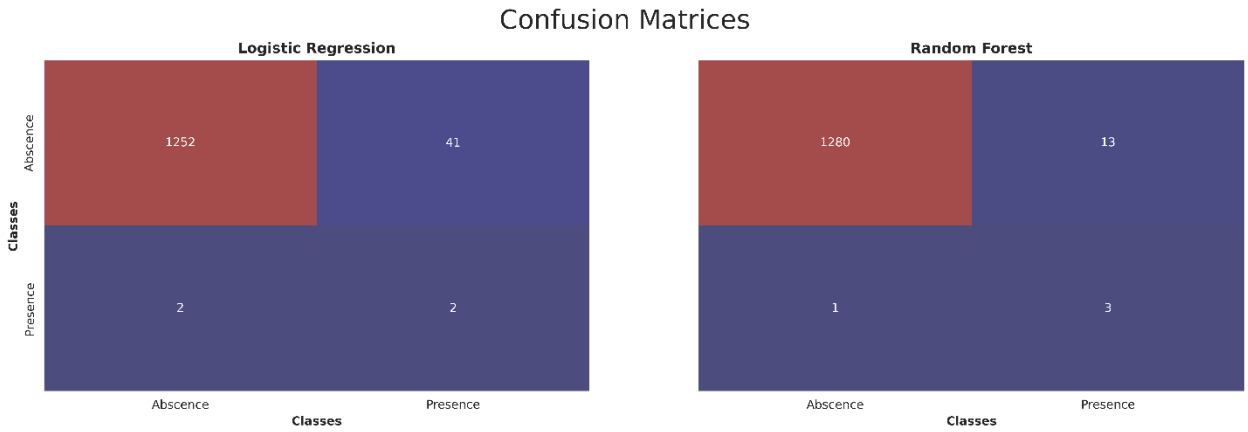


Figura 45. Matriz de confusión evaluación de modelos Regresión Logistica y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría CR.

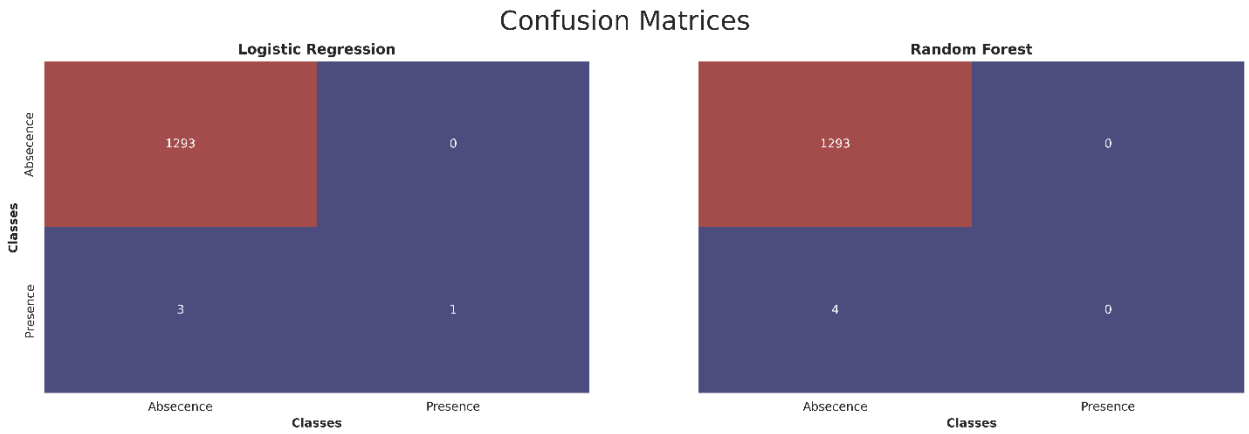


Figura 46. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría CR.

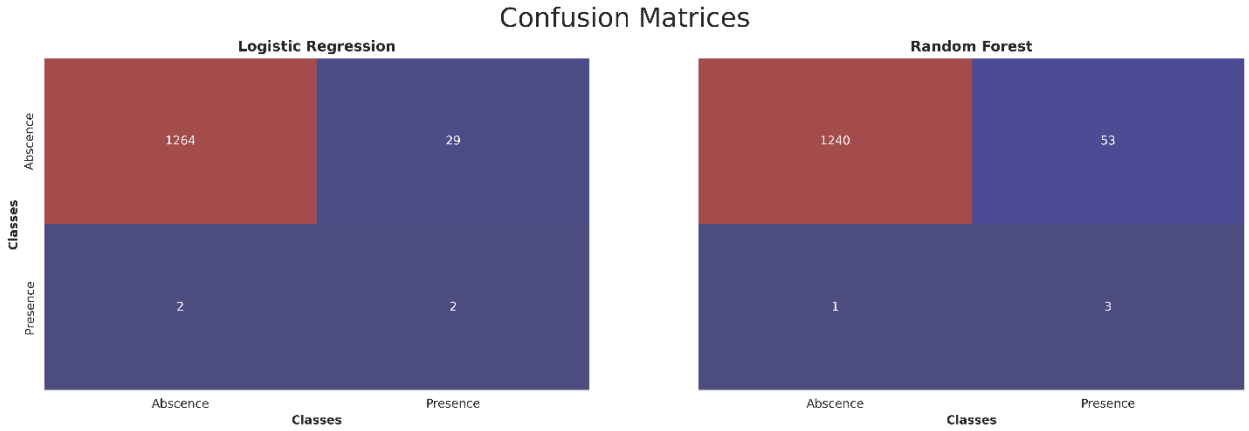


Figura 47. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas. Categoría CR.

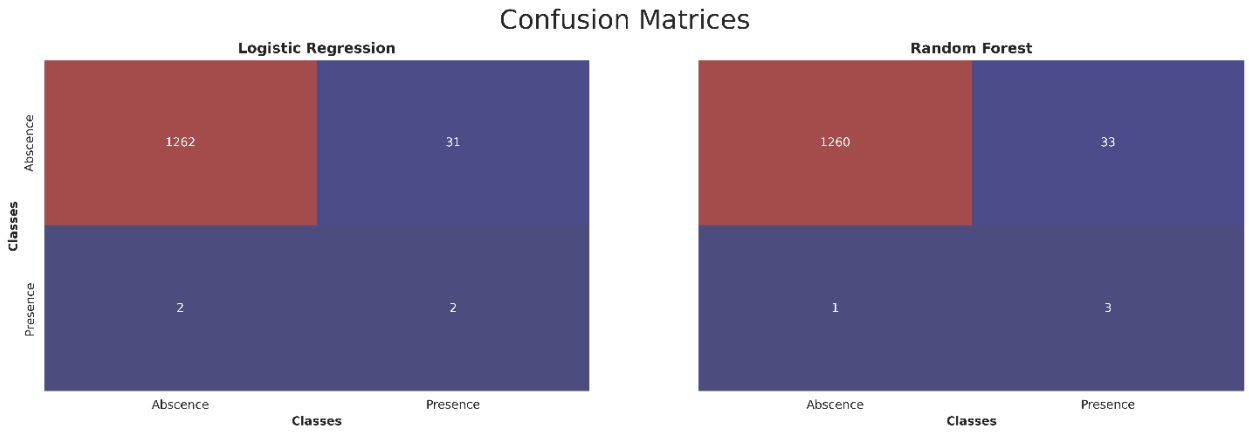


Figura 48. Matriz de confusión evaluación de modelos regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría CR.

Modelo	precision	recall	f1-score	accuracy
Regresión Logistica (variables completas)	99%	99%	99%	99%
Random Forest (variables completas)	99%	99%	99%	99%
Regresión Logistica + Random oversampling (variables completas)	99%	90%	94%	90%
Random Forest + Random oversampling (variables completas)	99%	98%	98%	98%
Regresión Logistica + SMOTE (variables completas)	99%	92%	95%	92%
Random Forest + SMOTE (variables completas)	99%	98%	98%	98%
Regresión Logistica (solo variables representativas)	99%	99%	99%	99%
Random Forest (solo variables representativas)	98%	99%	98%	99%
Regresión Logistica + Random oversampling (solo variables representativas)	99%	93%	96%	93%
Random Forest + Random oversampling (solo variables representativas)	99%	88%	93%	88%
Regresión Logistica + SMOTE (solo variables representativas)	99%	94%	96%	94%
Random Forest + SMOTE (solo variables representativas)	99%	95%	97%	95%

Tabla 10. Resultados de evaluación de los modelos Regresión Logística y Random Forest para la predicción de avistamiento de aves para categoría de peligro EN.

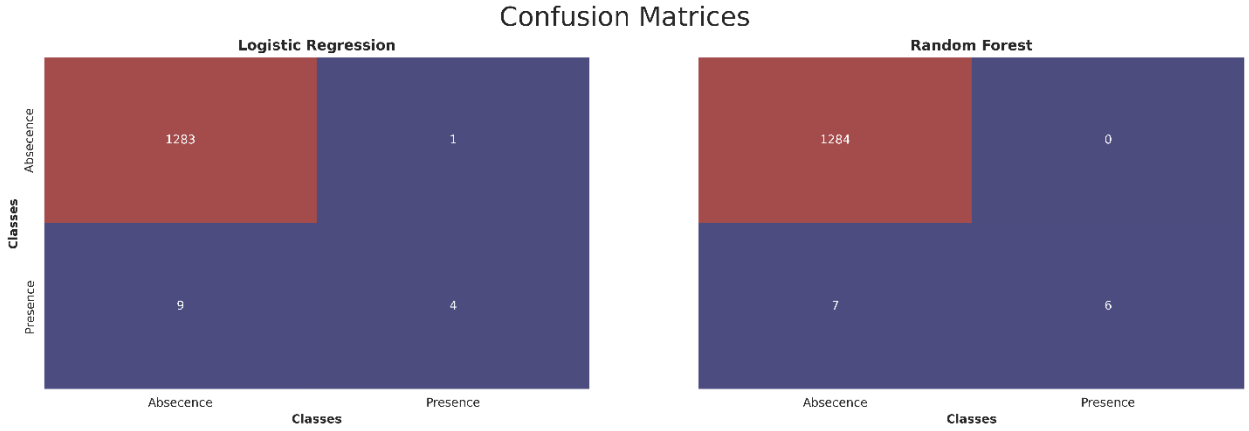


Figura 49. Matriz de confusión evaluación inicial de modelos Regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría EN.

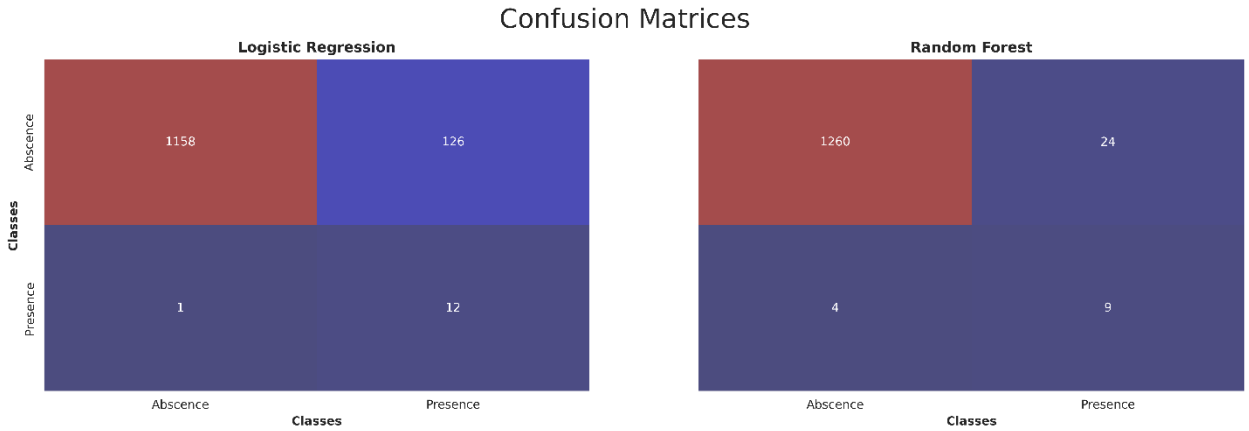


Figura 50. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con todas las variables disponibles. Categoría EN.

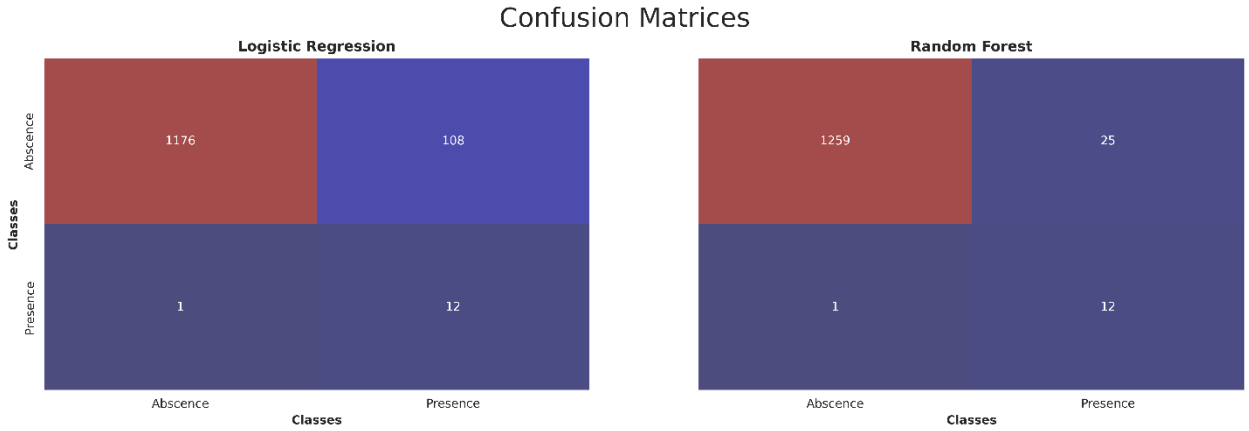


Figura 51. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría EN.

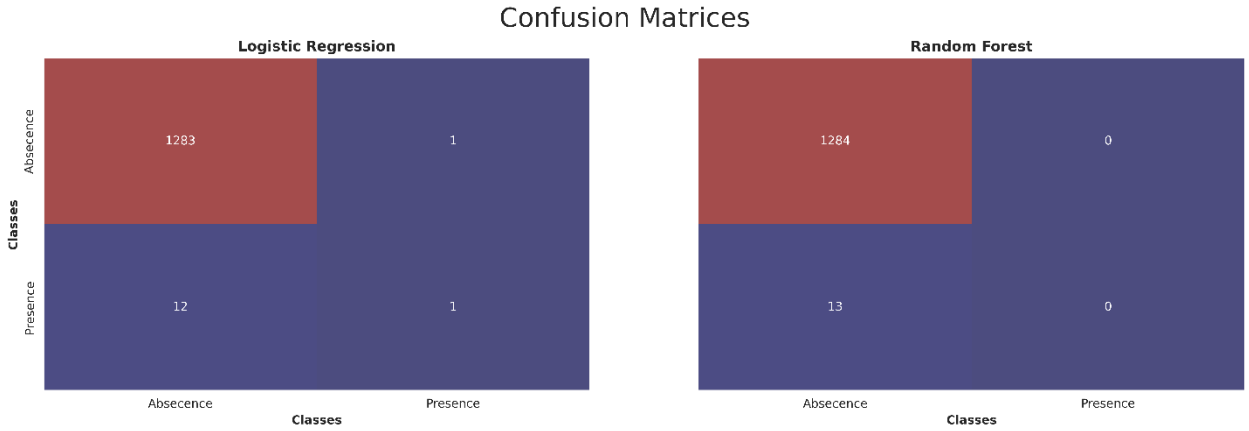


Figura 52. Matriz de confusión evaluación de modelos regresión Logistica y Random Forest sin realizar balanceo y con variables representativas. Categoría EN.

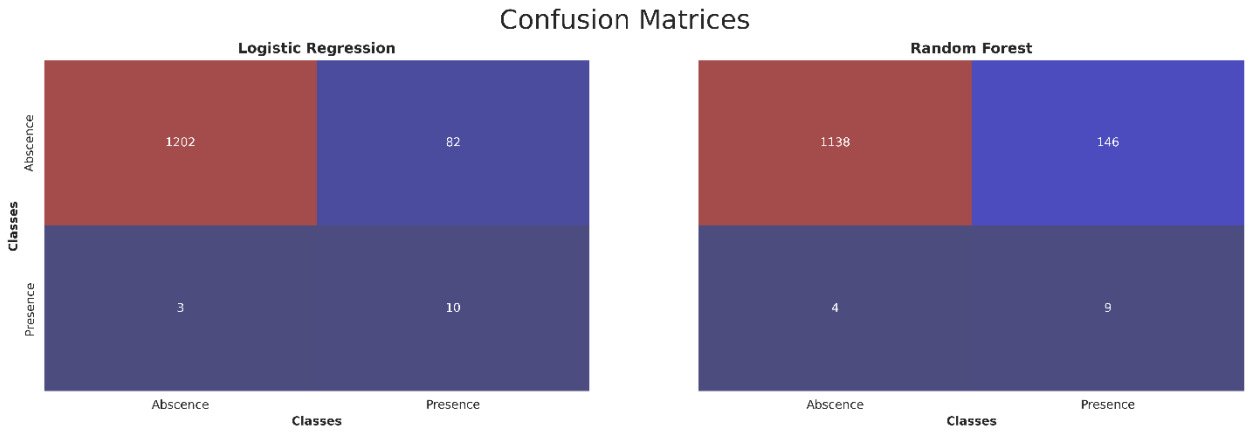


Figura 53. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas. Categoría EN.

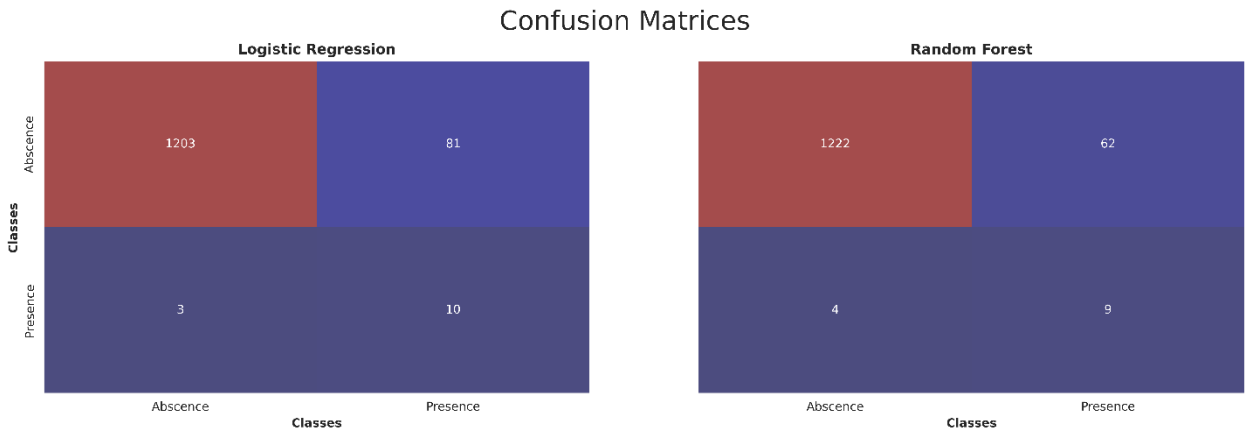


Figura 54. Matriz de confusión evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría EN.

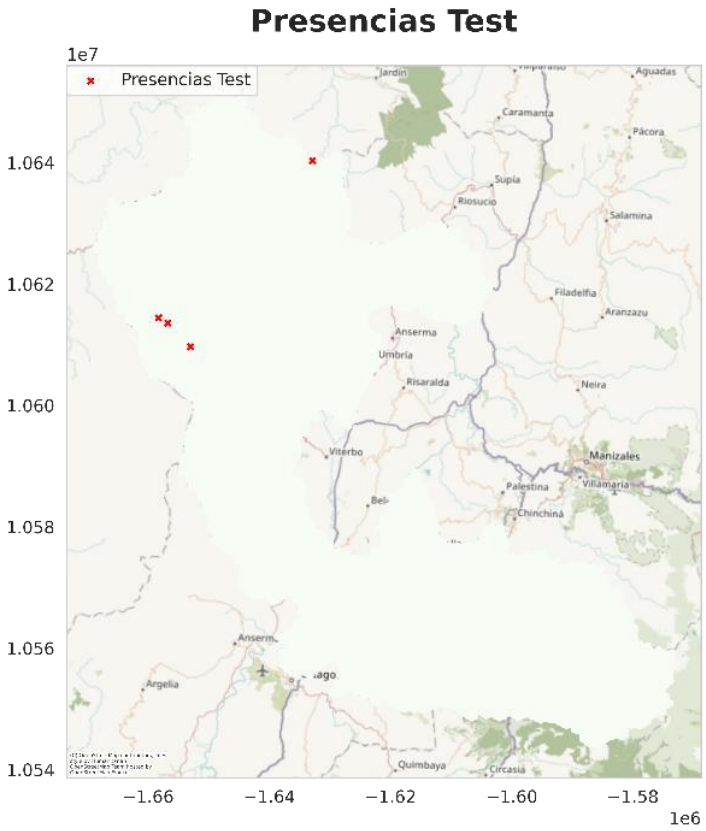


Figura 55. Mapa de presencias registradas en la evaluación del modelo. Categoría CR.

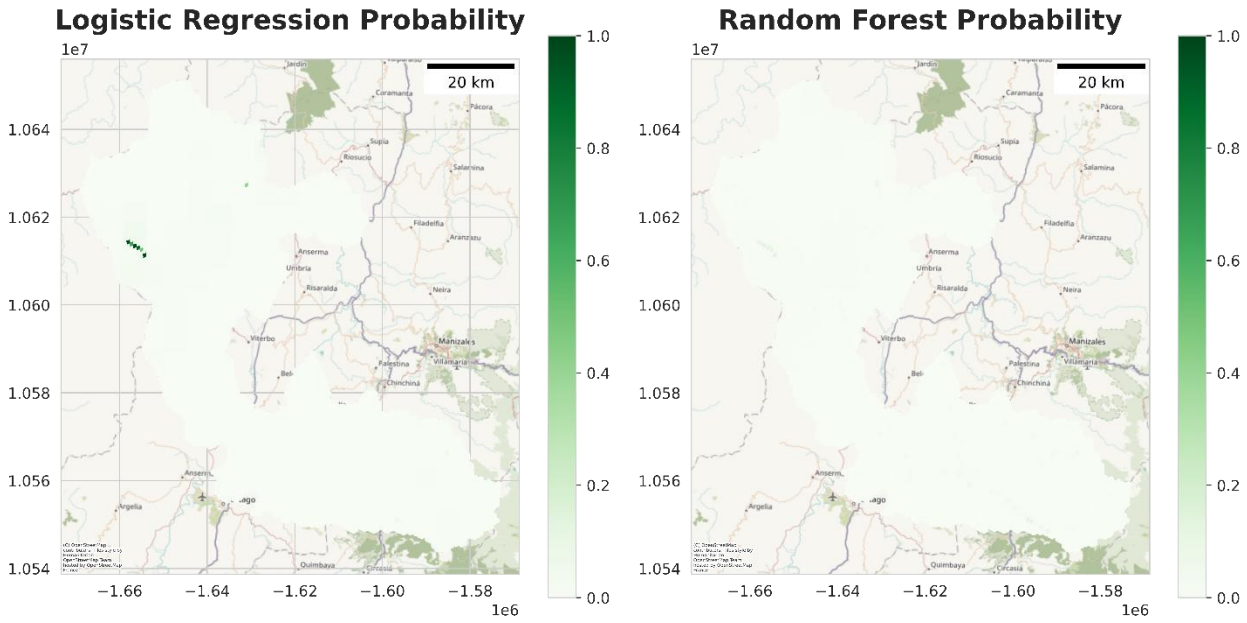


Figura 56. Mapa de probabilidad evaluación inicial de modelos Regresión Logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría CR.

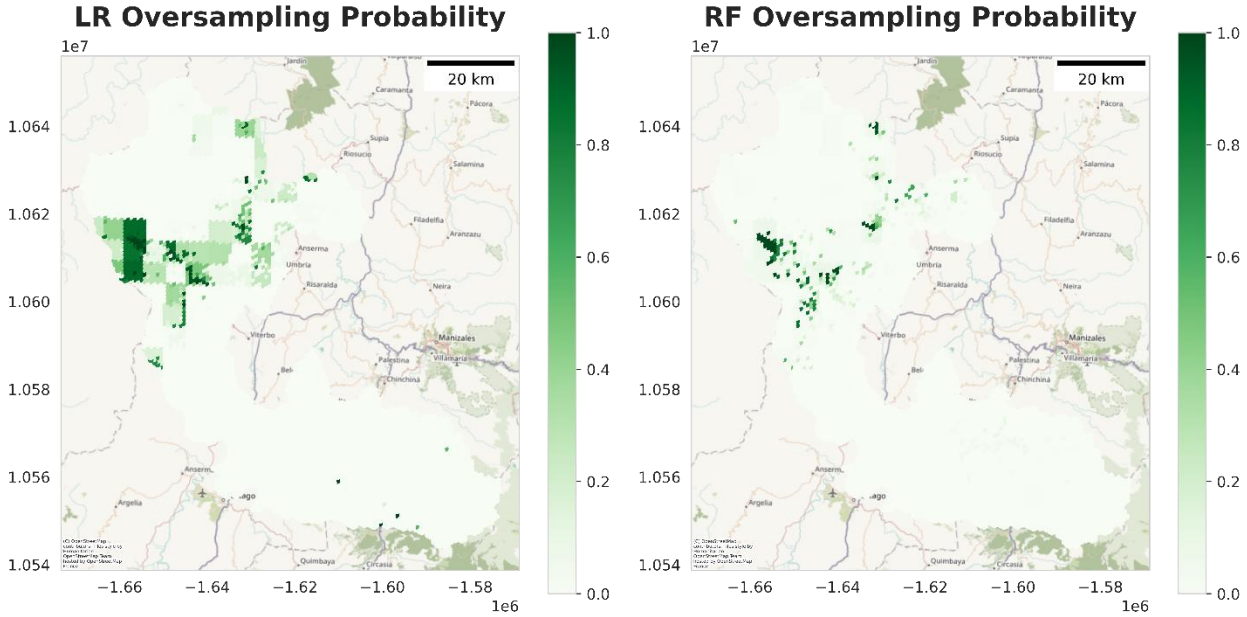


Figura 57. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con todas las variables disponibles. Categoría CR.

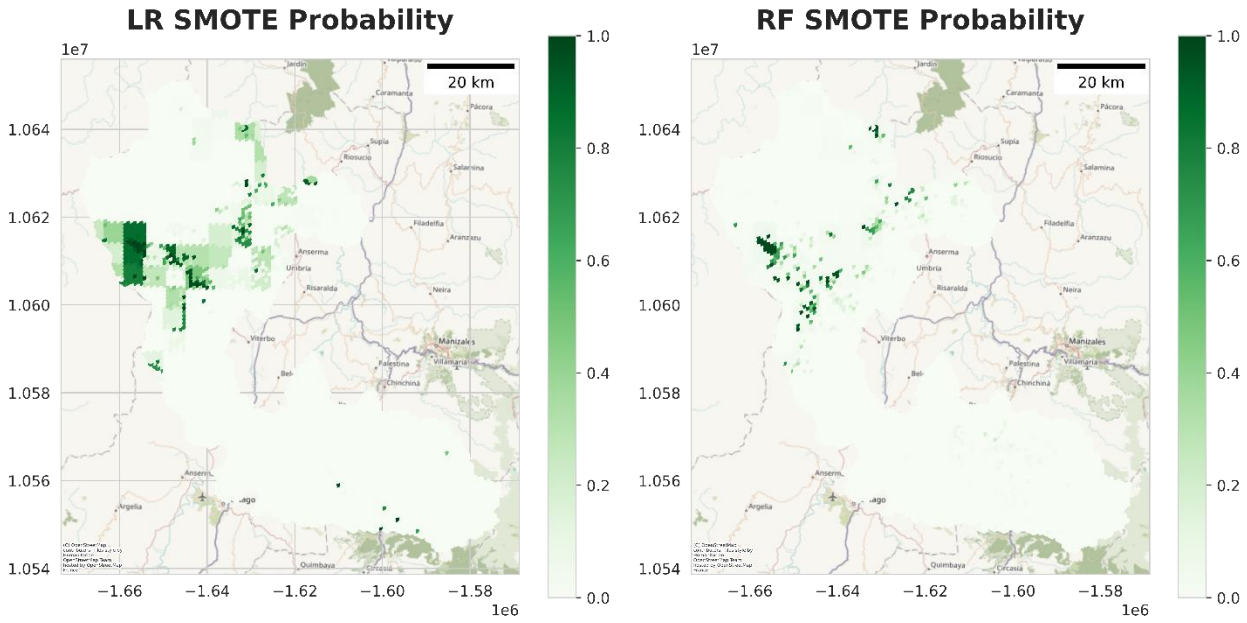


Figura 58. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría CR.

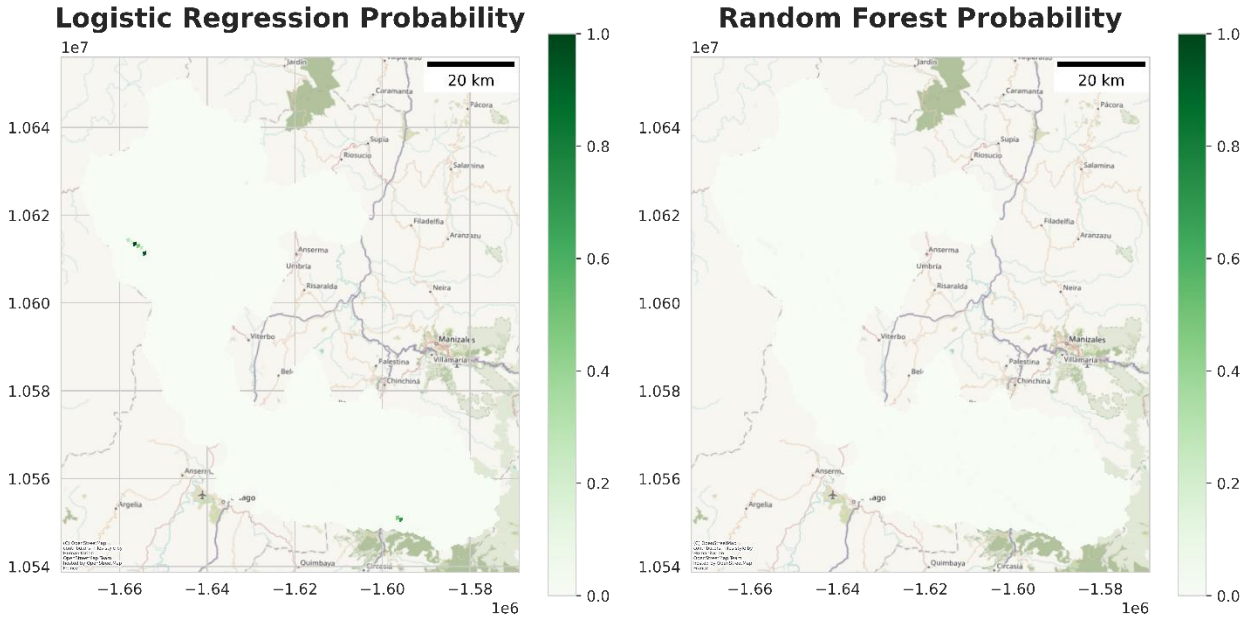


Figura 59. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría CR.

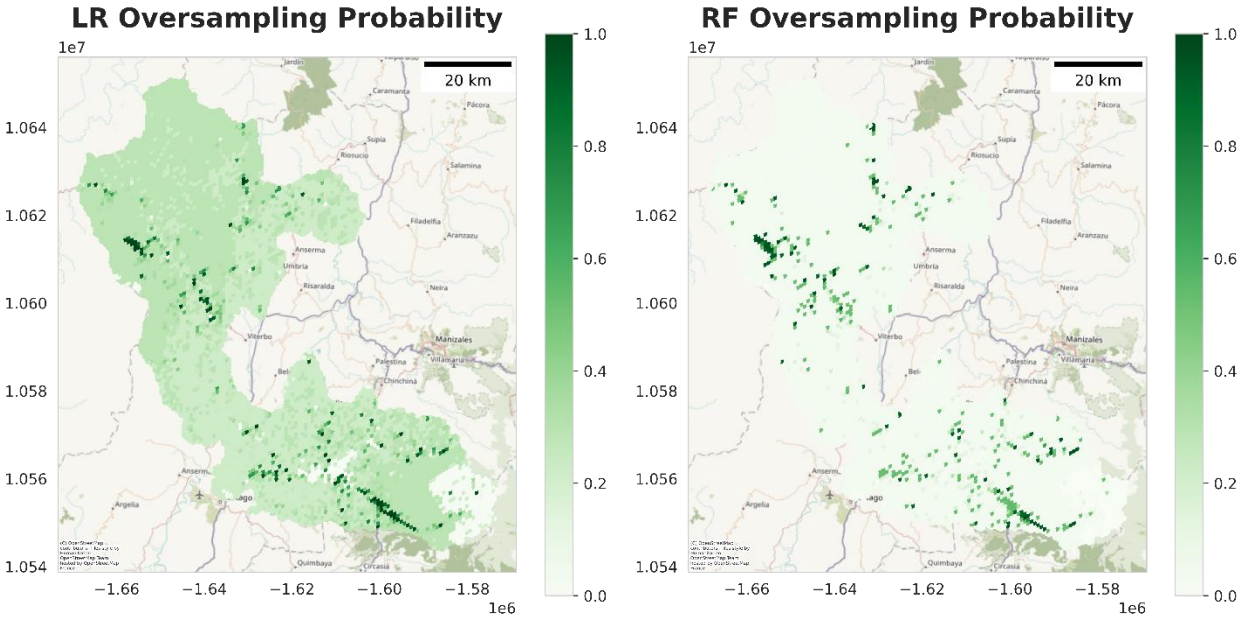


Figura 60. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con selección de variables representativas. Categoría CR.

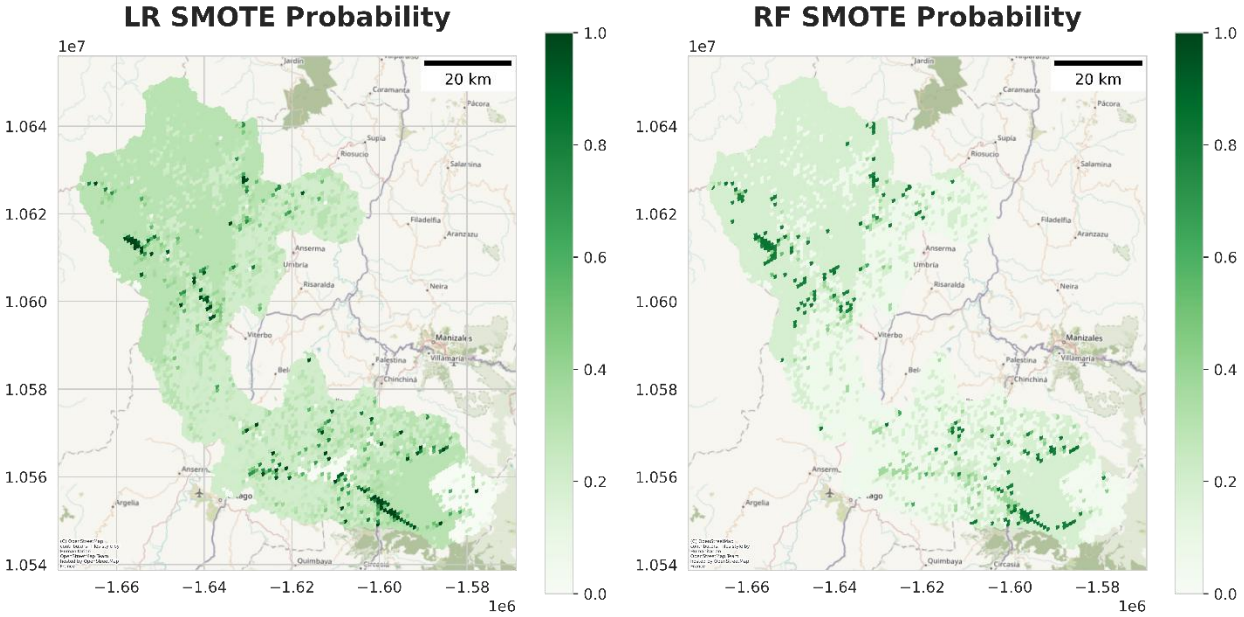


Figura 61. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría CR.

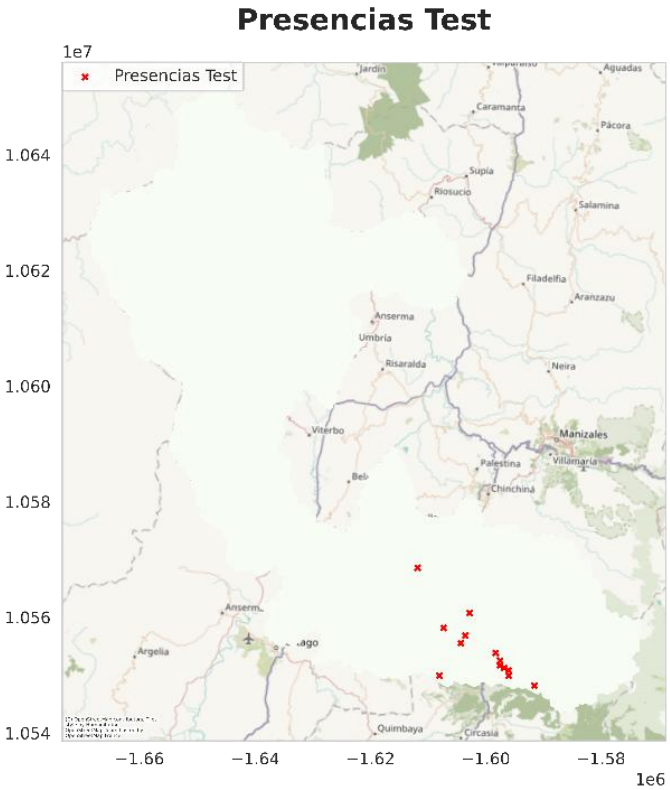


Figura 62. Mapa de presencias registradas en la evaluación del modelo. Categoría EN.

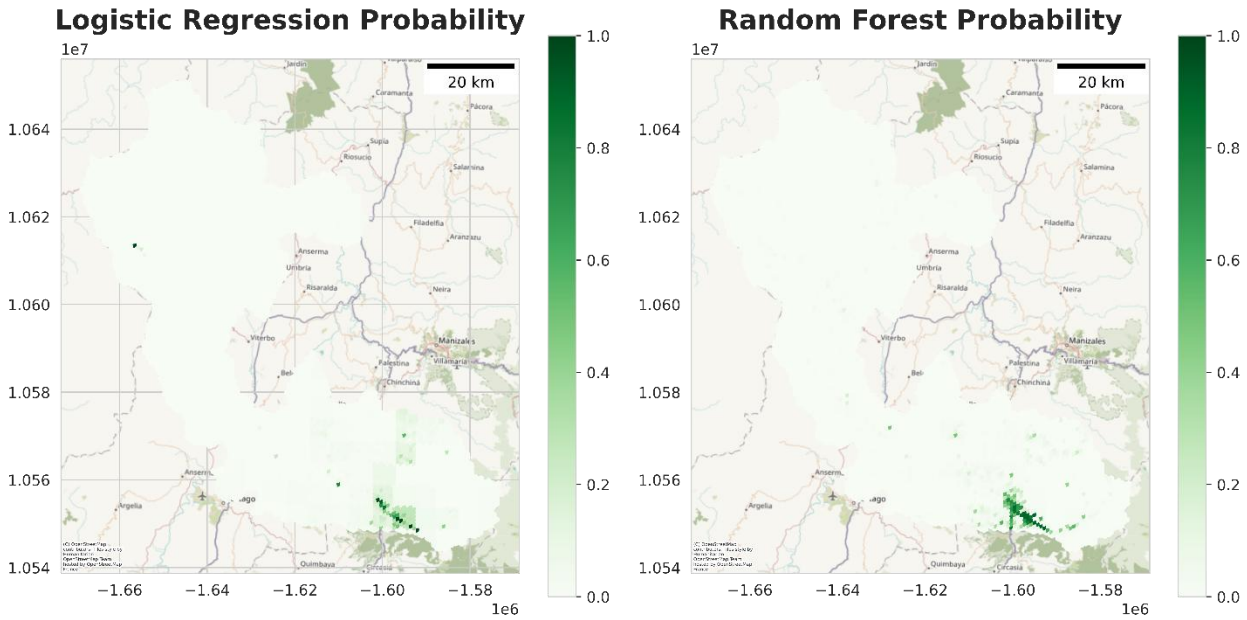


Figura 63. Mapa de probabilidad evaluación inicial de modelos Regresión logística y Random Forest sin realizar balanceo y con todas las variables disponibles. Categoría EN.

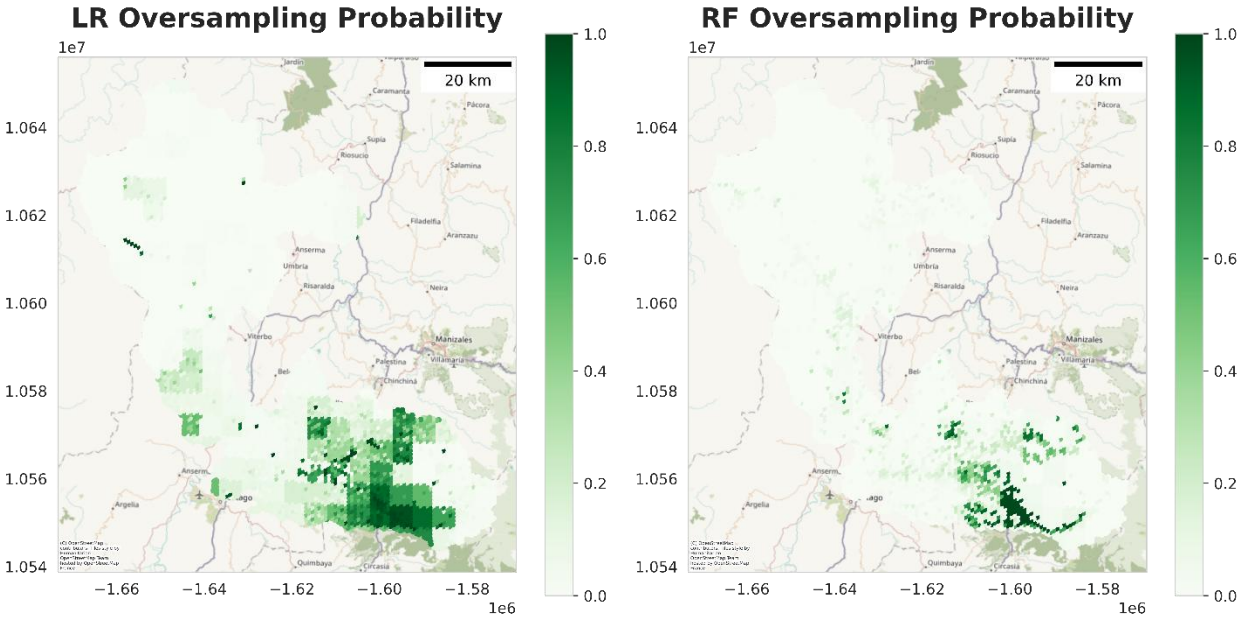


Figura 64. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con todas las variables disponibles. Categoría EN.

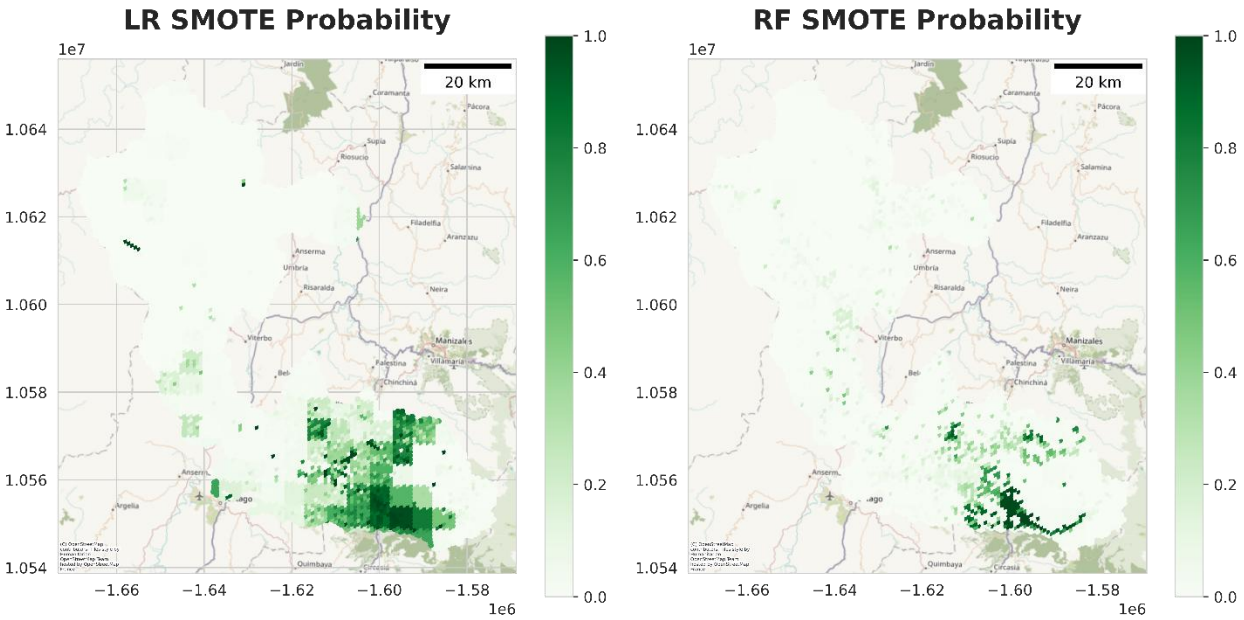


Figura 65. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con todas las variables disponibles. Categoría EN.

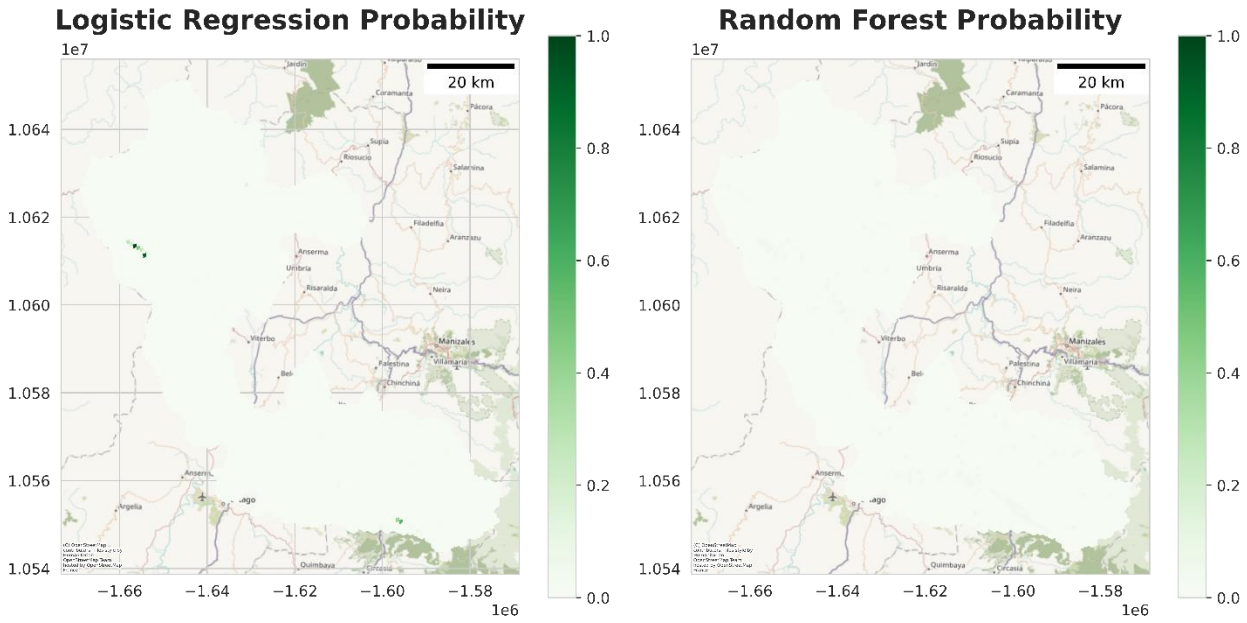


Figura 66. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest sin realizar balanceo y con variables representativas. Categoría EN.

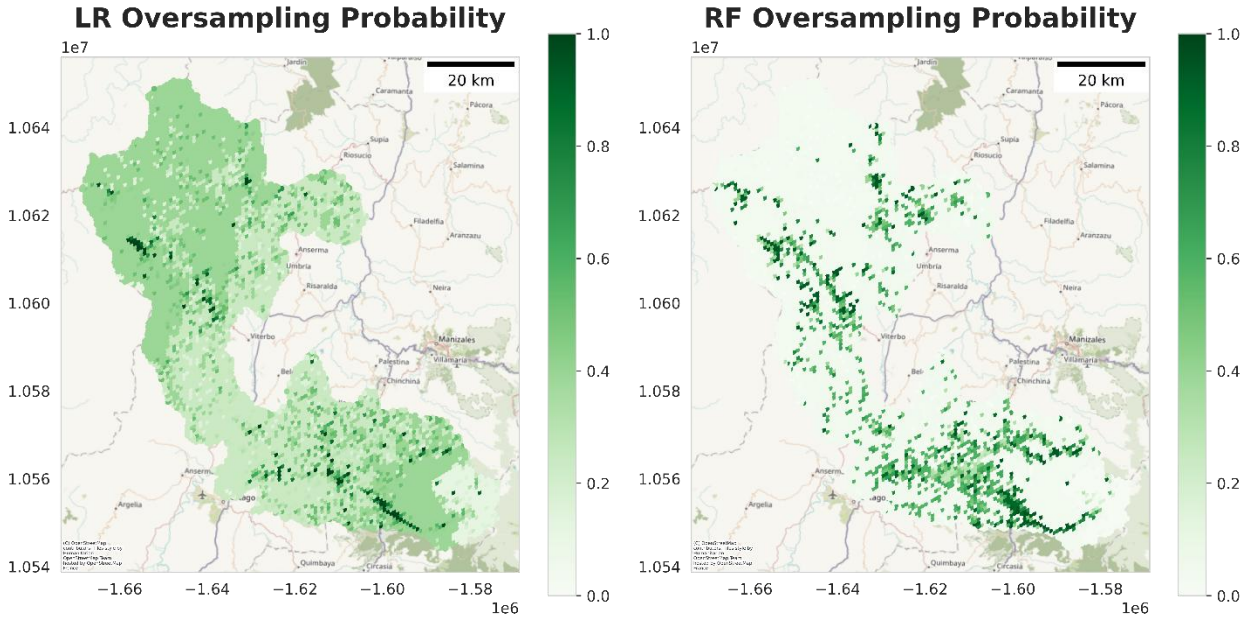


Figura 67. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica Random over sampling con selección de variables representativas. Categoría EN.

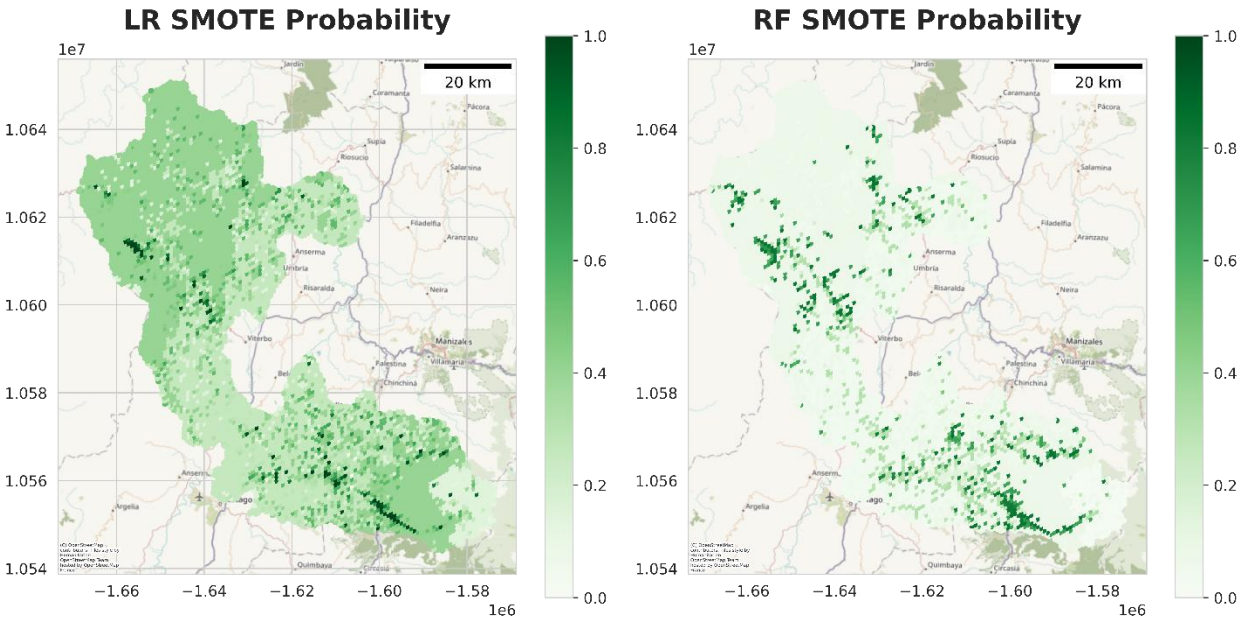


Figura 68. Mapa de probabilidad evaluación de modelos Regresión Logística y Random Forest posterior a realizar balanceo con técnica SMOTE con selección de variables representativas. Categoría EN.

Variable	MDI
number_observers	**0.223990**
lulc_2014_mineria	**0.007502**
lulc_2020_otra_formacion_natural_no_forestal	**0.002661**
lulc_2014_rio_lago_oceano	**0.001295**
lulc_2017_bosque	**0.000836**
lulc_2014_otra_area_sin_vegetacion	0.000079
lulc_2018_mosaico_de_agricultura_pasto	0.000054
lulc_2016_bosque	0.000049
lulc_2017_mosaico_de_agricultura_pasto	0.000042
lulc_2018_rio_lago_oceano	0.000040
lulc_2019_otra_area_sin_vegetacion	0.000039
lulc_2020_otra_area_sin_vegetacion	0.000033
lulc_2014_bosque	0.000001
lulc_2018_otra_formacion_natural_no_forestal	0.000000
lulc_2020_afloramiento_rocoso	0.000000
lulc_2020_rio_lago_oceano	0.000000

Tabla 11. Resultado de la importancia de las características del modelo Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas para categoría Peligro Critico (CR).

Variable	MDI
number_observers	**0.216984**
lulc_2014_mineria	**0.006381**

lulc_2020_otra_formacion_natural_no_forestal	**0.004196**
lulc_2014_otra_area_sin_vegetacion	**0.000958**
lulc_2014_rio_lago_oceano	**0.000707**
lulc_2017_bosque	0.000692
lulc_2017_mosaico_de_agricultura_pasto	0.000339
lulc_2018_mosaico_de_agricultura_pasto	0.000273
lulc_2020_otra_area_sin_vegetacion	0.000181
lulc_2019_otra_area_sin_vegetacion	0.000109
lulc_2016_bosque	0.000044
lulc_2018_rio_lago_oceano	0.000043
lulc_2020_afloramiento_rocoso	0.000003
lulc_2014_bosque	0.000000
lulc_2018_otra_formacion_natural_no_forestal	0.000000
lulc_2020_rio_lago_oceano	0.000000

Tabla 12. Resultado de la importancia de las características del modelo Random Forest posterior a realizar balanceo con técnica Random Over Sampling con selección de variables representativas para categoría En Peligro (EN).