MODELO PREDICTIVO PARA ESTIMAR EL CRECIMIENTO EN LA PUBLICACIÓN DE DATOS SOBRE BIODIVERSIDAD: UN ENFOQUE BASADO EN VARIABLES SOCIOECONÓMICAS Y DECISIONES GUBERNAMENTALES EN EL NODO GBIF COLOMBIA

Ricardo Ortiz Gallego
(Código 9013834)
Daniel Badillo Mojica
(Código 9015087)

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

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 19 DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

En el contexto global en el que la pérdida de biodiversidad compromete la estabilidad de los ecosistemas, disponer de datos abiertos y accesibles resulta esencial para diseñar estrategias de conservación más efectivas. La red GBIF desempeña un papel central en la movilización y publicación de datos abiertos sobre biodiversidad, apoyada por sus nodos nacionales. En el caso de Colombia, el Sistema de Información sobre Biodiversidad de Colombia (SiB Colombia), consolida un papel estratégico en la gestión y accesibilidad de datos biológicos. La ausencia de modelos predictivos limita la capacidad de definir metas de publicación más allá de la capacidad institucional, generando interrogantes sobre cómo los tomadores de decisiones pueden apoyar el avance hacia el cumplimiento de compromisos internacionales, como los establecidos en el Marco Mundial de Biodiversidad Kunming-Montreal, cuya meta 21 promueve la accesibilidad a datos, información y conocimiento para la toma de decisiones sobre la biodiversidad.

Ante este panorama, en este proyecto se propuso desarrollar un modelo predictivo para estimar el crecimiento en la publicación de datos sobre biodiversidad en el SiB Colombia bajo escenarios de cambio en variables socioeconómicas. Para ello, se realizó un análisis de información histórica y de acceso público, implementando y comparando distintos enfoques de modelamiento estadístico y de *machine learning*. El resultado fue la construcción de un modelo replicable, capaz de predecir el crecimiento futuro de los datos publicados en el nodo nacional, que no solo tiene en cuenta la inercia del crecimiento actual, sino el impacto de variables socioeconómicas y de gobernanza, lo cual podría servir como insumo para la toma de decisiones estratégicas y la formulación de políticas basadas en evidencia que contribuyan al fortalecimiento del SiB Colombia, y al cumplimiento de los compromisos globales de conservación y ciencia abierta.

El documento se organiza en ocho capítulos. El primero presenta el planteamiento del problema y las preguntas de investigación; el segundo expone los objetivos generales y específicos; y el tercero aborda el marco teórico y los antecedentes del estudio. A partir de este punto, se desarrollan las fases operativas del proyecto: la exploración de los datos (Capítulo 4); la preparación de datos y el diseño de los modelos estadísticos y de aprendizaje automático (Capítulo 5); la evaluación comparativa del desempeño de los modelos (Capítulo 6); y la implementación del modelo seleccionado para generar proyecciones estratégicas de crecimiento (Capítulo 7). Finalmente, se presentan las conclusiones generales, las implicaciones prácticas y las proyecciones futuras derivadas del estudio (Capítulo 8).

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

La evaluación de la Plataforma Intergubernamental Científico Normativa sobre Biodiversidad y Servicios Ecosistémicos (IPBES) estimó que alrededor del 25 por ciento de las especies de animales y plantas evaluados están amenazadas, lo que equivale a cerca de un millón de especies a nivel global que ya están en peligro de extinción, muchas en apenas decenios [1]. Ante este escenario, el Marco mundial Kunming-Montreal de la diversidad biológica (MGB-KM) plantea la necesidad de catalizar, facilitar e impulsar la acción urgente y transformadora de los gobiernos nacionales, departamentales y locales, con la participación de toda la sociedad, para invertir la pérdida de la biodiversidad [2]. Sin embargo, para que los gobiernos puedan tomar acciones, es necesario contar con datos e información que soporten el cumplimiento de sus Estrategias y Planes Nacionales sobre Biodiversidad (NBSAP).

Más de 60 países, entre ellos Colombia, hacen parte del Sistema Global de Información sobre Biodiversidad (GBIF) [3], una red creada en el año 2000, que busca facilitar la publicación, acceso y uso de los datos sobre biodiversidad de forma abierta (acceso libre). Esta es la red más grande de este tipo, y cuenta con nodos por país como el Sistema de Información sobre Biodiversidad de Colombia (SiB Colombia) [4]. No obstante, a pesar de los avances, a la fecha no existe una herramienta que permita identificar los factores gubernamentales que más influyen en el crecimiento de los datos compartidos en estos nodos, o que permita establecer y predecir futuras tendencias, lo que es particularmente relevante teniendo en cuenta que a partir de estos datos se calcularán y estimarán los indicadores de cumplimiento de las metas abordadas en el MGB-KM. La falta de una herramienta de este tipo impide una planificación estratégica efectiva para la generación y gestión de datos de biodiversidad a nivel de los nodos nacionales de GBIF.

Dada la importancia de contar con datos robustos para monitorear los cambios en ecosistemas y su biodiversidad, se planteó este proyecto, que desarrolla un modelo predictivo que permite estimar el crecimiento de los datos e información sobre biodiversidad disponibles en función de variables socioeconómicas, que son impactadas por las decisiones gubernamentales. Con estos resultados se espera apoyar la medición de indicadores de cumplimiento de compromisos internacionales como el Convenio de Diversidad Biológica y el MGB-KM, que tiene como meta al 2030 que “los mejores datos, información y conocimientos estén disponibles a los encargados de la toma de decisiones, los profesionales y el público, para que guíen una gobernanza eficaz y equitativa, una gestión integrada y participativa de la diversidad biológica” [2].

El planteamiento de este problema implica que existen múltiples factores que pueden explicar el crecimiento en los datos e información sobre biodiversidad y muchos de ellos están en el ámbito gubernamental, por ejemplo, el porcentaje del PIB dedicado a la investigación, la ausencia de políticas públicas de datos y ciencia abierta, normatividad ambiental enfocada en el libre acceso a los resultados de proyectos de investigación o licenciamiento ambiental, entre otros. Estas variables pueden explicar el comportamiento del crecimiento de los datos en las últimas décadas sirviendo a su vez como base para que los nodos de GBIF proyecten su crecimiento y se orienten las decisiones que se deben tomar para alcanzar las metas asociadas al aumento del volumen de datos sobre biodiversidad, que finalmente, son esenciales para construir un inventario de la biodiversidad y evaluar el estado de pérdida del mismo.

Dado lo anterior, resulta fundamental aplicar herramientas analíticas avanzadas para explorar y modelar el crecimiento de los datos sobre biodiversidad en los nodos nacionales de GBIF. Lo cual puede ser cubierto a partir del desarrollo de un modelo predictivo basado en técnicas de modelado estadístico y aprendizaje automático, sustentado en la recopilación, limpieza y transformación de diversas fuentes de datos (GBIF, gobiernos, organizaciones internacionales). Estas herramientas tienen el potencial de identificar las variables más relevantes, construir y evaluar modelos que expliquen el comportamiento del crecimiento de los nodos, y aplicarlo a un caso particular como el SiB Colombia. Este análisis buscó facilitar la comprensión de las dinámicas actuales y proyectar escenarios futuros que podrán orientar la toma de decisiones estratégicas basadas en evidencia.

Desde la perspectiva de la planificación estratégica, y empleando la metodología del Marco Lógico de la CEPAL [5], se identificó que la falta de un modelo predictivo para estimar el crecimiento de la publicación de datos sobre biodiversidad en los nodos nacionales de GBIF limita significativamente la capacidad de gestionar de manera efectiva los datos de biodiversidad. La ausencia de un enfoque estructurado que permita identificar variables clave, como la inversión en investigación, las políticas públicas de ciencia abierta o la normatividad ambiental, representa un desafío para que los nodos de GBIF contribuyan al cumplimiento de compromisos internacionales como el MGB-KM. Esta situación compromete la capacidad de los países para generar y gestionar datos críticos que permitan monitorear la biodiversidad, evaluar el progreso hacia los objetivos globales y garantizar una gobernanza efectiva y equitativa de la diversidad biológica.

1.2. FORMULACIÓN DEL PROBLEMA

Pregunta de investigación

¿Cómo predecir el crecimiento en la publicación de datos sobre biodiversidad en los nodos nacionales de GBIF y específicamente del SiB Colombia, mediante técnicas de modelado estadístico y de machine learning, considerando variables socioeconómicas y de políticas públicas derivadas de decisiones gubernamentales?

Preguntas de sistematización

- ¿Cuáles son las variables que explican en mayor medida el crecimiento de la publicación de datos sobre biodiversidad en el nodo SiB Colombia?

- ¿Cuáles son los métodos de modelado más efectivos para predecir el crecimiento en la publicación de datos de biodiversidad?

- ¿Cómo se pueden utilizar las proyecciones del modelo para apoyar el cumplimiento de los compromisos internacionales, como el Convenio de Diversidad Biológica?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Desarrollar un modelo predictivo que permita estimar el crecimiento de la publicación de datos sobre biodiversidad en los nodos nacionales de GBIF, específicamente en el nodo GBIF Colombia (SiB Colombia) mediante el análisis de variables socioeconómicas y decisiones gubernamentales, con el fin de optimizar la planificación estratégica y la gestión de datos de biodiversidad.

2.2. OBJETIVOS ESPECÍFICOS

- Identificar las variables socioeconómicas y gubernamentales que afectan el crecimiento en la generación y publicación de datos sobre biodiversidad en el nodo nacional SiB Colombia, a partir del análisis de datos históricos y de acceso público.
- Implementar diferentes métodos de modelado estadístico y de machine learning para predecir el crecimiento en la publicación de datos sobre biodiversidad en el nodo nacional de GBIF (SiB Colombia), configurando cada modelo con base en los datos disponibles y las necesidades del problema.
- Evaluar el desempeño de los modelos implementados mediante métricas de validación y comparación de resultados, con el propósito de identificar el enfoque más efectivo para la predicción.
- Aplicar el modelo predictivo para proyectar el crecimiento futuro de los datos de biodiversidad publicados en el nodo nacional de GBIF (SiB Colombia), apoyando el cumplimiento de compromisos internacionales, y mejorar la planificación estratégica en la gestión de la biodiversidad.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1. Ciencia abierta y nodos de GBIF

La ciencia abierta ha transformado la forma en que se recopilan, comparten y utilizan los datos sobre biodiversidad, impulsando una colaboración científica más amplia y eficiente a nivel global [6]. Esta se basa en la aplicación de prácticas abiertas y cooperativas a lo largo de todo el proceso de generación y divulgación del conocimiento, promoviendo la transparencia, la accesibilidad y la reutilización de los datos [7]. Este enfoque ha permitido maximizar el impacto científico y social de la investigación al favorecer la participación de múltiples actores.

En este marco, el Sistema Global de Información sobre Biodiversidad (GBIF) se establece como una infraestructura para el acceso y uso de datos abiertos de biodiversidad, fortaleciendo tanto la investigación científica como la toma de decisiones informadas en conservación [8]. Los nodos nacionales del GBIF (Fig. 1), como el SiB Colombia, desempeñan un papel clave dentro de esta red global al gestionar la estandarización y publicación de datos en el contexto nacional. El SiB Colombia integra información proveniente de múltiples fuentes como colecciones biológicas, proyectos de investigación y programas de monitoreo ambiental [9], de forma se aporte con datos relevantes para la formulación de políticas públicas y el cumplimiento de compromisos internacionales, como el Convenio sobre la Diversidad Biológica (CDB).

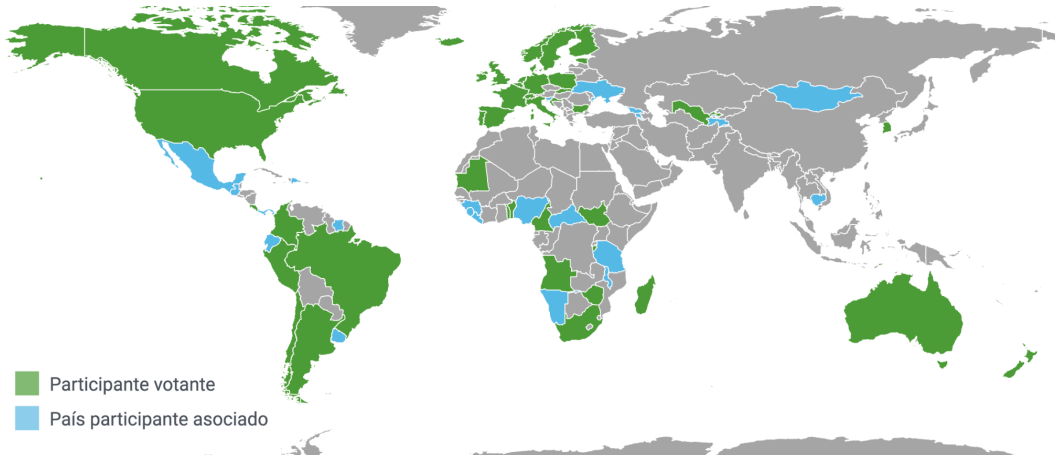


Figura 1. Distribución de los países miembros de la red de GBIF a nivel Global. Fuente: GBIF, 2024

La sostenibilidad y expansión del SiB Colombia dependen en gran medida de factores como la inversión gubernamental, el fortalecimiento de capacidades técnicas y la consolidación de colaboraciones interinstitucionales. Un fortalecimiento estratégico de este nodo permitirá continuar posicionándolo como un referente dentro de la red GBIF, contribuyendo a una mayor equidad en la disponibilidad y uso de datos para estrategias locales, nacionales y globales de conservación.

3.1.2. Convenios sobre la diversidad biológica e indicadores de biodiversidad

A escala global, los indicadores de biodiversidad se emplean principalmente para detectar y monitorear tendencias generales de la biodiversidad [10], identificar patrones en los datos e investigaciones generadas sobre biodiversidad [11], [12], y sensibilizar tanto al público como a los responsables políticos [13]. Sin embargo, estos indicadores siguen estando subutilizados en la toma de decisiones activas, como la evaluación de acciones pasadas de gestión o políticas, la orientación de decisiones políticas o el establecimiento de metas de biodiversidad [14].

El seguimiento del cumplimiento de convenios de diversidad biológica, como el MGB-KM [2], requiere un conjunto de indicadores predictivos fácilmente monitoreables construidos a partir de modelos explicativos [15]. Estos modelos permiten establecer relaciones claras entre los impulsores de la biodiversidad, como el cambio en el uso del suelo, los impactos del cambio climático, y otros motores de cambio como la gestión o políticas. Los indicadores predictivos no solo ofrecen herramientas para monitorear el progreso actual, sino que también anticipan escenarios futuros. Esto los convierte en instrumentos clave para guiar la planificación proactiva y la implementación de acciones estratégicas.

Los convenios internacionales, como el Marco Mundial Kunming-Montreal de la Diversidad Biológica (MGB-KM) y otros del Convenio sobre la Diversidad Biológica (CDB), tienen como objetivo principal promover la conservación de la biodiversidad, garantizar su uso sostenible y asegurar una distribución justa y equitativa de los beneficios derivados de los recursos naturales. Estas iniciativas abordan la crisis global de pérdida de biodiversidad a través de acciones integradas, transformadoras y basadas en compromisos multilaterales, buscando catalizar, facilitar e impulsar medidas urgentes y coordinadas por parte de los gobiernos nacionales,

subnacionales y locales, con la participación activa de toda la sociedad [2]. A pesar de los avances logrados con estos acuerdos [1], [16], en la práctica han surgido desafíos significativos, especialmente en la medición del progreso de sus objetivos debido a la falta de enfoque y especificidad en los indicadores utilizados [14]. Por ejemplo, el Índice de la Lista Roja (RLI), asignado para monitorear diversas Metas de Aichi bajo el marco del Global Biodiversity Framework (GBF), evidenció limitaciones en su capacidad para evaluar de manera efectiva las acciones implementadas. Este incumplimiento de las Metas de Aichi subrayó la necesidad de un enfoque renovado, llevando a la formulación del Marco Mundial de Biodiversidad Post-2020 (GBF) [17]. El GBF establece un marco ambicioso para "doblar la curva" de pérdida de biodiversidad mediante acciones inmediatas y transformadoras, con objetivos intermedios para 2030 y de largo plazo para 2050 [2].

3.1.3. Ciencia de datos y su integración en la gestión de biodiversidad

La ciencia de datos ofrece un conjunto de herramientas y metodologías que están revolucionando la investigación en biodiversidad, al potenciar la conexión entre las ciencias computacionales, las matemáticas y la biología. Estas herramientas permiten acelerar los procesos de captura, análisis e interpretación de información, ampliando las posibilidades para comprender los sistemas ecológicos [18]. Las investigaciones recientes demuestran nuevas aplicaciones en el monitoreo de poblaciones, el seguimiento de la biodiversidad a través del tiempo y el espacio, la planificación de la conservación y la protección de los ecosistemas mediante enfoques de aprendizaje profundo (*deep learning*) y el uso de gemelos digitales, con alta relevancia para la sostenibilidad y la transición ecológica.

El repositorio de datos GBIF, la red más grande de datos abiertos sobre biodiversidad, ha crecido más de un 1.000 % en la última década, reflejando el impacto de la digitalización de colecciones y la integración de grandes volúmenes de información biológica [10], [19]. Este proceso ha impulsado una nueva era de ciencia de datos intensiva, que combina información proveniente de múltiples campos como la biología evolutiva, la ecología, la conservación, la salud ambiental y las políticas públicas y fuentes como los datos derivados de ADN, cámaras trampa, monitoreo acústico, entre otros [19], [20]. Este enfoque permite construir teorías más integrales y aplicar modelos predictivos e inteligencia artificial en la evaluación de riesgos, la planificación de estrategias de conservación y la toma de decisiones basadas en evidencia sobre biodiversidad [19].

3.1.4. Modelos predictivos y técnicas de análisis aplicadas a la biodiversidad

Dentro del campo de la ciencia de datos, los modelos predictivos constituyen una herramienta esencial para anticipar comportamientos futuros a partir de patrones observados en datos históricos. En el contexto de la biodiversidad, estos modelos permiten identificar tendencias en la generación, publicación y disponibilidad de información, proporcionando bases cuantitativas para orientar decisiones estratégicas en conservación y gestión ambiental [21]. Entre las principales técnicas utilizadas se encuentran los modelos estadísticos tradicionales, como la *Regresión Lineal Múltiple* (RLM), que permite estimar relaciones entre una variable dependiente continua y múltiples variables explicativas, ofreciendo una primera aproximación para identificar factores clave que explican el crecimiento en la publicación de datos [22]. Por su parte, los modelos de series temporales ARIMA (*Autoregressive Integrated Moving Average*) resultan adecuados para analizar el comportamiento temporal de las publicaciones, capturando tendencias y patrones cíclicos a partir de datos históricos. A medida que los conjuntos de datos se vuelven más complejos y heterogéneos, los métodos de aprendizaje automático (machine learning) ofrecen alternativas más flexibles y precisas.

Entre ellos, los Bosques Aleatorios (*Random Forest*) [23] y XGBoost (*Extreme Gradient Boosting*) [24] se destacan por su capacidad para manejar relaciones no lineales, grandes volúmenes de variables y datos con estructuras mixtas. Estos modelos resultan especialmente útiles para analizar interacciones entre factores socioeconómicos, institucionales y ambientales, facilitando la identificación de las variables con mayor influencia en el crecimiento de los datos publicados.No obstante, la naturaleza de "caja negra" de estos algoritmos puede dificultar la comprensión detallada de las relaciones funcionales entre los predictores. Ante esta limitación, los Modelos Aditivos Generalizados (GAM) emergen como una solución intermedia que combina la flexibilidad del aprendizaje automático con la interpretabilidad de los modelos lineales [25]. Los GAM permiten modelar efectos no lineales mediante funciones de suavizado, lo que facilita visualizar y explicar cómo cada decisión gubernamental o variable socioeconómica impacta específicamente en la tendencia de publicación de datos, una característica esencial para la toma de decisiones estratégicas basadas en evidencia.

Por otra parte, los modelos de redes neuronales recurrentes con arquitectura *Long Short-Term Memory* (LSTM) son considerados una solución para capturar dependencias de largo plazo en secuencias, a partir del uso de "puertas" (gates) para gestionar la memoria de sucesos pasados [26]. En este sentido, estos modelos cuentan con capacidad para predecir secuencias de eventos futuros, marcas de tiempo y recursos asociados, lo cual es directamente aplicable a la lógica de

un proyecto de predicción de datos sobre biodiversidad a través del tiempo.

En los últimos años, los avances en aprendizaje por refuerzo (*Reinforcement Learning*, RL) han ampliado las posibilidades del modelado predictivo hacia escenarios dinámicos, donde las decisiones y sus consecuencias se retroalimentan de manera continua. A diferencia de los métodos supervisados, el RL aprende mediante la interacción entre un agente y su entorno, optimizando sus acciones a partir de recompensas acumuladas. Este enfoque es especialmente valioso para modelar sistemas ecológicos complejos y procesos de gestión adaptativa, en los cuales las decisiones de conservación modifican las condiciones futuras del sistema [27].

Entre las variantes más recientes, el algoritmo *Twin Delayed Deep Deterministic Policy Gradient* (TD3) [28] combina redes neuronales profundas con principios del aprendizaje por refuerzo continuo, mejorando la estabilidad y la precisión del entrenamiento en entornos con alta incertidumbre y múltiples variables interdependientes. Su potencial para la sostenibilidad ambiental y la proyección de escenarios ecológicos complejos abre nuevas oportunidades para la modelización de la biodiversidad y la evaluación de políticas ambientales [29].

3.2. ANTECEDENTES

Durante la primera parte de la revisión bibliográfica, se identificaron trabajos que destacan la relevancia de los indicadores para el seguimiento de las tendencias de la biodiversidad y la evaluación de las medidas destinadas a detener su pérdida, particularmente en el contexto de los convenios sobre la diversidad biológica. Leadley et al. (2022) [15] enfatizan la necesidad de fortalecer los marcos de monitoreo mediante estrategias integradas, como la incorporación de sistemas de detección y atribución que identifiquen los impactos de los impulsores de cambio en la biodiversidad, la integración y desagregación de datos a nivel nacional y global para evaluar el progreso de los países, y el desarrollo de indicadores predictivos que guíen acciones proactivas. Estas capacidades no solo permitirían rastrear el progreso, sino también respaldar políticas y acciones adaptativas esenciales para cumplir con los objetivos de los convenios internacionales. Por su parte, Stevenson et al. (2021) [14] abordan la falta de indicadores diseñados específicamente para la toma de decisiones, proponiendo un modelo conceptual que define el papel activo de los indicadores dentro del ciclo de políticas y su conexión explícita con la acción preventiva. Este enfoque es especialmente relevante en el marco de los convenios sobre diversidad biológica, ya que permite garantizar que los indicadores asociados sean efectivos para revisar y cumplir los objetivos establecidos, facilitando una gestión más informada y proactiva.

Los trabajos de Leadley et al. (2022) [15] y Stevenson et al. (2021) [14] ofrecen valiosos aportes al marco conceptual del presente proyecto. Leadley et al. refuerzan la importancia de conectar los datos de biodiversidad con herramientas predictivas adaptativas, aportando un enfoque crítico que el proyecto aplica al contexto del SiB Colombia, con el objetivo de proyectar el crecimiento de datos mediante variables específicas del entorno colombiano. Stevenson et al., por otro lado, proporcionan un marco conceptual para vincular los indicadores con decisiones estratégicas, lo que complementa el enfoque del proyecto al incorporar estos indicadores como insumos clave para modelar y proyectar tendencias en la gestión de datos de biodiversidad. De esta forma, el presente proyecto integrará estas perspectivas para avanzar hacia una implementación práctica que conecta compromisos internacionales con herramientas analíticas específicas para un nodo nacional.

Por otro lado, artículos como Zuccotto et al. (2024) [27] y Lapeyrolerie et al. (2022) [29] presentan aportes metodológicos sobre la aplicabilidad de modelos predictivos basados en técnicas avanzadas de aprendizaje automático para la sostenibilidad ambiental y la toma de decisiones en conservación, respectivamente. Zuccotto et al. (2024) [27] realizan una revisión exhaustiva de las aplicaciones del aprendizaje por refuerzo en la sostenibilidad ambiental, destacando su capacidad para abordar problemas complejos y dinámicos mediante la simulación de escenarios y la optimización de estrategias. El artículo analiza cómo las técnicas de aprendizaje por refuerzo han evolucionado desde aplicaciones iniciales en videojuegos hasta su uso en conservación, gestión de recursos naturales y mitigación de impactos climáticos. Sin embargo, subraya que la implementación efectiva de estas técnicas depende de la colaboración interdisciplinaria y del desarrollo de simuladores robustos que reflejen la complejidad de los sistemas ambientales. Por su parte, Lapeyrolerie et al. (2022) [29] exploran específicamente el uso del aprendizaje por refuerzo profundo en decisiones de conservación. El estudio presenta casos aplicados como la gestión de cuotas de pesca y la protección de puntos críticos ecológicos, demostrando que estas técnicas pueden superar estrategias tradicionales en escenarios dinámicos e inciertos. Estos estudios sugieren que el aprendizaje por refuerzo profundo no solo permite optimizar decisiones actuales, sino también generar soluciones adaptativas para escenarios futuros.

Los estudios de caso presentados por Zuccotto et al. (2024) [27] ofrecen un marco metodológico sólido para aplicar técnicas de aprendizaje por refuerzo en la modelación de dinámicas complejas en sistemas ambientales. Este enfoque se alinea con los objetivos del presente proyecto, especialmente en lo relacionado con la validación y comparación de métodos de modelado, proporcionando estrategias que podrían integrarse en la fase de modelado predictivo

para estimar el crecimiento de datos sobre biodiversidad. Por su parte, Lapeyrolerie et al. (2022) [29] resalta la aplicabilidad del aprendizaje por refuerzo profundo para abordar problemas complejos de toma de decisiones en conservación, al tiempo que identifica los desafíos inherentes a la implementación de estas metodologías en escenarios realistas. Aunque ambos artículos destacan la relevancia de las metodologías de aprendizaje por refuerzo y proponen alternativas prácticas para su implementación, ninguno aborda la integración de variables socioeconómicas o de políticas públicas en la proyección de datos de biodiversidad. Esto abre un espacio para explorar y probar estas metodologías en contextos previamente no examinados, permitiendo además formular nuevas preguntas sobre el uso del aprendizaje por refuerzo para informar decisiones estratégicas en conservación y gestión de la biodiversidad.

4 METODOLOGÍA BAJO EL MODELO CRISP-DM

El desarrollo de este proyecto se apoya en el modelo CRISP-DM (Cross-Industry Standard Process for Data Mining) [30], una metodología ampliamente utilizada en proyectos de ciencia de datos por su capacidad para estructurar procesos analíticos complejos y mantener coherencia entre los objetivos del problema, los datos disponibles y las técnicas de modelado empleadas. Esta metodología proporciona una estructura sistemática que guía el desarrollo del proyecto desde la comprensión inicial del problema y los objetivos del negocio hasta la implementación final de la solución analítica, permitiendo una adaptación iterativa basada en los hallazgos y requerimientos que surgen durante el proceso. CRISP-DM se compone de seis fases principales interrelacionadas, las cuales no se ejecutan de forma estrictamente secuencial, sino que pueden interactuar cíclicamente entre sí, permitiendo retroalimentación y ajustes a lo largo del ciclo de vida del proyecto [31] (Fig. 2).

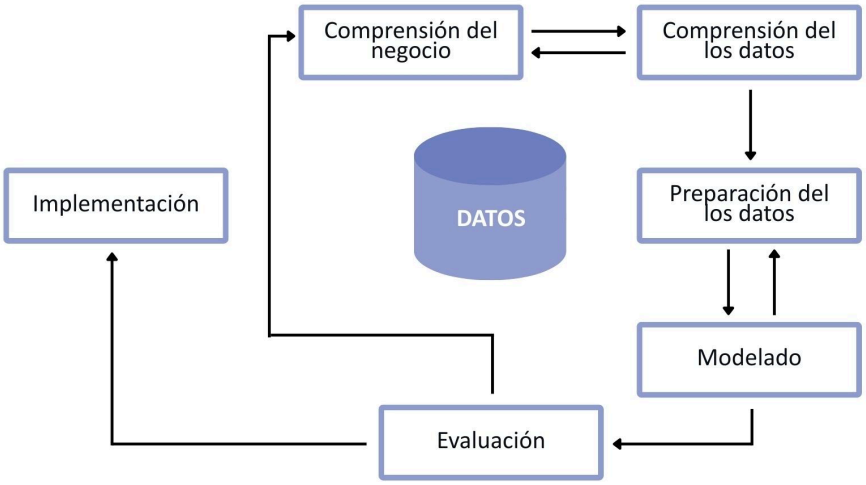


Figura 2. Esquema general del modelo CRISP-DM. Adaptado de CRISP-DM 1.0: Step-by-step data mining guide, por Chapman et al. (2000). Elaboración propia.

Desde CRISP-DM, la comprensión del negocio se centra en precisar el problema y traducirlo en objetivos analíticos: comprender y proyectar el crecimiento de la publicación de datos de biodiversidad, e identificar factores socioeconómicos e institucionales que puedan estar asociados a ese comportamiento, con un énfasis aplicado en el caso del nodo SiB Colombia. A partir de ello, la comprensión de los datos aborda la identificación de fuentes de información pertinentes, la definición de la variable objetivo vinculada a la publicación de registros, y la selección de variables explicativas que representen dimensiones socioeconómicas, de capacidades científicas y de gobernanza, priorizando series comparables y consistentes.

La preparación de los datos comprende la consolidación de un conjunto de datos que permita el análisis y el modelado, integrando datos sobre biodiversidad con indicadores socioeconómicos y variables institucionales. Esta fase incluye tareas típicas de un proyecto de ciencia de datos, como depuración y estandarización, tratamiento de valores faltantes, transformación de variables para su uso en algoritmos de modelado y ajustes necesarios para preservar la coherencia temporal del fenómeno analizado. Enseguida, la fase de modelado consiste en construir y contrastar modelos estadísticos y de aprendizaje automático adecuados para datos con estructura temporal, de manera que se puedan capturar tendencias y relaciones relevantes, manteniendo criterios de interpretabilidad y desempeño.

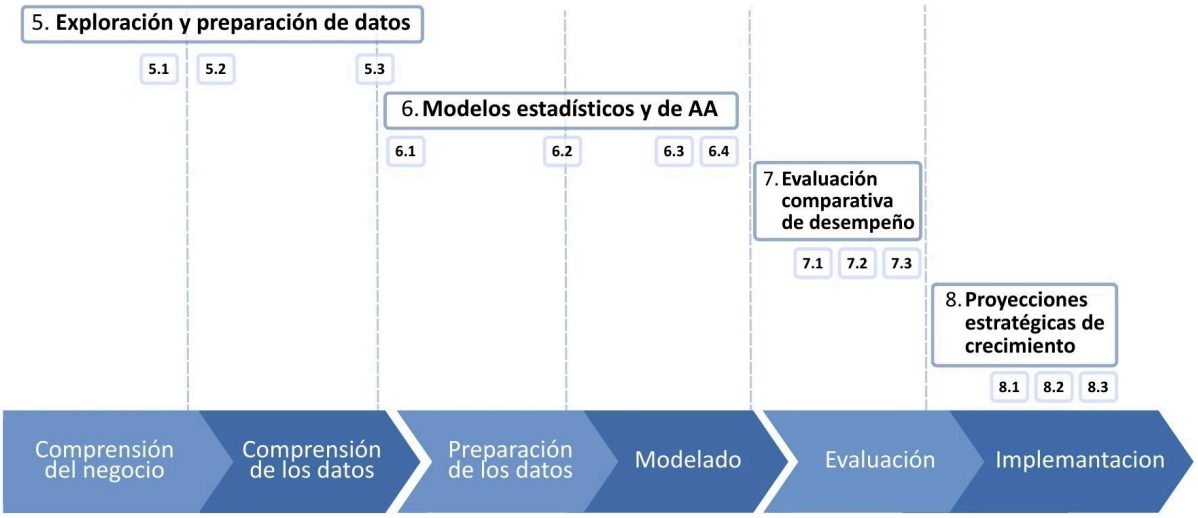


Figura 3. Correspondencia entre CRISP-DM y la estructura metodológica del proyecto. Fuente: elaboración propia.

La evaluación se orienta a comparar los modelos con métricas de desempeño apropiadas para el problema, bajo estrategias de validación que respeten el orden temporal, y a seleccionar el enfoque más consistente para generar proyecciones. Finalmente, la implementación se entiende como la aplicación del modelo seleccionado para construir escenarios de predicción que apoyen la reflexión estratégica sobre el crecimiento de publicación de datos, conectando el ejercicio analítico con posibles líneas de acción en el marco de la ciencia abierta y la gestión de información para biodiversidad. La Figura 3 presenta cómo estas fases se articulan con la estructura metodológica de los capítulos siguientes (5 al 8), que desarrollan cada etapa con mayor profundidad.

5 EXPLORACIÓN Y PREPARACIÓN DE DATOS

Este capítulo desarrolla el primer objetivo específico, orientado a identificar las variables socioeconómicas y gubernamentales que influyen en el crecimiento de la generación y publicación de datos sobre biodiversidad en el nodo nacional SiB Colombia, a partir del análisis de información histórica y de acceso público. Para ello, se presenta el proceso de exploración, obtención y depuración de los datos, el cual se organiza en tres partes: la selección de variables utilizadas en el estudio, el análisis exploratorio que permite comprender el comportamiento y relaciones, y la preparación final del conjunto de datos que sirvió de base para el desarrollo de los modelos predictivos.

5.1. SELECCIÓN DE DATOS Y VARIABLES

5.1.1. Datos socioeconómicos

La selección de variables socioeconómicas se basó en la disponibilidad y cobertura temporal de los datos, eligiendo aquellos indicadores que cuentan con series históricas suficientemente largas y consistentes para los países participantes y el periodo de análisis. En segundo lugar, se evaluó la confiabilidad de las fuentes, seleccionando únicamente variables provenientes de organismos reconocidos y con metadatos claros que permitieran su interpretación adecuada. En tercer lugar, se tuvo en cuenta la representatividad y comparabilidad internacional, privilegiando aquellos indicadores que se reportan bajo estándares homogéneos entre países. Adicionalmente, se procuró una cobertura temática equilibrada, seleccionando variables de diferentes ejes como ciencia y tecnología, economía, educación, ambiente, sector público e infraestructura, para capturar una visión multidimensional de los factores que pueden incidir en el fenómeno estudiado.

En este proceso de selección se evaluaron portales institucionales como el Departamento Administrativo Nacional de Estadística (DANE), el Banco Mundial, la Comisión Económica para América Latina y el Caribe (CEPAL) y otros organismos internacionales que consolidan estadísticas económicas, sociales y ambientales. Además, considerando el enfoque del proyecto en relación a la ciencia abierta, se exploraron indicadores relacionados con la apertura de datos por parte de los gobiernos. Para ello, se consultaron portales como el Open Data Inventory (ODIN) y la vinculación de los países a procesos como Open Government Partnership (OGP).

Como resultado de esta fase, se seleccionaron como fuentes principales el Banco Mundial, por la amplitud y consistencia de sus series temporales, y la Open Government Partnership (OGP), por su cobertura sobre los países participantes de la red GBIF y la disponibilidad de información sobre políticas de gobierno abierto. En el caso del Banco Mundial, la consulta de los indicadores se efectuó a través de su *Interfaz de Programación de Aplicaciones* (API), utilizando la librería *wbgapi* [32], a partir de la cual se obtuvieron los valores de las variables seleccionadas para los países miembros de GBIF en el periodo 2000 a 2024. Para el caso de los datos de gobierno abierto (OGP), se creó una nueva variable que registra el año de vinculación de cada país participante de la red GBIF a la iniciativa OGP. Esta variable permitió incorporar una dimensión institucional, al reflejar los años en los que los países adoptaron compromisos formales en materia de transparencia, participación y apertura de datos, factores que pueden influir directamente en la capacidad de publicación de información sobre biodiversidad.

La base de datos resultante contiene un total de 19 variables, de las cuales 14 corresponden a indicadores socioeconómicos distribuidos en 6 ejes temáticos y 6 categorías (Tabla 1).

Tabla 1. Descripción de las variables socioeconómicas (predictivas).

Variable	Unidad de medida	Descripción
Ciencia & Tecnología		
art_cientificos	Artículos en publicaciones científicas y técnicas (float64)	Los artículos en publicaciones científicas y técnicas se refieren a la serie de artículos científicos y de ingeniería publicados en los siguientes campos: física, biología, química, matemática, medicina clínica, investigación biomédica, ingeniería y tecnología, y ciencias de la tierra y el espacio. [https://datos.bancomundial.org/indicador/IP.JRN.ARTC.SC]
investigadores_RD	Número de investigadores dedicados a investigación y desarrollo por cada millón de personas (float64)	Los investigadores dedicados a investigación y desarrollo son profesionales que se dedican al diseño o creación de nuevos conocimientos, productos, procesos, métodos o sistemas, y a la gestión de los proyectos correspondientes. Se incluyen los estudiantes de doctorados (nivel 6 de la CINE 97) dedicados a investigación y desarrollo. [https://datos.bancomundial.org/indicador/SP.POP.SCIE.RD.P6]
gasto_RD_pib	Porcentaje del PIB en gasto en	Los gastos en investigación y desarrollo son gastos corrientes y de capital (público y privado) en trabajo creativo realizado

	investigación y desarrollo (float64)	sistemáticamente para incrementar los conocimientos, incluso los conocimientos sobre la humanidad, la cultura y la sociedad, y el uso de los conocimientos para nuevas aplicaciones. El área de investigación y desarrollo abarca la investigación básica, la investigación aplicada y el desarrollo experimental. [https://datos.bancomundial.org/indicador/GB.XPD.RSDV.GD.ZS]
tecnicos_RD	Número de técnicos de investigación y desarrollo por cada millón de personas (float64)	Los técnicos de investigación y desarrollo y el personal equivalente son personas cuyas tareas principales exigen conocimiento técnico y experiencia en ingeniería, ciencias naturales (técnicos), o ciencias sociales y humanidades (personal equivalente). Participan en investigación y desarrollo realizando tareas científicas y técnicas que abarcan la aplicación de conceptos y métodos operativos, por lo general supervisados por investigadores. [https://datos.bancomundial.org/indicador/SP.POP.TECH.RD.P6]
Educación		
gasto_educacion_pib	Porcentaje total del PIB en gasto público en educación (float64)	El gasto del gobierno general en educación (corriente, capital y transferencias) se expresa como porcentaje del PIB. Incluye gastos financiados por transferencias de fuentes internacionales al gobierno. El gobierno general generalmente se refiere a los gobiernos locales, regionales y centrales. [https://datos.bancomundial.org/indicador/SE.XPD.TOTL.GD.ZS]
inscripcion_primaria	Porcentaje bruto de inscripción escolar, nivel primario (float64)	Tasa bruta de matrícula, educación primaria, total. Corresponde al número total de estudiantes matriculados en educación primaria, independientemente de su edad, expresado como porcentaje de la población total en edad oficial de cursar enseñanza primaria. La TBM puede ser superior al 100% debido a la inclusión de estudiantes mayores y menores a la edad oficial ya sea por repetir grados o por un ingreso precoz o tardío a dicho nivel de enseñanza. [https://datos.bancomundial.org/indicador/SE.PRM.ENRR]
inscripcion_secundaria	Porcentaje bruto de inscripción escolar, nivel secundario (float64)	Tasa bruta de matrícula, enseñanza secundaria, todos los programas, total. Corresponde al número total de estudiantes matriculados en educación secundaria, independientemente de su edad, expresado como porcentaje de la población total en edad oficial de cursar la secundaria. La TBM puede ser superior a 100% debido a la inclusión de estudiantes mayores y menores

		a la edad oficial ya sea por repetir grados o por un ingreso precoz o tardío a dicho nivel de enseñanza. [https://datos.bancomundial.org/indicador/SE.SEC.ENRR]
inscripcion_terciaria	Porcentaje bruto de inscripción escolar, nivel terciario (float64)	Tasa bruta de matrícula, educación superior (niveles 5 y 6 de la Clasificación Internacional Normalizada de la Educación (CINE)), total. Corresponde al número total de estudiantes matriculados en educación superior (niveles 5 y 6 de la CINE), independientemente de su edad, expresado como porcentaje de la población total del grupo etario cinco años después de finalizar la enseñanza secundaria. [https://datos.bancomundial.org/indicador/SE.TER.ENRR]
gasto_educacion _gobierno	Porcentaje del gasto total del gobierno en educación (float64)	El gasto público en educación como porcentaje del gasto total del Gobierno corresponde al gasto público total (corriente y de capital) en educación, expresado como porcentaje del gasto total del Gobierno en todos los sectores en un año financiero determinado. El gasto público en educación incluye el gasto del Gobierno en instituciones educativas (públicas y privadas), administración educativa y subsidios para entidades privadas (estudiantes/hogares y otras entidades privadas). [https://datos.bancomundial.org/indicador/SE.XPD.TOTL.GB.ZS]
Cambio climático		
areas_protegidas	Porcentaje total de la superficie territorial de áreas protegidas terrestres y marinas (float64)	Las áreas protegidas terrestres son zonas total o parcialmente protegidas de por lo menos 1.000 hectáreas designadas por autoridades nacionales como reservas científicas con acceso público limitado, parques nacionales, monumentos nacionales, reservas naturales o santuarios de la naturaleza, paisajes protegidos y zonas manejadas principalmente para uso sostenible. Las áreas marinas protegidas son zonas de terreno intermareal o submareal, junto con sus aguas suprayacentes y su flora, fauna y características históricas y culturales conexas, que han sido reservadas por ley o por cualquier otro medio eficaz para proteger parte del entorno que encierra o su totalidad. No se incluyen las zonas protegidas conforme a leyes locales o provinciales. [https://datos.bancomundial.org/indicador/ER.PTD.TOTL.ZS]
area_selvatica_km2	Kilómetros cuadrados área	La superficie forestal se refiere a las tierras con agrupaciones de árboles naturales o plantados de por lo menos 5 metros in situ,

	selvática (float64)	sean estas para usos productivos o no, y excluye las poblaciones en los sistemas de producción agrícola (por ejemplo, en plantaciones frutales y sistemas agroforestales) y los árboles en los parques y jardines urbanos. [https://datos.bancomundial.org/indicador/AG.LND.FRST.K2]
Ambiente		
superficie_total_km2	Kilómetros cuadrados área de tierra (float64)	El área de tierra es la superficie total de un país, sin incluir la superficie cubierta por masas de agua interiores, los derechos del país sobre la plataforma continental ni las zonas económicas exclusivas. En la mayoría de los casos, la definición de masas de agua interiores incluye los principales ríos y lagos. [https://datos.bancomundial.org/indicador/AG.LND.TOTL.K2]
Sector público		
efectividad_gobierno	Estimación de la eficacia del gobierno (float64)	La Eficacia Gubernamental captura la percepción de la calidad de los servicios públicos, la calidad de la función pública y su grado de independencia frente a presiones políticas, la calidad de la formulación e implementación de políticas, y la credibilidad del compromiso del gobierno con dichas políticas. La estimación proporciona la puntuación del país en el indicador agregado, en unidades de una distribución normal estándar, es decir, con un rango aproximado de -2,5 a 2,5. [https://databank.worldbank.org/metadataglossary/worldwide-governance-indicators/series/GE.EST]
Infraestructura		
uso_internet	Porcentaje individuos que utilizan Internet (float64)	Los usuarios de Internet son personas que han utilizado Internet (desde cualquier lugar) en los últimos 3 meses. Internet se puede utilizar a través de una computadora, un teléfono móvil, un asistente digital personal, una máquina de juegos, una televisión digital, etc. [https://datos.bancomundial.org/indicador/IT.NET.USER.ZS]
Categóricos		
country	Nombre del país (object)	El término país, es utilizado de manera intercambiable con economía, no implica independencia política, sino que se refiere a cualquier territorio cuyas autoridades reportan

		estadísticas sociales o económicas por separado. [https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries].
countryCode	Códigos de país ISO de dos letras (object)	Corresponde al código de país basado en el estándar internacional ISO 3166-1 alfa-2, publicado por la Organización Internacional de Normalización (ISO). Este código de dos letras permite identificar de manera única a cada país o territorio.
región	Región geográfica (object)	Las agrupaciones se basan principalmente en las regiones utilizadas con fines administrativos por el Banco Mundial. [https://datatopics.worldbank.org/sdgatlas/archive/2017/the-world-by-region.html]
incomeLevel	Nivel de ingreso del país (object)	Las economías se dividen actualmente en cuatro grupos de ingresos: bajo, medio-bajo, medio-alto y alto. El ingreso se mide utilizando el ingreso nacional bruto (INB) per cápita, en dólares estadounidenses, convertido desde la moneda local mediante el método Atlas del Banco Mundial. Las estimaciones del INB son obtenidas por economistas de las oficinas de país del Banco Mundial; y el tamaño de la población es estimado por los demógrafos del Banco Mundial a partir de diversas fuentes, incluyendo las Perspectivas de Población Mundial bienales de las Naciones Unidas. [https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups]
ogp_membership	Membresía OGP (object)	Indica si el país contaba con una política nacional de gobierno abierto, reflejada por su membresía a la Open Government Partnership (OGP) en el año correspondiente. Se asigna “Sí” a los países que eran miembros de la OGP durante el año del registro, y “No” en caso contrario. La membresía a la OGP implica la existencia de compromisos gubernamentales para mejorar la transparencia, la rendición de cuentas y la participación ciudadana a través de planes de acción co-creados con la sociedad civil. [https://www.opengovpartnership.org/our-members/#national]

5.1.2. Datos sobre biodiversidad

Los datos de biodiversidad constituyen la variable dependiente o a predecir en el modelo. Para responder a la pregunta de investigación, fue necesario disponer de información que describiera el comportamiento del volumen de publicación de datos sobre biodiversidad para los países miembros de la red GBIF durante el periodo 2000 a 2024.

GBIF ofrece diversas fuentes de acceso a la información, entre las que destacan su API [33], el Portal de datos [34], que permite descargar los aproximadamente 3 mil millones de registros disponibles en la red, y los Cubos de Datos [35], que entregan información agregada por dimensiones taxonómicas y geográficas. Sin embargo, para este estudio se optó por utilizar los Archivos de Análisis de GBIF (GBIF Analytics Files) [36], [37], los cuales consolidan reportes anuales con fecha de corte uniforme. Estos archivos proporcionan una fuente más eficiente y coherente para recopilar los datos de manera sistemática, al reflejar los resultados anuales del proceso de publicación realizado por los nodos y miembros de la red.

El objetivo de esta fase fue identificar y consolidar los datos que representarán el estado de la publicación de biodiversidad a nivel nacional durante la ventana temporal definida. A partir de este procedimiento se obtuvieron dos indicadores principales: (i) el número total de registros de biodiversidad publicados asociados a cada país (*occurrenceCount*) por año, y (ii) el número de registros publicados exclusivamente por instituciones nacionales (*occurrenceCount_publisher*). Estos indicadores permitieron analizar tanto el volumen total de datos movilizados por cada país como el nivel de participación de sus propias instituciones en la generación de información. Dado que el propósito del proyecto fue comprender cómo los indicadores socioeconómicos y de desarrollo influyen en la publicación de datos, se definió *occurrenceCount_publisher* como la variable objetivo principal del modelo predictivo.

El proceso de adquisición y preparación de los datos biológicos se automatizó mediante un *script* en Python, que realizó la ingestión de los archivos de análisis de GBIF a través de sus endpoints de descarga. Para ello, se emplearon las librerías *requests* [38], para la descarga, y *pandas* [39], para el procesamiento y consolidación de una serie temporal país–año ajustada a la fecha de corte más cercana al cierre de cada año. Este procedimiento aseguró la coherencia temporal de las observaciones y constituye la base para los análisis de evolución temporal, detección de puntos de cambio y posterior correlación con las variables socioeconómicas. En la tabla 2 se describen las variables definidas para los datos de biodiversidad.

Tabla 2. Descripción de las variables biodiversidad.

Variable	Unidad de medida	Descripción
occurrenceCount	Número de registros biológicos publicados para el país (float64)	Registro de una observación de la biodiversidad con detalles acerca de la ubicación de un individuo en el tiempo y el espacio, autoría de la observación y descripción de la clasificación taxonómica de la especie (u otro taxón). Los registros biológicos representan la mayoría de los datos publicados a través de GBIF y comprenden, por ejemplo, especímenes y fósiles en colecciones biológicas, observaciones producto de inventarios, proyectos de ciencia participativa, cámaras trampa, información genética o sensores remotos, entre otros. En esta variable se tienen en cuenta los registros aportados por organizaciones del país donde se genera la observación, así como los generados por organizaciones externas al país de la observación. [https://analytics-files.gbif.org/]
occurrenceCount _publisher	Número de registros biológicos publicados por país (float64)	Registro de una observación de la biodiversidad con detalles acerca de la ubicación de un individuo en el tiempo y el espacio, autoría de la observación y descripción de la clasificación taxonómica de la especie (u otro taxón). Los registros biológicos representan la mayoría de los datos publicados a través de GBIF y comprenden, por ejemplo, especímenes y fósiles en colecciones biológicas, observaciones producto de inventarios, proyectos de ciencia participativa, cámaras trampa, información genética o sensores remotos, entre otros. En esta variable se tienen en cuenta los registros aportados únicamente por organizaciones con base en el país donde se genera la observación. [https://analytics-files.gbif.org/]
gbif_member	Miembro de la red de GBIF (object)	Indicador booleano que indica por cada año si el país era miembro de la red de GBIF. [https://www.gbif.org/the-gbif-network]

El proceso automatizado para la extracción de los datos se documentó en un *pipeline* disponible en el repositorio *GitHub* del proyecto.

• Pipeline de recopilación y compilación de datos : https://github.com/rortizgeo/Maestria_CD_Proyecto-Aplicado/blob/main/1_PA_DataExtraction.ipynb

5.2. ANÁLISIS EXPLORATORIO DE DATOS

Con el propósito de comprender la estructura del conjunto de datos y orientar las decisiones para la fase de modelado, se realizó un análisis exploratorio que incluyó la caracterización general de las variables, el análisis de correlaciones, la identificación de valores atípicos, la evaluación de datos faltantes y la revisión de la serie temporal.

5.2.1. Descripción general de los datos

El análisis exploratorio evidenció una alta heterogeneidad entre países en las variables analizadas, coherente con las diferencias en tamaño territorial, nivel de ingreso y desarrollo científico. Las medidas de centralidad (Tabla 3) muestran amplias variaciones en la superficie total, que va de 89 km² a más de 15 millones km², en la proporción de áreas protegidas, que oscila entre 0,1 % y 55,8 %, y en los indicadores de ciencia y tecnología, como la inversión en investigación y desarrollo, el número de investigadores y las publicaciones científicas, que reflejan las asimetrías estructurales entre países.

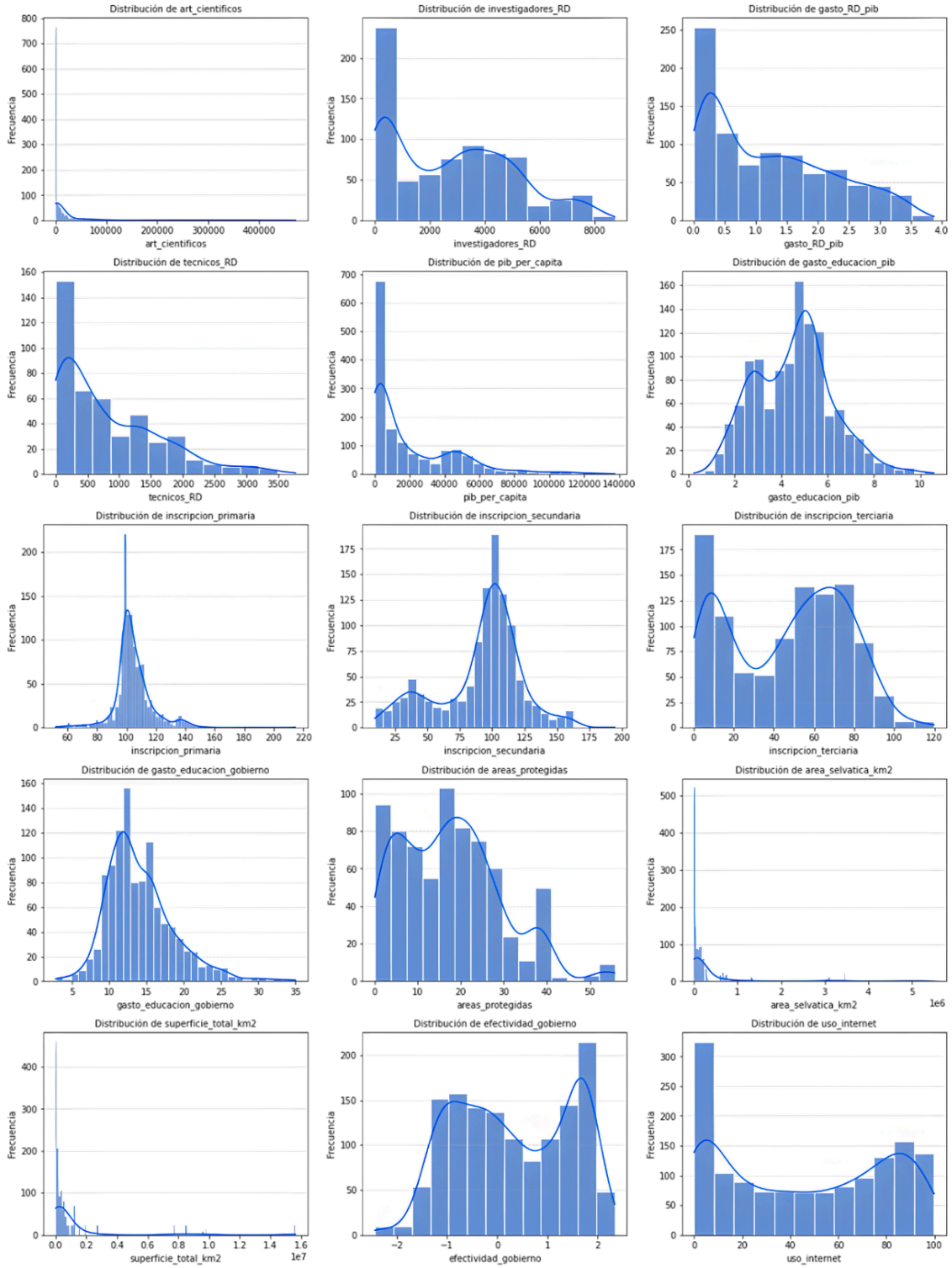
Tabla 3. Análisis descriptivo de medidas de tendencia central y dispersión.

Variable	Media	std	min	25%	50%	75%	max
year	2012,00	7,21	2000,00	2006,00	2012,00	2018,00	2024,00
area_selvatica_km2	330847,50	879032,56	89,50	12689,49	66399,50	176825,20	5510886,00
superficie_total_km2	1083790,81	2684184,37	470,00	69700,00	243610,00	624499,25	15640500,00
areas_protegidas	17,67	11,67	0,10	7,60	17,40	24,30	55,80
gasto_RD_pib	1,23	1,01	0,01	0,30	1,06	1,98	3,87
efectividad_gobie	0,31	1,15	-2,44	-0,70	0,22	1,50	2,35

rno							
art_cientificos	17877,98	55542,10	0,00	65,27	677,45	11458,42	472448,44
uso_internet	46,15	34,24	0,02	10,00	46,25	80,00	99,80
pib_per_capita	20783,14	25477,89	111,41	1713,42	8060,45	37940,34	137516,59
inscripcion_primaria	104,78	12,85	52,00	98,99	102,56	109,52	214,67
inscripcion_secundaria	91,14	32,08	9,64	76,94	99,06	110,25	194,46
inscripcion_terciaria	46,51	29,23	0,32	15,73	51,85	70,96	119,68
gasto_educacion_gobierno	14,14	4,52	2,82	11,07	13,10	16,30	35,01
gasto_educacion_pib	4,54	1,67	0,24	3,17	4,69	5,51	10,59
investigadores_RD	2807,98	2253,76	13,28	514,90	2824,70	4491,29	8735,60
tecnicos_RD	824,37	799,00	2,35	168,23	606,18	1332,35	3766,86
occurrenceCount	15770421,35	65935991,68	65,00	110035,25	835354,50	7666272,25	1100336258,00
occurrenceCount_publisher	21560125,45	80493145,32	52,00	96247,50	1582525,00	13415924,50	1177656893,00

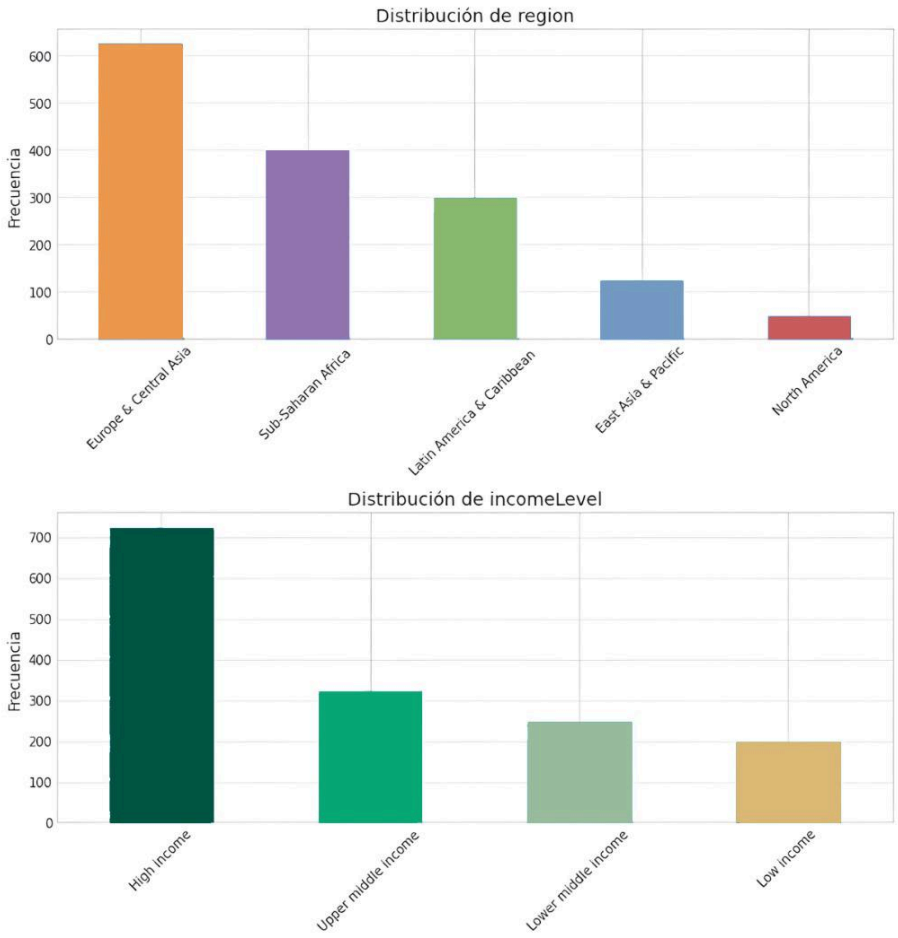
La distribución de las variables numéricas (Fig. 4) presenta un sesgo positivo, con concentraciones altas en un pequeño grupo de países y valores bajos en la mayoría, especialmente en los indicadores de investigación y desarrollo, PIB per cápita y publicaciones científicas.

Figura 4. Distribución general de las variables numéricas.



Algunas variables muestran distribuciones bimodales relacionadas con la región y el nivel de ingreso, evidenciando contrastes entre el norte y el sur global. El análisis de las variables categóricas (Fig. 5) confirma que la mayor proporción de datos para este análisis proviene de Europa y Asia Central, seguidas de África Subsahariana y Latinoamérica y el Caribe, regiones donde se concentran los países con mayor antigüedad en la red y con estructuras científicas consolidadas. La mayoría de los países con datos completos corresponden a economías de ingresos altos.

Figura 5. Distribución general de las variables categóricas.



Desde una perspectiva temporal, los países participantes registran en promedio 16 años de permanencia en la red GBIF (Fig. 6), con una desviación estándar de 8.5 años, lo que evidencia una amplia variabilidad en la duración de su participación.

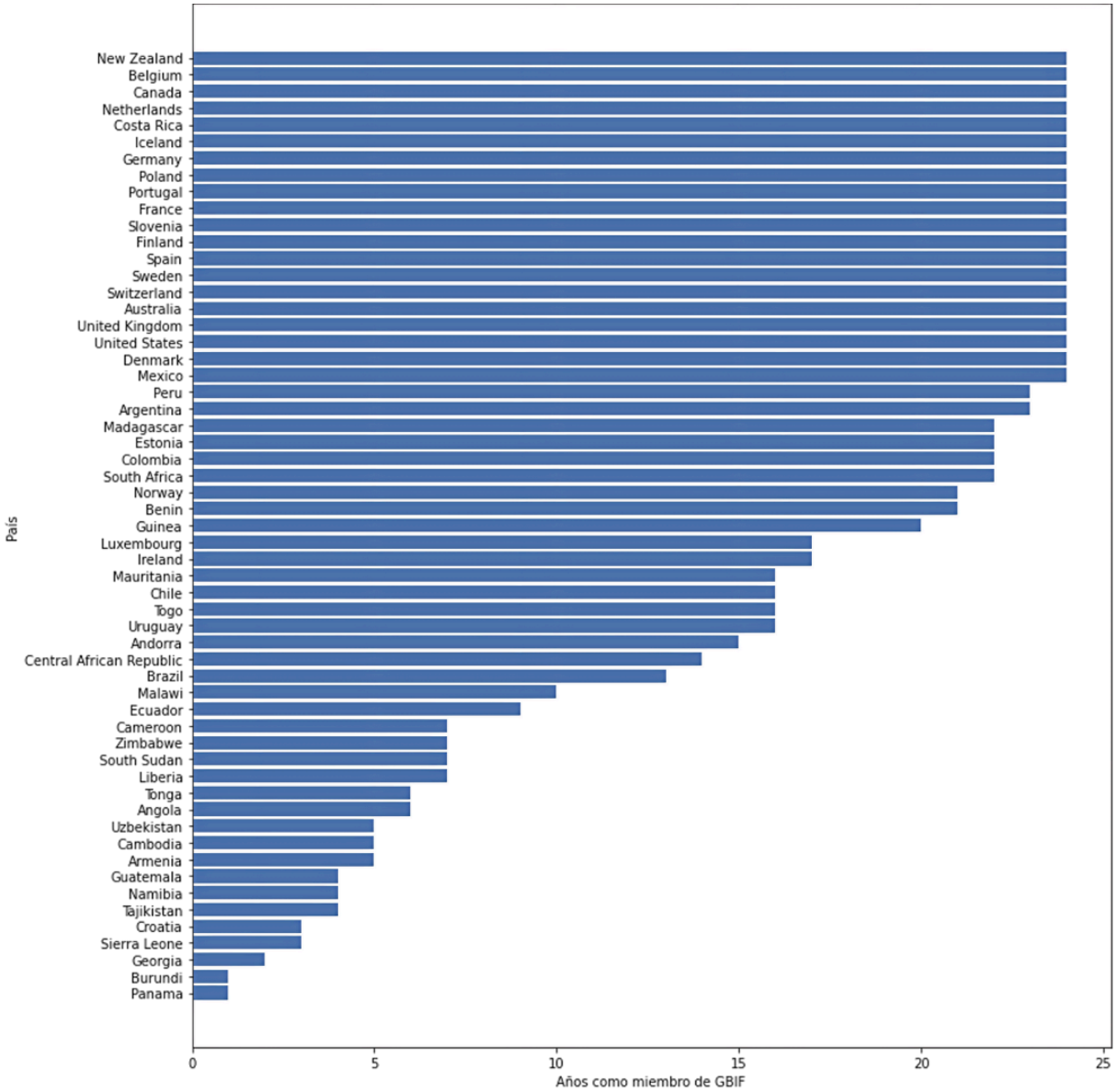


Figura 6. Tiempo de participación de países en GBIF.

5.2.2. Análisis de correlaciones

La matriz de correlación entre las variables numéricas (Fig. 7) evidencia una asociación fuerte entre los indicadores de ciencia, educación e innovación, en particular entre el gasto en investigación y desarrollo (*gasto_RD_pib*), el número de investigadores dedicados a I+D (*investigadores_RD*), el número de técnicos en I+D (*tecnicos_RD*) y la producción de artículos científicos (*art_cientificos*), con coeficientes superiores a 0.8. Este patrón refleja que los países con mayor inversión en investigación y una comunidad científica más amplia presentan también una producción académica más alta, lo que refuerza la coherencia interna del conjunto de datos y su correspondencia con el comportamiento esperado de los sistemas de ciencia y tecnología a nivel nacional. Además, las variables de educación y conectividad, como el PIB per cápita (*pib_per_capita*), la inscripción en educación terciaria (*inscripcion_terciaria*) y el uso de internet (*uso_internet*), muestran correlaciones moderadas con los indicadores de ciencia y desarrollo, sugiriendo que el fortalecimiento educativo y tecnológico es un factor estructural asociado al crecimiento de las capacidades científicas.

En relación con las variables de publicación de datos de biodiversidad, la correlación entre el número de registros publicados por instituciones nacionales (*occurrenceCount_publisher*) y los indicadores de ciencia y educación es positiva y significativa, alcanzando valores de 0.76 con *art_cientificos*, 0.74 con *investigadores_RD* y *gasto_RD_pib*, y 0.69 con *inscripcion_terciaria*. Esto sugiere que la capacidad de movilización de datos sobre biodiversidad depende en gran medida del desarrollo científico y la infraestructura institucional de los países. En contraste, las variables ambientales como *superficie_total_km2*, *area_selvatica_km2* y *areas_protegidas* presentan correlaciones bajas o negativas con las variables de GBIF, lo que indica que ni la extensión territorial ni la proporción de áreas naturales garantizan por sí solas una mayor disponibilidad de información.

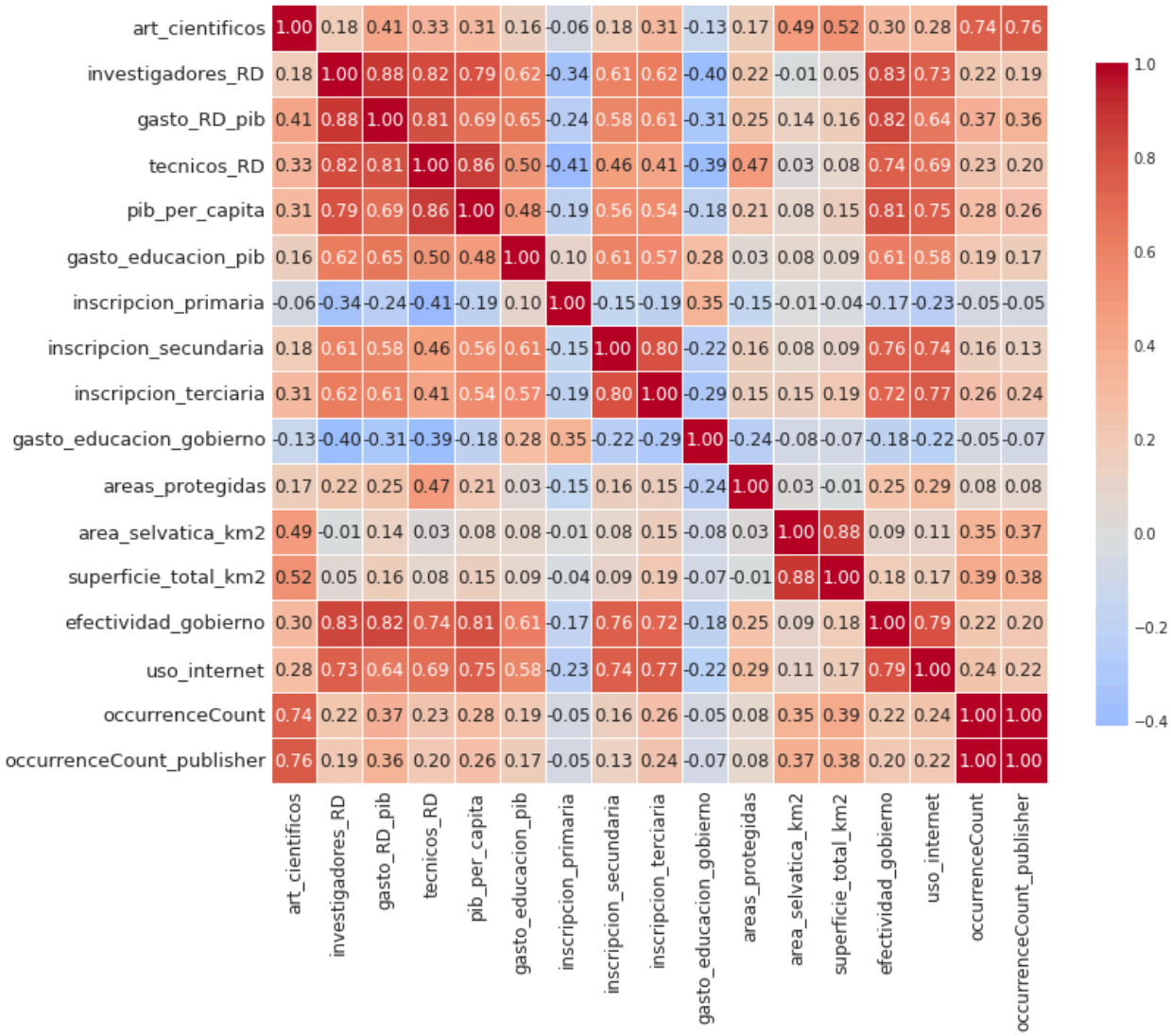


Figura 7. Correlación de variables numéricas en el dataset.

5.2.3. Análisis de datos faltantes

El análisis de completitud reveló que el 93,7 % de las observaciones del conjunto de datos presentan al menos un valor faltante, hecho que refleja la heterogeneidad en la disponibilidad de información entre países y años. Las variables más afectadas corresponden a los indicadores de ciencia y tecnología, entre ellos *tecnicos_RD* (70 % de valores faltantes), *investigadores_RD* (49 %) y *gasto_RD_pib* (41 %). Asimismo, las variables de cobertura ambiental como *areas_protegidas* muestran vacíos cercanos al 52 %, mientras que las de educación y

gobernanza presentan una menor proporción de ausencias, en especial *pib_per_capita* y *uso_internet*, cuyos vacíos son inferiores al 10 %. En conjunto, estos resultados reflejan diferencias estructurales en la capacidad de reporte y en la disponibilidad histórica de estadísticas nacionales (Fig. 8a, 8b).

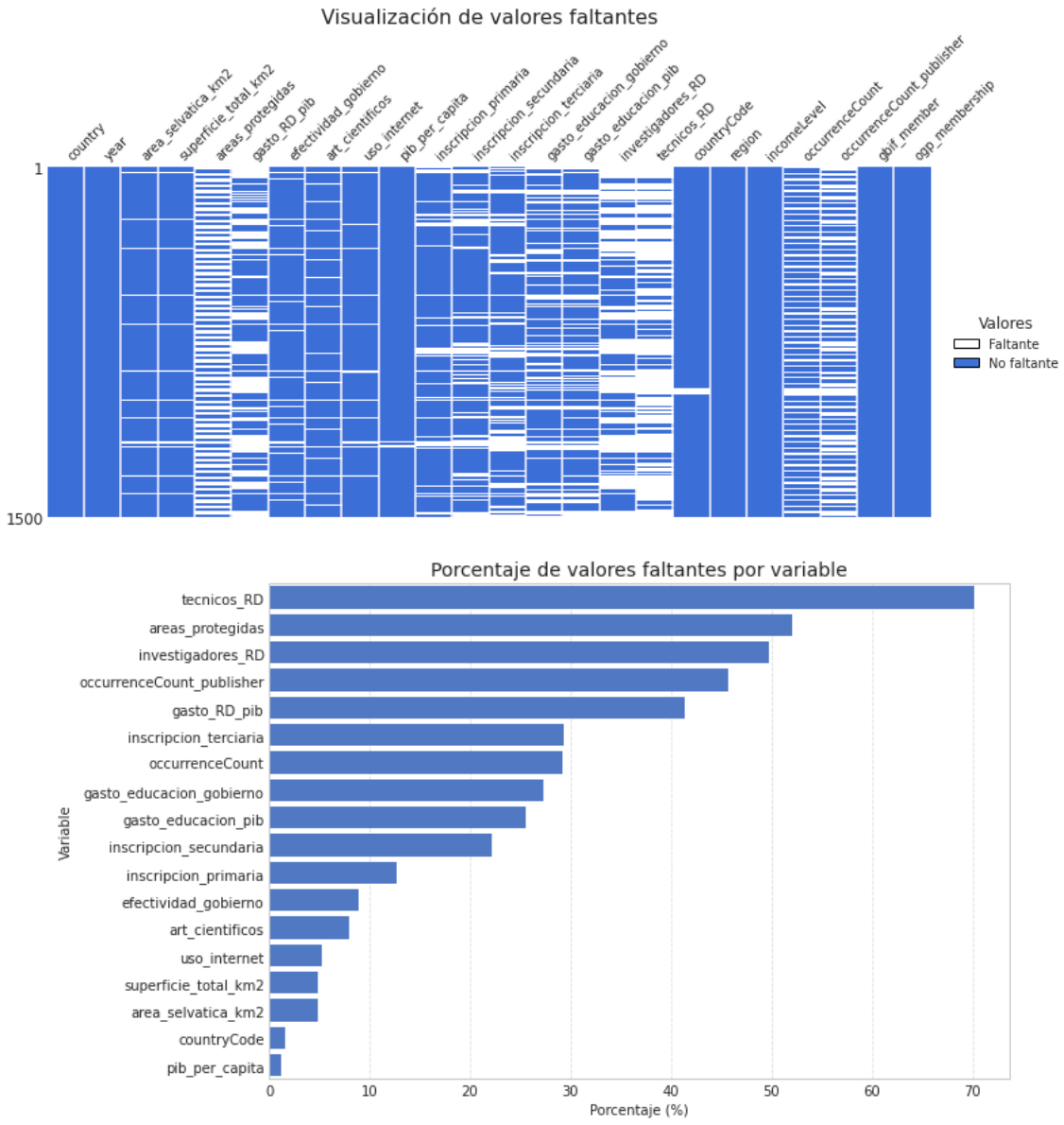


Figura 8. Datos faltantes por variable (a) Arriba. Visualización comparativa de datos faltantes por variable. (b) Abajo. Visualización de distribución porcentual de datos faltantes

La distribución temporal de los valores faltantes evidencia un patrón coherente con la evolución de la red GBIF. Los primeros años de la serie (2000–2006) concentran altos porcentajes de vacíos, superiores al 20 %, debido a la ausencia de datos de publicación durante la fase inicial de consolidación de la red. De manera similar, los años más recientes (2023–2024) presentan los mayores niveles de incompletitud, cercanos al 50 %, atribuibles a la falta de actualización de las bases del Banco Mundial y otras fuentes globales. Este comportamiento está alineado con la trayectoria de vinculación de los países a GBIF (Fig. 9). Así, los países con incorporación más reciente, particularmente en África Subsahariana y América Latina (Anexo 1), exhiben vacíos sistemáticos que reflejan tanto su menor permanencia en la red como el rezago en la consolidación de indicadores nacionales.

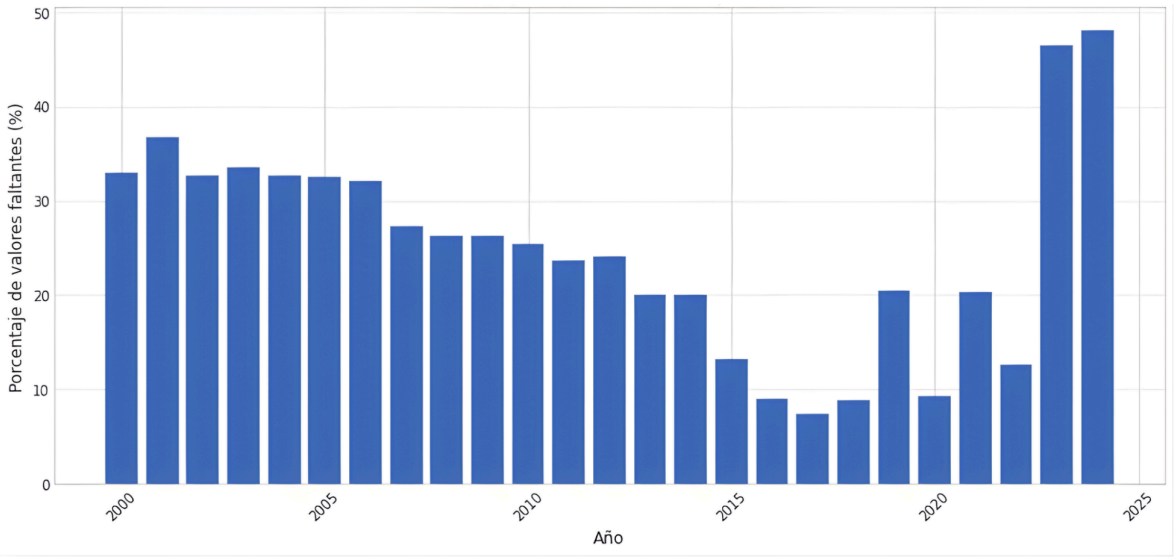


Figura 9. Porcentaje de datos faltantes por año.

5.2.4. Análisis de la serie temporal

El análisis de la serie temporal de la variable *occurrenceCount_publisher* revela un crecimiento sostenido y acelerado en la publicación de datos de biodiversidad desde la consolidación de la red GBIF (Fig. 10). Entre 2000 y 2006 no se registra publicación de datos, reflejando la fase inicial de estructuración del sistema global. A partir de 2007, la tendencia adquiere un comportamiento claramente exponencial, alcanzando más de 2.8 mil millones de registros en 2024. Este patrón evidencia la expansión progresiva de la infraestructura global de datos abiertos y la consolidación de capacidades institucionales para la movilización de información biológica a escala mundial.

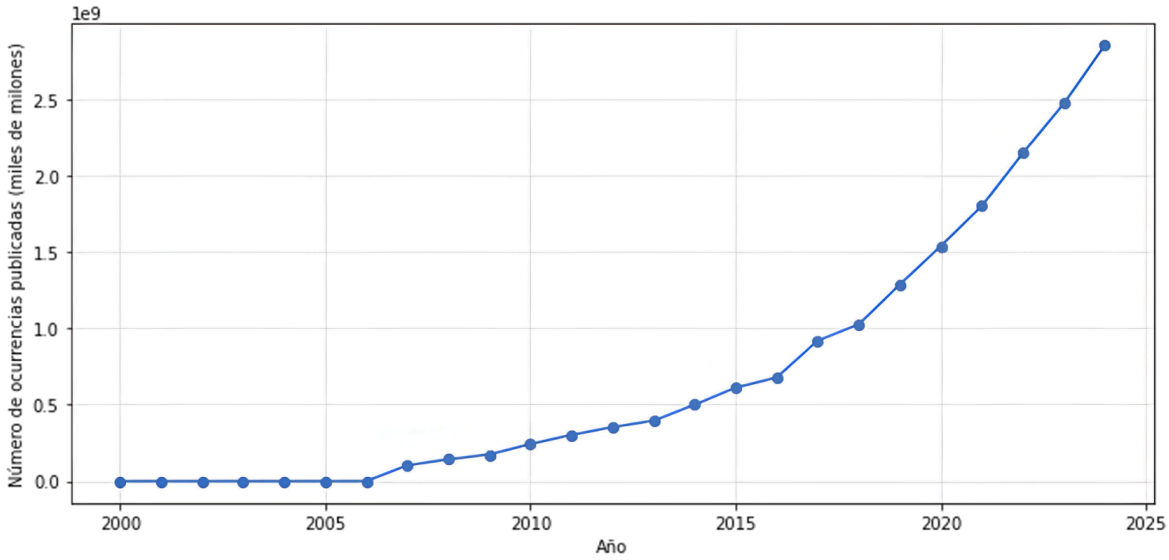


Figura 10. Evolución anual de la publicación de datos de biodiversidad a nivel global

Sin embargo, este comportamiento es variable a nivel regional. En la Figura 11 se observa una diferenciación en la tendencia de publicación en las regiones de Europa, Asia central y Norte América.

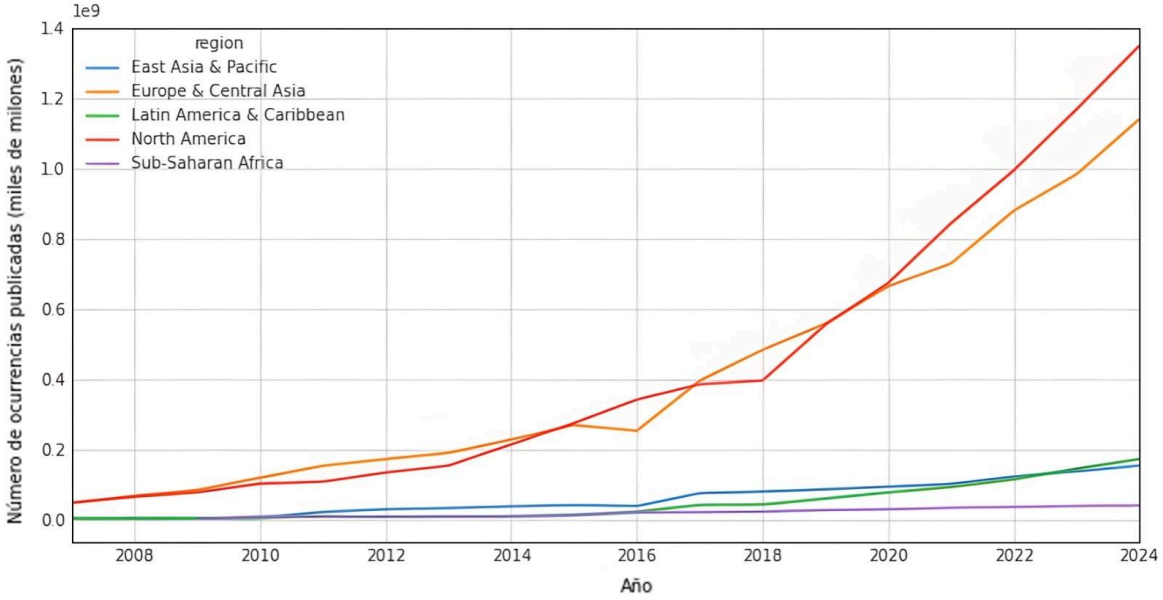


Figura 11. Evolución anual de la publicación de datos de biodiversidad a nivel regional.

El análisis de puntos de cambio en la serie (Fig. 12) identificó tres agrupaciones temporales

consistentes: un primer grupo de países con cambios significativos alrededor de los diez años de participación en la red, un segundo entre los catorce y dieciséis años, y un tercero hacia los veinte años. Estos resultados sugieren que la permanencia prolongada, especialmente más allá de una década, constituye un factor determinante para alcanzar una publicación sostenida y estable de datos. Asimismo, refuerzan la necesidad de incorporar la madurez temporal de los países como variable clave en el modelado y proyección del crecimiento futuro. De forma complementaria, los análisis adicionales (Figura 11) confirman que el crecimiento global se concentra principalmente en países de ingresos altos, responsables de la mayor parte de los registros publicados, mientras que las trayectorias nacionales presentan comportamientos más heterogéneos, asociados posiblemente a las diferencias en desarrollo científico e infraestructura institucional.

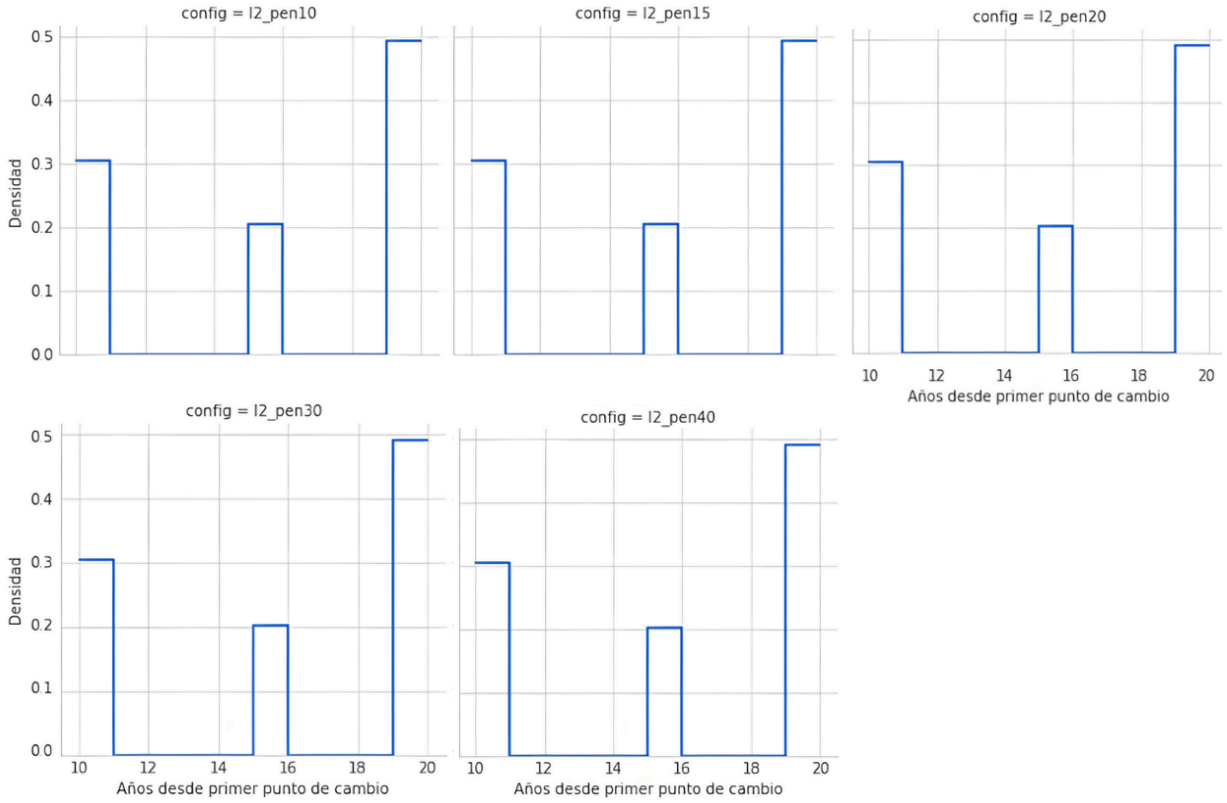


Figura 12. Detección de puntos de cambio en la serie temporal.

5.3. PREPARACIÓN DE DATOS

5.3.1. Datos faltantes

En coherencia con la estructura temporal del fenómeno analizado, se realizó un ajuste en la serie de tiempo para excluir los años con información incompleta o no representativa. Los años 2023 y 2024 se eliminaron debido a la ausencia de datos actualizados en el Banco Mundial, mientras que el periodo 2000–2006 se descartó porque ningún país había publicado registros en GBIF, etapa correspondiente a la fase inicial de consolidación de la infraestructura global de datos. Como resultado, la serie final comprendió un rango de 2007 a 2022, equivalente a dieciséis años de observaciones.

Asimismo, se filtraron los países con menos de diez años de permanencia en la red GBIF, al no presentar trayectorias suficientemente estables para modelar tendencias. Este ajuste excluyó diecinueve países recientes y redujo el conjunto final a 41 países, disminuyendo la ausencia de datos (Fig. 13).

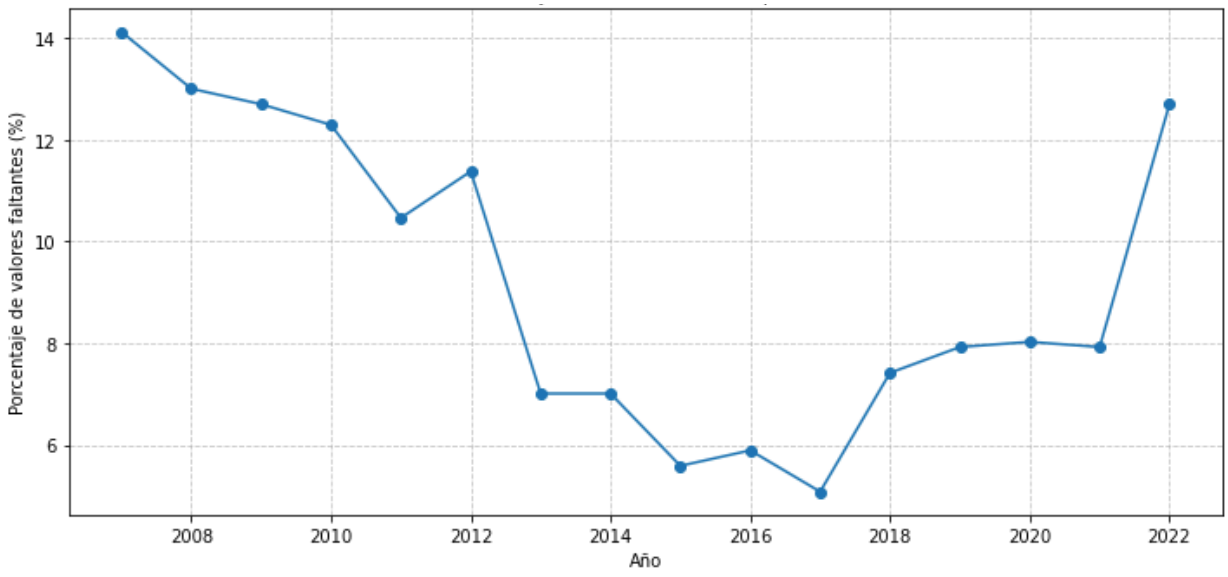


Figura 13. Porcentaje de datos faltantes tras filtros temporales y geográficos.

Adicionalmente, para el tratamiento de los valores faltantes, se implementó un proceso de *imputación múltiple en las variables numéricas*, exceptuando las variables objetivo. Este método comprende tres etapas: (i) generación de múltiples versiones del conjunto de datos completadas mediante modelos probabilísticos; (ii) análisis independiente de cada versión; y (iii) combinación

estadística de los resultados, lo que permite capturar la incertidumbre y producir estimaciones más robustas. La Figura 14 muestra la disminución de valores faltantes tras este proceso, persistiendo únicamente vacíos en las variables objetivo.

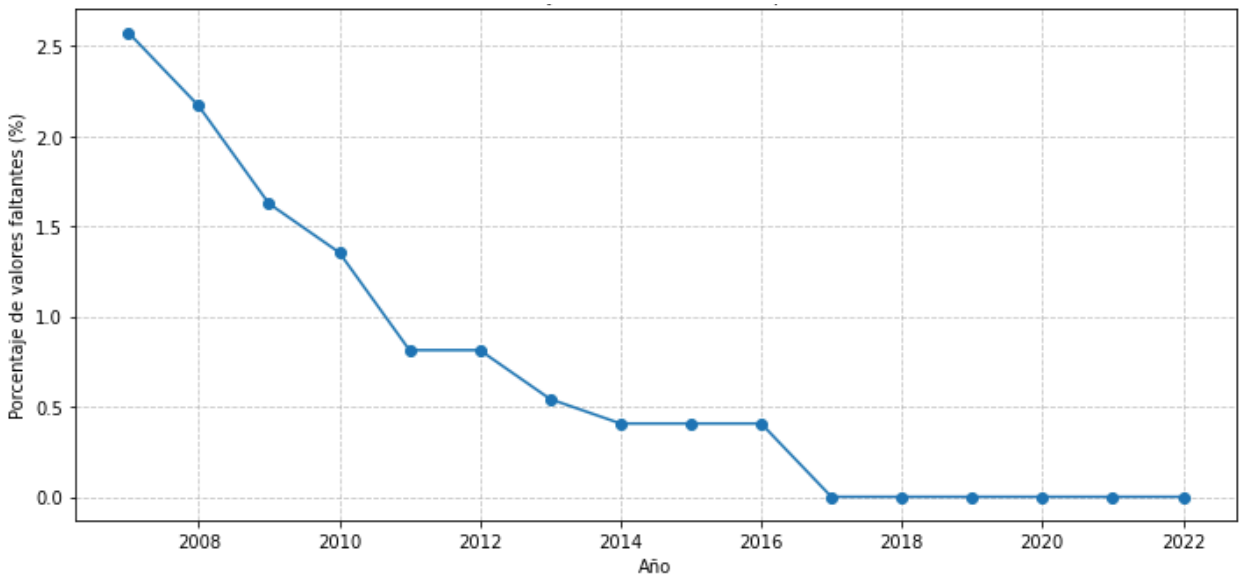


Figura 14. Porcentaje de datos faltantes por año tras imputación de valores.

5.3.2. Multicolinealidad

Inicialmente, para abordar la multicolinealidad se contempló el uso de Análisis de Componentes Principales (PCA) para generar nuevas variables a partir de combinaciones lineales de las originales, proyectando los datos en un subespacio de menor dimensión y reteniendo la mayor parte de la varianza explicada. Sin embargo, estas metodologías suelen generar una pérdida del poder explicativo de las variables, lo que dificulta cumplir el primer objetivo del proyecto sobre la identificación de variables significativas para el crecimiento de los datos. De igual forma, esto también invalida el objetivo específico 4 sobre proyectar escenarios de políticas pues dificulta simular un escenario sobre variables conocidas pues ya no existen de forma independiente en el modelo. En este sentido, se asume la presencia de la multicolinealidad durante el modelado para la selección de los modelos a implementar, por ejemplo, los modelos de árbol (*Random Forest*, XGBoost) que son inherentemente robustos a ella. Adicionalmente, se contempla el uso de regularizadores (como LASSO) para seleccionar las variables más fuertes al momento de proyectar escenarios futuros.

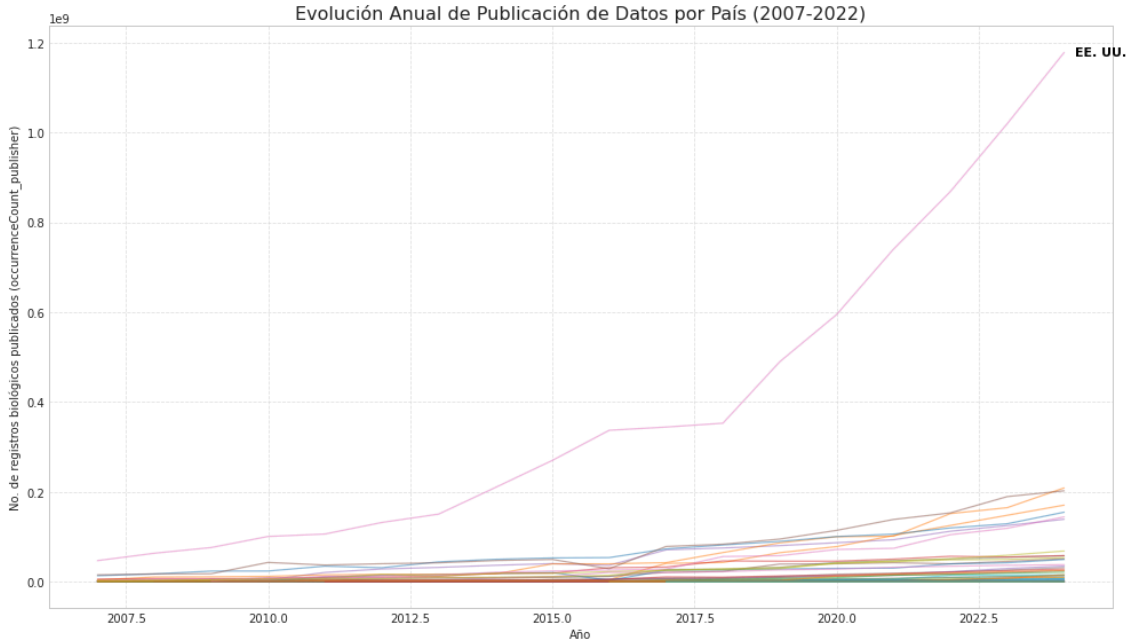
5.3.3. Transformación de variables categóricas

Las variables categóricas se transformaron a formato numérico para garantizar compatibilidad con los algoritmos de modelado. Las variables binarias *gbif_member* y *ogp_membership* se codificaron en valores 0 y 1, mientras que las variables multinomiales *incomeLevel* y *región* se representaron mediante *un Label Encoder*, para evitar la generación de columnas adicionales para cada categoría.

5.3.4. Datos atípicos

Tras una primera evaluación de los modelos (Capítulo 5) e identificar una falta de convergencia en los folds de periodos de tiempo más recientes, se optó por revisar nuevamente los datos y se eliminaron países con comportamientos que son coherentes y reales pero por su comportamiento extremo generan inestabilidad en la predicción. Este fue el caso de EE.UU. que como se observa en la Figura 15 tiene una tendencia de publicación mucho más alta que otros países en la región y a nivel global.

Figura 15. Evolución anual de la publicación de datos de biodiversidad por país.



El conjunto de datos se ordenó cronológicamente por país y año, consolidando una base final de 656 registros y 21 variables, incluyendo la variable objetivo (*occurrenceCount_publisher*) a la

cuál se le aplicó una última transformación con la cual se cambiaba por cero (0) los campos vacíos que representaba que el país aún no se unía a la red o publicaba datos a través de ella. Este dataset depurado y estandarizado constituye la base final para la construcción y evaluación de los modelos predictivos del capítulo siguiente.

Para validar estadísticamente la exclusión del nodo de EE.UU. se realizaron pruebas de Z-score y *Distancia Cook*. Como resultado, este nodo presentó un Z-score de 7.15, situándose a más de siete desviaciones estándar del promedio mundial. Por otro lado, la *Distancia de Cook* alcanzó un valor de 0.882, superando en más de 13 veces el umbral de relevancia global (4/n=0.067). Este hallazgo confirma que la inclusión de Estados Unidos ejercía una influencia sobre los coeficientes del modelo, distorsionando la pendiente de crecimiento hacia escalas de miles de millones de registros que no corresponden a la realidad operativa de los nodos emergentes. En contraste, aunque otros nodos como Reino Unido o Francia mostraron volúmenes superiores al promedio, sus métricas de influencia se mantuvieron bajo los umbrales críticos, permitiendo su permanencia en el estudio.

6 MODELOS ESTADÍSTICOS Y DE APRENDIZAJE AUTOMÁTICO

El presente capítulo aborda un enfoque que integra la inferencia causal a partir de técnicas de modelado de econometría como los Efectos Fijos (EF), y se combinan con técnicas del aprendizaje automático (*Machine Learning*) para la predicción del comportamiento a futuro. Esta metodología responde directamente al cumplimiento de los dos primeros objetivos específicos del proyecto.

En una primera fase, se implementa un análisis de inferencia estadística mediante modelos de datos de panel con *efectos fijos* y técnicas de regularización como LASSO (*Least Absolute Shrinkage and Selection Operator*). El propósito es identificar y validar estructuralmente las variables socioeconómicas y gubernamentales que actúan como determinantes causales en la publicación de datos de biodiversidad, controlando la heterogeneidad no observada entre países y mitigando problemas de multicolinealidad al momento de predecir y seleccionar las variables predictivas (Capítulo 8). Posteriormente, en el enfoque predictivo se detalla la construcción, entrenamiento y optimización de algoritmos avanzados de modelos de Machine Learning (ML) cubriendo modelos como *Random Forest (RF)*, *XGBoost* y redes neuronales recurrentes de Memoria a Corto-Largo Plazo (*LSTM),* capaces de capturar patrones no lineales y dependencias temporales complejas.

6.1. ANÁLISIS DE DETERMINANTES Y SELECCIÓN DE VARIABLES

Con el fin de identificar la relevancia de las variables a incorporar en el modelo y de seleccionar aquellas que son más relevantes para la construcción de escenarios, por su significancia en el crecimiento del volumen de publicación de datos, se aplica un enfoque econométrico basado en *datos de panel*. Se asume que los datos corresponden a una configuración de *datos longitudinales o de panel*, ya que la variable respuesta, orientada a la publicación de datos sobre biodiversidad, es observada varias veces en diferentes momentos para cada país. Adicionalmente, ya que los datos de los países no se observan necesariamente en los mismos tiempos o el mismo número de veces, los datos pueden considerarse desbalanceados [40]. Por este motivo, los datos son susceptibles a analizarse de forma global, a partir del promedio de sus co-variables, o de forma específica, según el comportamiento de cada sujeto, que en este caso, corresponden a los países miembros de la red de GBIF [40]. Una ventaja de realizar el análisis como datos de panel y no como series de tiempo, es precisamente, que estas últimas no manejan de forma adecuada la heterogeneidad [41].

6.1.1. Especificación del modelo de datos de panel (FE vs RE).

Los modelos de datos de panel más usados son los de efectos fijos (FE) y efectos aleatorios (RE), ya que permiten capturar heterogeneidad no observada entre unidades (por ejemplo, diferencias estructurales entre países o nodos) y aprovechar la variación temporal [42]. Los modelos de FE asumen que cada país tiene características propias que no cambian en el tiempo y que pueden estar correlacionadas con las variables explicativas [42], es decir, se asume que hay un efecto individual no observado en las variables que conforman el conjunto de datos de entrada, pero que dichas variables están correlacionadas con aquellas que sí están presentes en los datos. Su especificación formal es:

$$Yit = \beta Xit + \alpha i + \epsilon it$$

Donde αi representa el efecto fijo individual interceptado que captura la heterogeneidad de cada país, eliminando el sesgo por variables omitidas e invariantes.

Al estimar el modelo FE, se obtuvo un R^2 *within* de 0.477, lo que indica que las variables seleccionadas explican el 47.7% de la variabilidad en el crecimiento de datos al interior de los países. Asimismo, el estadístico F (51.33) con un P-value de 0.0000 rechaza contundentemente la hipótesis nula de irrelevancia global de los regresores. Aunque el modelo RE presentó un ajuste global ligeramente superior, su validez depende del supuesto estricto de no correlación entre los efectos individuales y los regresores, lo cual debe verificarse estadísticamente.

6.1.2. Resultados del *Test de Hausman* y validación de hipótesis.

Para elegir estadísticamente entre FE y RE, se aplicó el *Test de Hausman*, el cual compara los estimadores de ambos modelos bajo la hipótesis nula de que *la diferencia en sus coeficientes no es sistemática*. Los resultados revelaron discrepancias significativas. Específicamente, se observó una diferencia de 1.22 unidades en el coeficiente de la variable *'efectividad del gobierno'* y cambios drásticos en el intercepto (-0.95). Estas diferencias confirman que los efectos específicos de cada país (αi) están correlacionados con las variables explicativas. En consecuencia, el modelo RE produce estimaciones sesgadas, mientras que el modelo FE logra aislar el impacto causal neto, eliminado el “ruido” generado por las características del país que son inobservables.

6.1.3. Importancia de las variables.

Una vez validada la estructura de *Efectos Fijos*, se procedió a identificar y confirmar la importancia de las variables mediante una regresión LASSO, utilizada como un mecanismo de confirmación de la selección de características. Cabe resaltar que LASSO se aplicó a nivel global sobre todos los datos y analizando país por país. El análisis arrojó dos hallazgos fundamentales para la estrategia de modelado:

1. El modelo LASSO Global corroboró la importancia crítica de la variable *uso_internet,* ya identificada como significativa por el modelo FE, pero adicionalmente rescató y asignó coeficientes relevantes a variables estructurales como *art_científicos, inscripción_secundaria* e *inscripcion_primaria* (Tabla 4).

   Si bien el modelo FE diluyó la significancia estadística de estas últimas debido a su baja variabilidad temporal, su selección por parte del algoritmo LASSO confirma que poseen un alto poder predictivo global. Esta complementariedad técnica valida la inclusión de estas variables como insumos esenciales para las fases de predicción, independientemente de las limitaciones de estimación de cada método.

2. Al intentar aplicar LASSO a nivel individual por país, se observó que la escasez de datos históricos (N < 16 años) impedía contar con resultados estables. Esto justifica la decisión metodológica de utilizar modelos globales (que aprenden patrones compartidos entre países) en las fases subsiguientes de predicción, en lugar de entrenar modelos aislados para cada nodo.

El *pipeline* implementado para el test de los datos de Panel con FE Y LASSO se encuentra disponible en el repositorio en GitHub a través del siguiente enlace:

- Análisis de datos de panel: https://github.com/rortizgeo/Maestria_CD_Proyecto-Aplicado/blob/main/2a_Panel_Datasets_test.ipynb

Tabla 4. Coeficientes estimados y selección de variables mediante Regresión LASSO global.

Variable	Coeficiente
uso_internet	2.8356712
art_cientificos	0.9974186
inscripcion_secundaria	0.9331545
inscripcion_primaria	0.4265470
gasto_RD_pib	0.0000001

efectividad_gobierno	-0.0000004
pib_per_capita	-0.0000007
inscripcion_terciaria	0.0000008
gasto_educacion_gobierno	0.0000009
gasto_educacion_pib	-0.00000010
investigadores_RD	0.000000

La aplicación de la regresión LASSO Global, tras optimizar el parámetro de regularización con un α de 0.259, operó como un filtro de selección de características, reduciendo a cero los coeficientes de variables redundantes. Los resultados (Tabla 4) destacan de manera contundente la variable de *uso_internet* como el predictor dominante con un coeficiente de 2.84, sugiriendo que la infraestructura digital es el *driver* más fuerte de la publicación de datos, magnitud que supera casi por el triple a cualquier otra variable.

Adicionalmente, el modelo rescató la importancia de *art_cientificos* (1.00) e *inscripcion_secundaria* (0.93) como determinantes secundarios clave. Es crucial notar que variables como *efectividad_gobierno* y *gasto_educacion_gobierno* fueron penalizadas hasta prácticamente cero. Esto no implica que no sean importantes, sino que su poder explicativo ya ha sido "absorbido" por las variables dominantes debido a la alta multicolinealidad. En términos predictivos, LASSO sugiere que un país con alto acceso a internet y producción científica ya contiene la información implícita de una buena gestión gubernamental, por lo que el modelo descarta esta última para ganar eficiencia y evitar el sobreajuste.

6.2. PREPARACIÓN Y PREPROCESAMIENTO DE DATOS PARA EL APRENDIZAJE AUTOMÁTICO

Para el desarrollo del presente trabajo, y en concordancia con los fundamentos expuestos en el marco teórico (Ver 3.1.4.), se seleccionó una terna de modelos que representan el estado del arte en aprendizaje supervisado y redes neuronales, elegidos estratégicamente por su complementariedad:

- *Modelos tabulares (Random Forest y XGBoost)*: Se optó por estos algoritmos debido a su robustez frente a la multicolinealidad (identificada durante el análisis exploratorio de los datos) y su capacidad para manejar relaciones no lineales sin necesidad de transformaciones previas de las variables independientes. Su arquitectura basada en árboles permite capturar interacciones complejas entre variables que modelos lineales tradicionales podrían ignorar.

- Modelos de aprendizaje profundo secuencial (Memoria a Corto-Largo Plazo - LSTM): Dada la estructura de panel de los datos, donde el orden cronológico de las observaciones es fundamental, se integró una red LSTM. Esta elección se sustenta en su capacidad nativa para modelar dependencias temporales de largo plazo y capturar la persistencia de los fenómenos sociales y biológicos, superando la limitación de los modelos tabulares que requieren una ingeniería de características(*lags y rolling windows*) para interpretar el tiempo.

Tras la identificación de los determinantes estructurales en la fase anterior (Sección 5.1), el *pipeline* de datos se enfocó en transformar el conjunto de datos de panel en una estructura compatible con los algoritmos de aprendizaje automático seleccionados. Este proceso es crítico para permitir que modelos tabulares capturen la dinámica temporal y que los modelos de aprendizaje profundo procesen la información en formato de tensores.

El *pipeline* implementado para el desarrollo de los modelos que se presentará en los siguientes capítulos se encuentra disponible en el repositorio en GitHub a través del siguiente enlace:

Flujo de trabajo para análisis de datos de panel:
https://github.com/rortizgeo/Maestria_CD_Proyecto-Aplicado/blob/main/3_PA_Workflow_Paneldata.ipynb

6.2.1. Transformación de la variable objetivo

Se aplicó de forma universal la transformación logarítmica $log(1 + x)$ a la variable objetivo *occurrenceCount_publisher*. Esto fue implementado para todos los modelos del *pipeline* (*Random Forest, XGBoost* y LSTM), para cumplir tres propósitos:

- Reducir la heterocedasticidad intrínseca de los datos de panel [43].
- Evitar que países con volúmenes de publicación excepcionalmente altos, sesguen el entrenamiento por la presencia de datos extremos (no atípicos), como es el caso de Estados Unidos (ver Figura 15).
- Asegurar que al revertir la transformación mediante la función inversa $(y' = ln(1 + x))$, los resultados nunca arrojen valores negativos de biodiversidad.

6.2.2. Ingeniería de características temporales y datos faltantes

El proceso de ingeniería de características se diseñó para capturar tanto las dinámicas temporales como las diferencias estructurales entre países. Dado que los modelos considerados (tabulares y secuenciales) requieren distintas representaciones de los datos, se adoptó un enfoque diferenciado según el tipo de modelo (Tabla 5).

- Modelos tabulares (*Random Forest y XGBoost*): Para estos modelos se generaron explícitamente rezagos (*lags*) y ventanas móviles (*rolling windows*) sobre las variables explicativas, incluyendo variables explicativas, indicadores socioeconómicos y la propia variable objetivo. Los rezagos permiten incorporar la información de años previos como predictores, mientras que las ventanas móviles (aplicadas a 3 y 5 años) resumen la evolución reciente de cada variable. Estas transformaciones convierten la dependencia temporal en atributos tabulares que los algoritmos de árboles pueden procesar, sin romper la estructura temporal de la serie. Aunque los árboles son menos sensibles a la escala, el escalado se aplicó de forma general para mantener la consistencia del pipeline y facilitar la comparación.
- Modelos secuenciales (*LSTM*): En el caso de la red LSTM, no se utilizaron los rezagos ni ventanas móviles explícitas, ya que este tipo de red está diseñado para aprender dependencias temporales directamente de las secuencias. La red neuronal LSTM requiere una preparación diferenciada para aprovechar su capacidad de memoria a corto y largo plazo, por esta razón, los datos se estructuraron en secuencias de longitud $lookback = 3$. Esto significa que para predecir el año t, la red analiza simultáneamente los patrones de los años $t-1, t-2$ y $t-3$. A diferencia de los árboles de decisión, la red LSTM es sensible a la magnitud de las entradas, por lo que se utilizó el método *StandardScaler* de *scikit Learn* [44] para normalizar las características y la variable objetivo a una distribución con media cero y desviación estándar igual a uno

6.2.3. Tratamiento de datos faltantes mediante Imputación Iterativa

El uso de rezagos (*lags*) y ventanas móviles (*rolling windows*), genera datos faltantes al calcular cada rezago en los datos. En estos casos se optó por la eliminación de las filas previo al entrenamiento. Sin embargo, considerando que los indicadores socioeconómicos luego de la estructuración de los datos seguían presentando algunos vacíos para ciertos países (panel desbalanceado), se implementó un esquema de Imputación iterativa [45]. En lugar de eliminar filas con datos faltantes que, en el caso de los datos de panel, rompe la continuidad temporal, la

imputación iterativa permitió usar esas observaciones, manteniendo la integridad histórica del país. Además, se toma la decisión de usar este método porque se tiene en cuenta que el conjunto de datos no es necesariamente grande y que la serie temporal de cada país es corta, más aún, si el ingreso del país a la red de GBIF es inferior a 10 años.

A diferencia de una imputación simple por la media o mediana, este método estima los datos faltantes basándose en las demás variables, manteniendo así las relaciones originales entre ellas [46]. Este proceso se aplicó directamente en el proceso de entrenamiento, como se explicará más adelante en este capítulo

Tabla 5. Descripción del manejo de datos para cada tipo de modelo.

Proceso	Modelos tabulares (RF / XGB)	Modelos secuenciales (LSTM)
Variable objetivo	$log(1 + x)$	$log(1 + x)$ + Escalado estándar
Tratamiento de datos vacíos	Imputación iterativa	Imputación iterativa
Dependencia temporal	Rezagos explícitos (*Lags 1, 3, 5*) y Medias móviles (3, 5)	*Look-back* (ventana de 3 pasos) en formato de tensor 3D
Escalado de variables	StandardScaler	StandardScaler
Resultado final	Matriz tabular de 87 dimensiones	Secuencias multivariadas (3 pasos de tiempo)

Es importante precisar que, si bien en la sección 5.1, se identificó un subconjunto de variables con significancia causal mediante LASSO y FE, los modelos de aprendizaje automático se entrenaron sobre la totalidad del conjunto de datos expandido por los rezagos y ventanas móviles (87 dimensiones). Esta decisión se fundamenta en la capacidad de los algoritmos de ensamble y redes recurrentes para gestionar la alta dimensionalidad y capturar interacciones no lineales que podrían ser omitidas en un proceso de eliminación de variables más rígido. No obstante, las variables priorizadas en la fase anterior conservan su importancia como determinantes durante la simulación de escenarios de política pública para Colombia, vinculando así la robustez predictiva del *Machine Learning* con la fundamentación estructural de la econometría

6.3. MODELADO Y ESTRATEGIA DE VALIDACIÓN CRUZADA

6.3.1. Estructuración de la matriz de aprendizaje y tensores secuenciales

Debido a la naturaleza de los modelos a implementar y la ingeniería de características explicada anteriormente, se generaron dos estrategias para la estructuración de los datos de entrenamiento de cada modelo, como se describen a continuación.

6.3.1.1. Definición del espacio predictor para los modelos tabulares

Para los modelos *Random Forest* y *XGBoost*, se requiere una estructura bidimensional donde cada fila representa un estado temporal único. En el código se establecieron los siguientes pasos:

- Se filtran exclusivamente las columnas de tipo numérico, garantizando la compatibilidad con los algoritmos de árboles.
- Se definen la variable *cols_exclude*, que incluyen variables como *year, countryCode*, *iso2*, y los nombres de los países. La exclusión de estas variables evita que el modelo aprenda sesgos geográficos o temporales específicos y se generaliza basándose únicamente en las variables socioeconómicas y de biodiversidad.

Segmentación de X y y: El resultado es una matriz X de 87 dimensiones (features) y una variable objetivo transformada.

6.3.1.2. Generación de tensores secuenciales para LSTM

El modelado con redes recurrentes LSTM exige una transformación profunda de los datos hacia una estructura de tensores 3D con la forma (muestras, pasos de tiempo, características).

Para lograr esto, se desarrolló la función *create_lstm_sequences_global*, la cual sigue este procedimiento:

- El algoritmo recorre cada país de forma independiente, asegurando que la secuencia de un país no se mezcle con otro.
- Se definió un parámetro $look\ back\ =\ 3$. Matemáticamente, para predecir el valor en el año t, la función captura el bloque de información de los años $t\ -\ 3.\ t\ -\ 2,\ t\ -\ 1.$
- Los datos se transforman a *float32*, el estándar de precisión requerido por los motores de cálculo de *TensorFlow* [47] para optimizar el uso de memoria y la velocidad de procesamiento en la red neuronal.

El resultado final son los conjuntos *X_seq* y *y_seq*, donde cada observación no es una fila, sino una ventana de tiempo que contiene la evolución multivariada de los determinantes de biodiversidad.

6.3.2 Validación cruzada de ventana expandida

La evaluación de modelos predictivos en datos con componentes temporales requiere un enfoque distinto al de los problemas de clasificación o regresión tradicionales. En los métodos convencionales, como la validación cruzada por pliegues (*K-fold*), los datos se dividen aleatoriamente, lo que asume que las observaciones son independientes entre sí. Sin embargo, en el estudio de la biodiversidad a través del tiempo, esta independencia no existe debido a la autocorrelación temporal, pues el estado de un sistema en un año determinado depende del crecimiento de la publicación de datos en años anteriores.

Para abordar esta particularidad, se implementó una estrategia de *validación cruzada de ventana expandida* mediante la técnica *TimeSeriesSplit* [48] con cinco particiones o folds. A diferencia de un método tradicional, que dividiría los datos, por ejemplo, en un 80% de entrenamiento y 20% de prueba, esta técnica respeta el orden cronológico de los registros entrenándose con el pasado para predecir el futuro [48].

6.3.2.1. Funcionamiento del proceso de validación

La implementación en el flujo de trabajo implementado en *Python*, sigue una arquitectura de aprendizaje incremental. En lugar de dividir el conjunto de datos en bloques aislados de igual tamaño, el algoritmo expande progresivamente los datos de entrenamiento del modelo:

- Primer ciclo: El modelo se entrena con la etapa inicial de la serie equivalente a aproximadamente el primer 16% de los datos, y se evalúa en el periodo inmediatamente posterior.
- Ciclos sucesivos: En cada nueva iteración, el conjunto de entrenamiento incorpora los datos de prueba del ciclo anterior. Por ejemplo, en el último ciclo, el modelo utiliza aproximadamente el 84% de la historia disponible para proyectar el comportamiento del último 16% del tiempo registrado.

Este enfoque permite verificar si el modelo mantiene su precisión a medida que las condiciones históricas evolucionan, proporcionando una métrica de error confiable para proyecciones a largo plazo.

6.3.2.2. Integración del preprocesamiento y prevención de fuga de datos

Un aspecto crítico de este *pipeline* es la localización del preprocesamiento. El rigor metodológico exige que cualquier transformación de los datos se base exclusivamente en la información disponible en el momento del entrenamiento, evitando una "fuga de datos" (*data leakage*) [49].

Dentro de cada ciclo de validación, el código ejecuta de forma aislada los siguientes pasos:

- Se crean los subconjuntos de entrenamiento y prueba respetando la ventana expandida.
- El algoritmo de imputación iterativa (*IterativeImputer*) y el escalador de variables (*StandardScaler*) calculan sus parámetros (como la media o el valor para rellenar vacíos) basándose únicamente en el conjunto de entrenamiento de ese ciclo específico.
- Aplicación del mismo preprocesamiento, usando los parámetros aprendidos del entrenamiento, al conjunto de prueba mediante la función *transform*.

6.4. CONFIGURACIÓN DE MODELOS Y OPTIMIZACIÓN DE HIPERPARÁMETROS

Para garantizar que los modelos alcancen su máximo potencial predictivo y eviten el sobreajuste, se implementó una estrategia de validación cruzada anidada *(Nested Cross-Validation)* [50]. Dentro de cada pliegue (*fold*) de la *validación cruzada* principal, se ejecutó una búsqueda automatizada mediante *RandomizedSearchCV* [51]. Esta búsqueda interna utilizó también una validación cruzada de cinco pliegues ($cv = 5$) para seleccionar la configuración de parámetros que minimizara el Error Absoluto Medio (MAE).

6.4.1. Configuración de búsqueda de hiperparámetros

La siguiente tabla detalla los espacios de búsqueda configurados en el script. Estos valores fueron seleccionados para equilibrar la complejidad del modelo con la capacidad de generalización en series de tiempo cortas:

Tabla 6. Configuración del espacio de búsqueda de hiperparámetros.

Modelo	Parámetro	Valor/Rango búsqueda	Descripción
Random Forest	*n_estimators*	[100, 300, 500, 700]	Cantidad de árboles para promediar predicciones.
	max_depth	[10, 20, 40, None]	Límite de profundidad de los árboles.

	max_features	[0.7, 0.9, 'sqrt']	Cantidad de variables a evaluar en cada división.
	min_samples_leaf	[1, 2, 5, 10]	Cantidad mínima de muestras en un nodo terminal.
XGBoost	*n_estimators*	[100, 300, 500]	Ciclos de entrenamiento del algoritmo de boosting.
	learning_rate	[0.01, 0.05, 0.1]	Tasa de aprendizaje que controla la corrección de errores.
	max_depth	[3, 5, 7, 9]	Complejidad de los árboles individuales.
	reg_alpha (L1)	[0, 0.01]	Penalización para fomentar la simplicidad del modelo.
	subsample	[0.7, 0.9]	Fracción de datos usada para entrenar cada árbol.
LSTM	*units*	[32, 50, 64]	Representa la "*capacidad de memoria*". Un valor mayor (64) permite capturar relaciones más complejas pero aumenta el riesgo de sobreajuste (overfitting); un valor menor (32) es más robusto pero puede ignorar matices sutiles.
	learning_rate	[0.001, 0.005]	Controla qué tan rápido se actualizan los pesos de la red. Un *valor* muy alto puede hacer que el modelo nunca converja, mientras que uno bajo requiere más épocas para aprender.
	activation	['relu', 'tanh']	Función de transferencia en la capa recurrente. La función *tanh* es el estándar para LSTM porque mantiene los valores entre -1 y 1, ideal para la celda de memoria. La opción *relu* ayuda a que el

			entrenamiento sea más rápido en redes profundas.
	dropout	[0.2, 0.3]	Es una técnica de regularización que "*apaga*" aleatoriamente un porcentaje de neuronas durante el entrenamiento. Esto obliga a la red a no depender de una sola variable y mejora la generalización en datos nuevos.
	batch_size	16, 32	Define cuántas muestras usa la red antes de actualizar sus errores.

6.4.2. Enfoque específico para LSTM

Se definieron límites superiores de 20 y 30 épocas para las distintas configuraciones de la red LSTM. No obstante, el número efectivo de ciclos de entrenamiento fue gestionado dinámicamente mediante la técnica de *Parada Temprana* (*EarlyStopping*) [52], configurada con una paciencia de 5 épocas. Esta herramienta detiene el entrenamiento automáticamente si el error en el subconjunto de validación interna no muestra mejoría, restaurando los pesos del mejor modelo obtenido hasta ese punto.

A diferencia de los modelos de árboles, donde la búsqueda se hace por validación cruzada interna ($cv = 5$), la LSTM utiliza un *Split* de validación fijo (X_v, y_v). Una vez finalizado el entrenamiento de una configuración específica, el modelo realiza una predicción sobre el conjunto de validación: `mae = mean_absolute_error(y_v, preds)`

Para la búsqueda de hiperparámetros en los modelos tabulares y secuenciales, se utiliza el *Error Absoluto Medio* (MAE) porque mide la magnitud promedio de los errores en las unidades logarítmicas de los registros, siendo menos sensible a valores atípicos que el error cuadrático.

El código para la implementación de los modelos se encuentra disponible en e GitHub:

- https://github.com/rortizgeo/Maestria_CD_Proyecto-Aplicado/blob/main/3_PA_Workflow_Paneldata.ipynb

7 EVALUACIÓN COMPARATIVA DE DESEMPEÑO

Este capítulo desarrolló el tercer objetivo específico del proyecto, orientado a evaluar y comparar el desempeño de los modelos implementados para predecir el crecimiento en la publicación de datos de biodiversidad en el nodo nacional de GBIF (SiB Colombia). El proceso de evaluación se diseñó para contrastar la efectividad de los modelos de ensamble tabular (*Random Forest* y *XGBoost*) frente a la arquitectura secuencial (LSTM). La comparación se basa, principalmente, en la capacidad de los modelos para capturar la varianza de los datos de panel y minimizar el error absoluto en las proyecciones anuales.

7.1. DEFINICIÓN DE LAS MÉTRICAS DE DESEMPEÑO

Para asegurar que los resultados sean interpretables en la magnitud real de los fenómenos de biodiversidad, todas las métricas fueron calculadas revirtiendo la transformación logarítmica inicial mediante la función $f - 1(y) = ey - 1\ (expm1)$. Las métricas utilizadas fueron: i) Error Absoluto Medio (MAE); Error Cuadrático Medio (RMSE), Coeficiente de determinación (R^2) y el Error Porcentual Absoluto Medio Simétrico (SMAPE).

7.2. ANÁLISIS COMPARATIVO DE RESULTADOS

Tras la ejecución de los cinco pliegues de validación cruzada, los resultados (Tabla 7) revelan una mayor estabilidad en los modelos de ensamble tabular sobre la red neuronal secuencial para esta estructura de datos, sobre todo en el *fold 5*, que es relevante por ser el pliegue que representa el horizonte de evaluación más reciente y se entrenó con la mayor cantidad de datos.

Tabla 7. Rendimiento de los modelos con mejor optimización por fold.

Modelo	Fold	MAE	RMSE	R^2	MAPE	SMAPE
RandomForest	1	7.705.197,65	10.419.874,27	0,46	1.054,16	83,96%
RandomForest	2	7.869.939,01	17.866.775,69	0,78	46,57	38,02%
RandomForest	3	3.087.252,26	5.880.395,44	0,95	63,79	43,41%
RandomForest	4	1.251.780,12	1.715.579,59	0,98	54,97	38,50%
RandomForest	5	5.611.122,85	12.465.234,42	0,94	31,30	35,60%
XGBoost	1	8.726.740,21	11.699.562,10	0,32	1.133,25	114,59%
XGBoost	2	8.152.770,59	17.863.622,01	0,78	41,13	36,68%
XGBoost	3	3.634.514,73	8.156.177,23	0,91	63,58	44,56%
XGBoost	4	1.381.068,82	1.980.196,26	0,98	83,91	46,45%

XGBoost	5	6.204.400,16	13.138.950,74	0,93	33,03	44,44%
LSTM*	1	8.919.166,00	18.865.162,20	0,69	2,03E+10	50,75%
LSTM*	2	7.983.145,00	17.278.504,96	0,74	1,23E+10	48,96%
LSTM*	3	12.147.585,00	21.671.543,89	0,59	3,81E+13	76,27%
LSTM*	4	14.668.214,00	34.127.472,82	-0,01	9,18E+13	77,80%
LSTM*	5	12.441.073,00	22.936.543,79	0,54	7,80E+13	77,57%

*Se observa una inestabilidad numérica en el MAPE del modelo LSTM, producto de valores reales cercanos a cero en la variable objetivo tras la transformación, debido a que los primeros años de varios nodos no hay publicación. Estos valores distorsionan por completo la métrica global. Por consiguiente, se prioriza el SMAPE como la métrica de error relativo válida para la comparación inter-modelos.

El modelo *Random Forest* destaca con un R^2 promedio de 0.8226 (Tabla 8), lo que implica que logra explicar más del 82% de la variabilidad observada en la publicación de registros a nivel global. Por su parte, la red LSTM presenta un desempeño moderado (R^2=0.5127), sugiriendo que, para el volumen y la estructura actual del panel, la captura de interacciones no lineales mediante árboles de decisión resulta más efectiva que el modelado de dependencias temporales profundas.

Tras ejecutar la optimización mediante *RandomizedSearchCV* bajo un esquema de validación cruzada anidada , se determinó que la configuración óptima para el modelo Random Forest fue: *n_estimators: 700, min_samples_leaf: 1, max_features: sqrt y max_depth: 40.* Estos resultados evidencian una selección de parámetros orientada a la robustez y la precisión granular: el elevado número de árboles (*n_estimators*) asegura la estabilidad de las predicciones al promediar el error y reducir la varianza global. Por su parte, la profundidad de 40 niveles (*max_depth*) permite capturar interacciones no lineales complejas entre las variables socioeconómicas y la publicación de datos, mientras que un *min_samples_leaf*: 1 otorga al modelo la flexibilidad necesaria para alcanzar su máxima especificidad en los nodos terminales. Esta configuración resulta fundamental para la Etapa 2 del marco híbrido como se explicará en el próximo capítulo 7.

En el último pliegue, el modelo *Random Forest* alcanza un nivel de precisión destacado con un R^2 de 0.94, demostrando una robustez significativa para capturar las tendencias de publicación más recientes. Esto se logró a partir de la depuración de datos extremos en el dataset, antes de la modelación. En análisis preliminares, los modelos tabulares presentaban un colapso en su capacidad explicativa en el *Fold 5* (con R^2 cercanos a 0.10). Cabe resaltar que tras eliminar los datos extremos de publicación, específicamente los datos de publicación de EE.UU., el modelo

Random Forest mejoró alcanzando los valores antes mencionados.

Tabla 8. Resumen de rendimiento promedio de los modelos implementados.

Modelo	MAE (Promedio)	RMSE (Promedio)	R^2 (Promedio)	Estabilidad (R^2 Std)
Random Forest	5.105.058,38	9.669.571,88	0.8226	±0.2157
XGBoost	5.619.898,90	10.567.701,67	0.7831	±0.2672
LSTM	11.231.836,60	22.975.845,53	0.5127	±0.3014

Contrario a lo esperado, la red LSTM no se benefició de igual manera que los modelos tabulares ante la remoción de los datos extremos. Si bien su MAE promedio bajó a 11.2 millones, su R^2 promedio descendió a 0.51, llegando incluso a mostrar una capacidad predictiva casi nula en algunos Folds (R^2=−0.0081).

Este comportamiento sugiere que la arquitectura LSTM, aunque robusta para capturar tendencias generales de largo plazo con ruido, requiere de volúmenes de datos masivos (como los proporcionados por el outlier eliminado) para ajustar sus pesos de manera eficiente en secuencias cortas. Ante un dataset más homogéneo y depurado, los modelos tabulares con ingeniería de características (rezagos y ventanas móviles) demuestran ser superiores para capturar la varianza local de los países.

Con base en la evidencia recolectada en las métricas de evaluación, el modelo con mejor desempeño fue *Random Forest* por las siguientes razones:

- Presenta el menor MAE promedio (5.1 millones de registros), que aunque es alto, asegura proyecciones más cercanas a la realidad.
- Su alto R^2 (0.82 promedio y >0.93 en el último *fold*) garantiza que la mayoría de los factores determinantes han sido integrados correctamente en la estructura del bosque de decisión.
- A diferencia de la red LSTM, los modelos de ensamble tabular mostraron una mayor estabilidad y una mejor respuesta ante las dinámicas de los años más recientes, lo cual es vital para el cumplimiento del objetivo de proyectar escenarios hacia el año 2030.

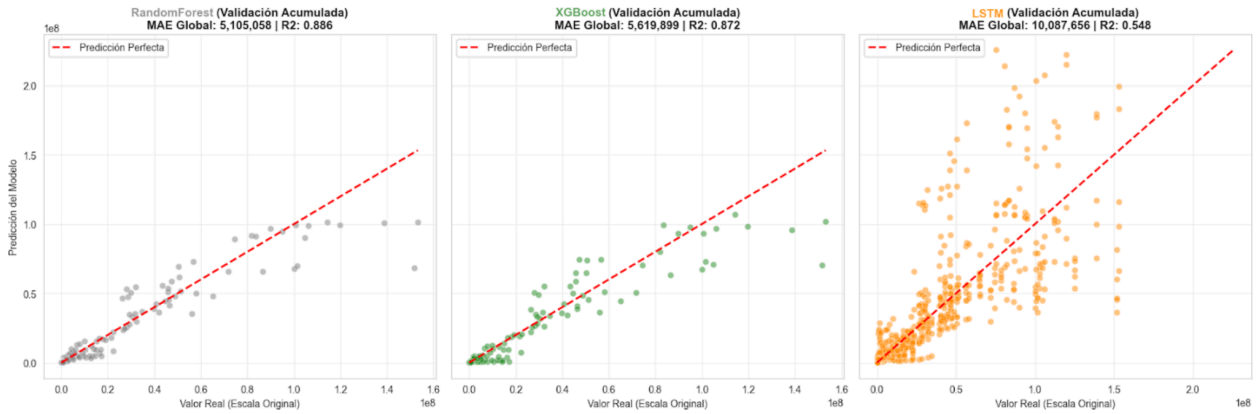


Figura 16. Comparación de los resultados de predicción para cada uno de los 5 folds en los 3 modelos (Random Forest, XGBoost y LSTM). *Nota:* La diferencia entre los R^2 de la Tabla 8 y los gráficos se debe a que la tabla toma el R^2 de los 5 *fold* por separado y saca un promedio simple, mientras que el gráfico toma todas las predicciones y las compara contra todos los valores reales juntos, calculando un solo R^2 sobre esa nube de puntos masiva.

Tras graficar cada una de las predicciones que hizo el modelo cuando se enfrentó a datos desconocidos en los 5 *folds* (Fig. 16) se realizó una validación visual de las predicciones acumuladas, observando que la nube de puntos del modelo Random Forest (izquierda) presenta la dispersión más compacta y alineada a la diagonal de "predicción perfecta" (línea roja), evitando la alta varianza dispersa observada en LSTM.

7.3. LIMITACIONES DEL MODELO

Al realizar un análisis de la incidencia de las variables en la predicción a través de SHAP (*SHapley Additive exPlanations*) [53] para explicar los resultados del modelo (Fig. 17), se observa que la variable (*occurrenceCount_publisher_lag1*) es la que más aporta al resultado del modelo, mientras que las variables socioeconómicas tienen un impacto casi nulo. En particular, al modificar variables como el gasto en investigación y desarrollo, la efectividad gubernamental o el acceso a internet, o las sugeridas por los modelos de efectos fijos, el modelo tiende a mantener la trayectoria histórica de los registros publicados, reduciendo la sensibilidad a cambios en dichas variables.

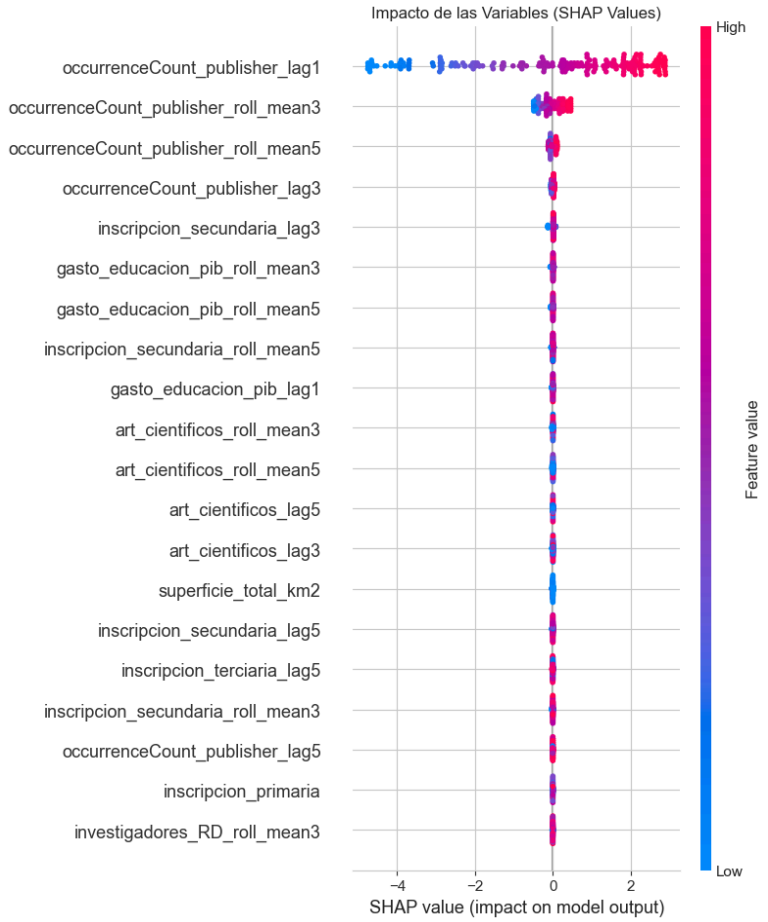


Figura 17. Gráfico resultante a partir de la validación con SHAP del modelo Random Forest ejecutado haciendo uso de las variables objetivo con sus respectivos rezagos y ventanas móviles.

En ese sentido, aunque el modelo *Random Forest* presenta los mejores resultados para modelar el crecimiento de los datos incluyendo tanto la inercia del crecimiento de los datos (*occurrenceCount_publisher*) como las variables exógenas (socioeconómicas), para el cumplimiento de los objetivos del presente proyecto, se hace necesario considerar una aproximación que permita capturar el efecto de los modelos sin la inferencia de los los rezagos y ventanas móviles de la variable objetivo.

En ese sentido, se re-entrenaron nuevamente los modelos de *Machine Learning* eliminando los rezagos y ventanas móviles de la variables objetivo (*ocurrenceCount_publisher*) para observar el impacto real de las variables socioeconómicas en el crecimiento del volumen de datos de las redes de GBIF. En la tabla 9 y figura 18 se muestran los resultados. Se destaca que estos cambios

afectan sobre todo a los modelos tabulares (*Random Forest* y XGBoost) ya que LSTM no usa rezagos, pues cuenta con su propia definición de estructura secuencial.

Tabla 9. Resumen de rendimiento promedio de los modelos sin rezagos en la variable objetivo.

Modelo	MAE (Promedio)	RMSE (Promedio)	R^2 (Promedio)	Estabilidad (R^2 Std)
Random Forest	13.086.380,77	20.395.001,22	0.4411	±0.2596
XGBoost	14.704.846,19	23.490.010,68	-0,031	±0.7547
LSTM	11.231.836,60	22.975.845,53	0.5127	±0.3014

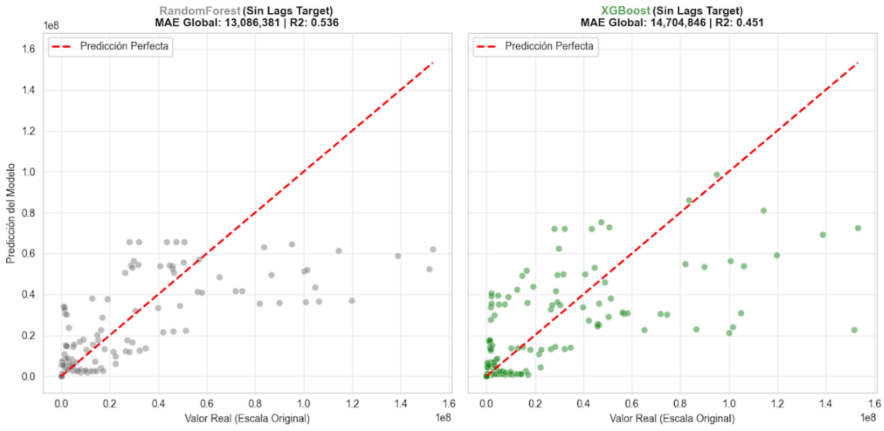


Figura 18. Comparación de los resultados de predicción para cada uno de los 5 folds en los 2 modelos (Random Forest, XGBoost) sin incluir rezagos o ventanas móviles en la variable objetivo. *Nota:* La diferencia entre los R^2 de la Tabla 9 y los gráficos se debe a que la tabla toma el R^2 de los 5 *fold* por separado y saca un promedio simple, mientras que el gráfico toma todas las predicciones y las compara contra todos los valores reales juntos, calculando un solo R^2 sobre esa nube de puntos masiva.

Como se esperaba, el rendimiento de los modelos tabulares baja considerablemente por que ya no cuentan con el aporte de la variable objetivo. Por lo que el R^2 de *Random Forest* cae a cerca de 0.44 y *XGBoost* a valores negativos debido a valores extremadamente bajos de predicción en los primeros folds. Sin embargo, estos resultados están indicando que casi la mitad (44%) de la variabilidad en la publicación de datos de biodiversidad puede explicarse por factores estructurales (I+D, internet, investigadores, etc.), sin mirar el pasado. Lo que potencia su aplicación en un *modelo híbrido,* discutido en el capítulo siguiente.

8 PROYECCIONES ESTRATÉGICAS DE CRECIMIENTO

8.1. IMPLEMENTACIÓN DEL MODELO HÍBRIDO

Cómo se discutió en el capítulo anterior, se consideró necesario ajustar el enfoque metodológico para separar la dinámica inercial del crecimiento histórico de los efectos estructurales asociados a variables socioeconómicas. Esto con el fin de dar cumplimiento al objetivo 4 del proyecto y poder generar escenarios de predicción de crecimiento en la publicación del nodos del SiB Colombia basado en las variables socioeconómicas. Para abordar esta limitación, se adoptó un enfoque de *modelado híbrido* compuesto por dos etapas. En la primera etapa, se ajustó un modelo de regresión utilizando exclusivamente variables exógenas socioeconómicas y gubernamentales, con el fin de capturar relaciones estructurales lineales y estables. En la segunda etapa, se entrenó el modelo de *Machine Learning (basado en los resultados del capítulo anterior)* sobre los residuos del modelo que capturara los patrones no lineales, interacciones complejas y efectos no explicados por la estructura lineal inicial.

8.1.1. Etapa 1 del modelo híbrido

Siguiendo el principio de parsimonia para este nuevo enfoque, se evaluó la implementación de un modelo de regresión simple [54]que explicara el comportamiento de la publicación de datos en el Nodo del SiB Colombia, sin embargo, este generó una curva de publicación de registros que crecía de forma exponencial, llegando a los 170 millones de registros biológicos publicados para Colombia al 2030, lo cuál es inverosímil de acuerdo al contexto de los datos y conocimiento del comportamiento de publicación del nodo. En ese sentido, se optó por la implementación de un modelo de regresión *Ridge*, que es una variante de la regresión lineal que utiliza una técnica llamada *regularización L2* [55]. A diferencia de una regresión lineal simple (Mínimos Cuadrados Ordinarios o OLS), que solo busca minimizar la suma de los errores al cuadrado, el modelo *Ridge* añadió una penalización a la función de pérdida. Esta penalización es proporcional al cuadrado de la magnitud de los coeficientes de las variables [56] (en este caso, los valores pasados del número de registros publicados).

Formalmente, este modelo se expresa como:

$$\hat{y}\_t\hat{}(estructura) \;=\; X\_t\,\beta$$

Sujeto a la penalización L2:

$$min\_\beta \{ \sum\{t = 1\}\^{}\{T\} (y\_t - X\_t \beta)\^{}2 + \alpha \sum\{j = 1\}\^{}\{p\} \beta\_j\^{}2 \}$$

Donde $\hat{y}\_t$ representa el número de registros de biodiversidad publicados en el año t, X_t corresponde al vector de variables socioeconómicas y gubernamentales observadas, β es el vector de coeficientes a estimar y α es el parámetro de regularización que controla el grado de penalización.

En el modelo *Ridge* se implementó un parámetro α (Alpha) de penalización. Con su implementación con un valor 7.5, se logró un comportamiento esperado y se redujo la sobre-predicción de registros, con valores cercanos a los 60 millones de registros biológicos. Para la implementación del modelo, las variables explicativas fueron estandarizadas para garantizar que la penalización se ejecutara de forma homogénea sobre todos los coeficientes.

Este modelo fue entrenado solo con la variable objetivo (*occurrenceCount_publisher*) y sus respectivos rezagos ($lag\,1$, $lag2$, $lag3$) y ventanas móviles ($rollmean3$ *y* $rollmean5$) de forma que se capturara adecuadamente la dependencia del comportamiento de los años previos para entender el comportamiento actual.

Adicionalmente, se realizó un análisis de los residuos del modelo, específicamente se aplicó el Test de *Shapiro-Wilk* para evaluar la normalidad de los residuos y como resultado se obtuvo un *p-value* de 0.00, lo que indica que se rechaza la hipótesis nula de normalidad y los errores no se distribuyen de forma acampanada. Esto confirma que los registros de biodiversidad presentan eventos extremos que una tendencia lineal no puede suavizar, lo que refuerza la necesidad de la segunda etapa en el modelo híbrido. En el histograma con la distribución de residuos (normalidad) de la figura 19 se muestra una concentración en torno a cero, pero con una asimetría positiva (cola larga hacia la derecha) y presencia de valores extremos que superan el 1.0. Así mismo, en el *Q-Q Plot,* los puntos se desvían significativamente de la línea roja de 45° en ambos extremos, mostrando la presencia de colas pesadas y confirmando los resultados de la prueba *Shapiro-Wilk*.

De igual forma, se analizó la autocorrelación con la aplicación de una prueba *Durbin-Watson*, la cuál arrojó un significativamente menor de 2.0 (1,100), lo que indica una autocorrelación positiva fuerte y los errores de un año están correlacionados con los del año anterior. En la figura 19, en el gráfico de Residuos vs. Predicciones (Homocedasticidad), se observa que aunque los puntos están dispersos, hay una mayor concentración de errores a medida que aumentan los valores predichos de la inercia (eje X).

Por último, a partir de la aplicación de *Breusch-Pagan* para analizar la homocedasticidad, se rechaza la homocedasticidad con un *p-value* de 0,00. Esto significa que la varianza del error cambia a medida que aumenta el volumen de datos (posiblemente los errores son mayores en nodos grandes). En la Figura 19, se muestra la presencia de un pico prominente en el *Lag 1* que sobresale significativamente del área sombreada de confianza. Esto confirma el estadístico Durbin-Watson con una autocorrelación positiva fuerte.

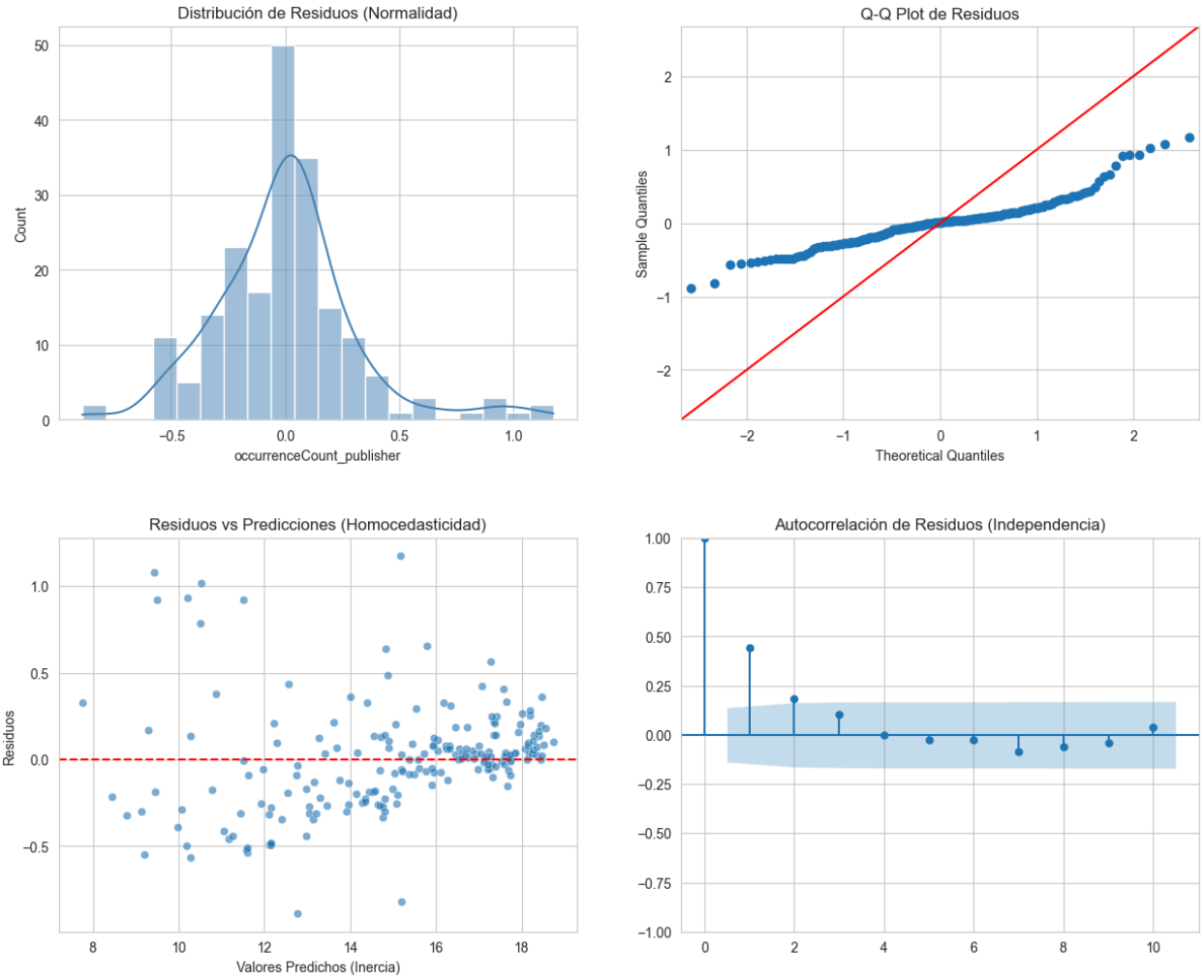


Figura 19. Diagnóstico estadístico de los residuos de la Etapa 1 (Regresión Ridge)

8.1.2. Etapa 2 del modelo híbrido

En la segunda etapa del marco híbrido, se implementó un modelo de *Random Forest Regressor*

diseñado específicamente para capturar la dinámica de innovación latente en los residuos generados por la etapa de inercia. Metodológicamente, este modelo no intenta predecir el volumen total de registros, sino que se entrena utilizando la diferencia matemática entre el valor real observado y la predicción de la regresión *Ridge* $y_{residuos} = y_{target} - y_{inercia}$, aislando así el componente estructural del peso del pasado.

La construcción del modelo emplea exclusivamente el conjunto de variables socioeconómicas, de inversión y gobernanza (*cols_estructural*), permitiendo que el algoritmo identifique interacciones no lineales y patrones complejos que la inercia temporal no logra explicar por sí misma. Para garantizar la precisión y robustez del sistema, el modelo utiliza los hiperparámetros previamente optimizados mediante validación cruzada y se apoya en los transformadores globales de escalado e imputación, asegurando que el análisis de la varianza residual sea técnicamente consistente con la tendencia base del sistema.

Cabe resaltar que aunque en el capítulo 6 se observa que LSTM, mostró un desempeño relativamente mejor que *Random Forest*, de acuerdo a métricas como el MSE y R^2, *Random Forest* fue considerado más idóneo para el planteamiento del *modelo híbrido*, pues El LSTM es una arquitectura diseñada para aprender secuencias y dado que la dependencia temporal ya fue extraída por el modelo autorregresivo (Etapa 1), no es necesario hacer uso de este enfoque nuevamente.

8.2. DESEMPEÑO DEL MODELO

Al evaluar el comportamiento del modelo de forma separada con el fin de evaluar el aporte de cada modelo al objetivo, se obtienen los siguientes resultados:

- Etapa 1 (*Ridge*): R^2=0.8993. El 90% del comportamiento de los datos se explica simplemente por la tendencia histórica y con un MAE (Error en registros reales) de 1,555,357 registros biológicos.
- Etapa 2 (*Random Forest*): R^2=0.5004. El Random Forest logra explicar la mitad (50%) de todo lo que la tendencia en el crecimiento de los datos, de la etapa 1, no puede explicar. Lo que sugiere que la inversión en ciencia es un motor real que impulsa el crecimiento de los datos.

De igual forma, para la etapa 2 se generó un análisis de SHAP con el objetivo de identificar la importancia de las variables en la explicación de los residuos (Fig. 20), donde se destacan

variables asociadas a la educación como *inscripción_terciaria*, *inscripción primaria*, y nuevamente *efectividad_gobierno* (que también había sido destacada en el modelo de Efectos Fijos).

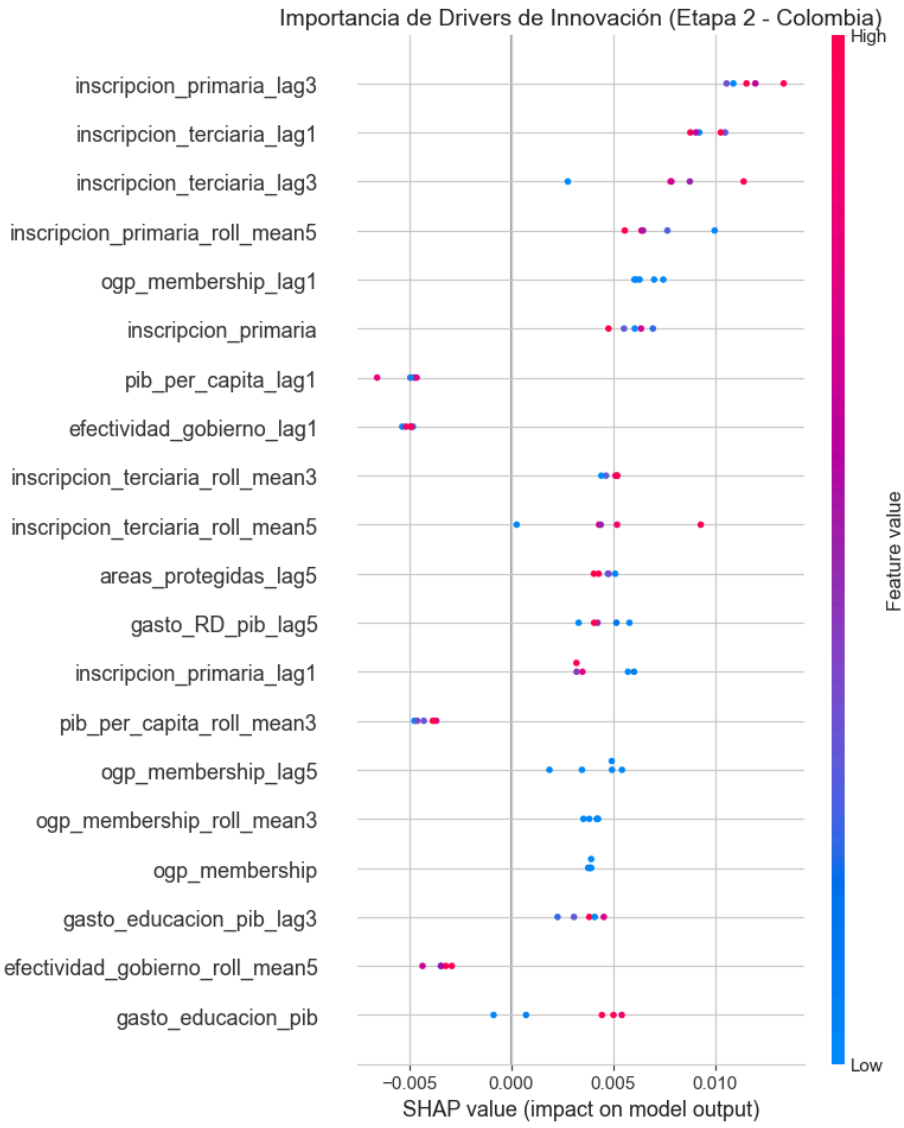


Figura 20. Gráfico resultante a partir de la validación con SHAP del modelo Random Forest ejecutado sobre los residuos de la etapa 1 de modelo híbrido. Los valores altos (puntos rojos) tienen un impacto positivo significativo en la producción de datos

La Figura 20 muestra que la variable *inscripcion_primaria_lag3* es el predictor que más aporta,

seguido por la educación terciaria. Esto sugiere que el fortalecimiento de la educación genera un retorno en la capacidad de publicación científica. De igual forma, la *inscripcion_terciaria_lag1* como sus promedios móviles (*roll_mean3, roll_mean5*) muestran que la inversión en educación superior es un *driver* de mediano plazo. Los puntos rojos a la derecha confirman que una base educativa sólida se correlaciona con un aumento sostenido en los registros de biodiversidad años después.

Por otro lado, la variable *efectividad_gobierno_lag1* muestra que los valores bajos (puntos azules) tienen un impacto negativo fuerte (desplazados hacia la izquierda en el eje SHAP). Esto sugeriría que la debilidad institucional resta potencial al crecimiento de la ciencia de datos en el país. Sin embargo, la pertenencia a la Alianza para el Gobierno Abierto (OGP) (*ogp_membership_lag1*) impacta positivamente, reforzando la idea de que las políticas de datos abiertos facilitan el flujo de información científica.

Finalmente, se evaluaron las métricas de la ejecución del *modelo híbrido, lo que mostró una mejora en del desempeño combinado con* un **MAE** de **651,015** registros, un **R^2** de **0,9739** y un **MAPE** de **5.04%**. Estos resultados muestran un alto comportamiento predictivo del modelo lo que lo hace adecuado para las siguientes etapas de predicción y simulación para Colombia aplicando diferentes escenarios de crecimiento en las variables explicativas.

Para validar el nivel de ajuste del modelo, se realizó adicionalmente una evaluación *Ex-post* con valores reales de los años 2023 y 2025 que no fueron tenidos en cuenta en el entrenamiento y pruebas del modelo (Fig. 21). Como resultado, se observa que la predicción es muy cercana al valor real. Acá se resalta que la precisión depende en gran medida del nivel de ajuste del α (Alpha) de penalización del modelo *Ridge* de la Etapa 1.

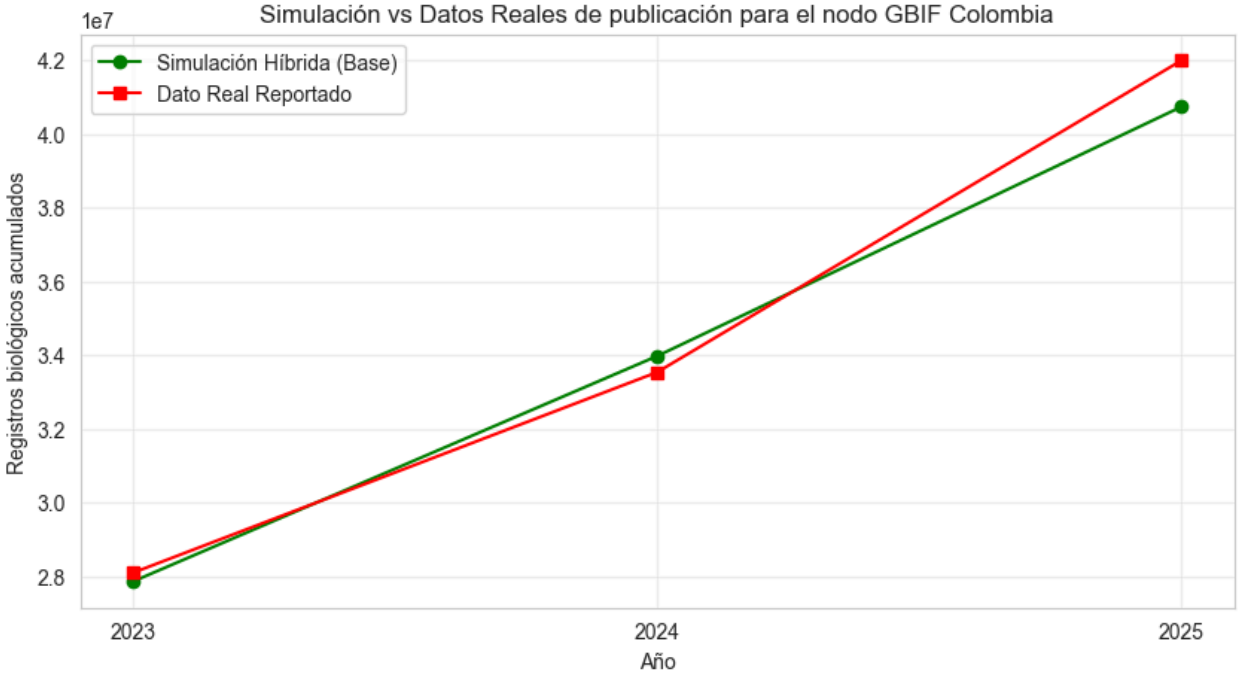


Figura 21. Validación de la simulación vs los datos reales de Colombia para los años 2023 a 2025 con un Alpha de 7.5 para el modelo Ridge de la etapa 1 del modelo híbrido.

8.3. PREDICCIÓN PARA COLOMBIA

El paso siguiente fue aplicar la definición del *modelo híbrido* para el caso del Nodo SiB Colombia. Sin embargo, para poder hacer la predicción fue necesario primero proyectar el comportamiento de las variables hacia el año 2030. En ese sentido, se implementó un algoritmo de simulación paso a paso que realiza una extrapolación lineal de las variables exógenas (socioeconómicas) mediante las funciones *polyfit* [57] y *poly1d* [58], basada en sus últimos cinco años de trayectoria.

Asimismo, se aplicaron restricciones a través de la función *clip* [59] para las variables *'gasto_RD_pib', 'gasto_educacion_pib', 'inscripcion_primaria', 'inscripcion_secundaria', 'inscripcion_terciaria','gasto_educacion_gobierno', 'areas_protegidas', 'uso_internet', de forma que las extrapolaciones no generarán valores mayores a la unidad de porcentaje*. Este ajuste

resultó ser de vital importancia pues afectó directamente las métricas del modelo luego de implementarlas.

Además, se integran en el *pipeline* el *IterativeImputer* para el manejo de datos faltantes y *StandardScaler* para el escalado de los datos a una escala común con media 0 y varianza 1. En esta simulación recursiva cada año proyectado alimenta los insumos del periodo siguiente, garantizando que el sistema evolucione de forma coherente con su propia historia.

8.3.1. Evaluación de escenarios estratégicos

Se definieron tres escenarios prospectivos para evaluar la sensibilidad del sistema ante diferentes palancas de política pública con diferentes niveles de intensidad en el cambio de los indicadores, con incrementos del 2, 5, 10 y 20% (Fig. 22):

- Base (Tendencial): Proyecta el crecimiento siguiendo la inercia actual sin intervenciones adicionales.
- Variables de simulación: Simula un incremento en las métricas de '*uso_internet', 'efectividad_gobierno'* y '*gasto_educacion_pib*', 'inscripcion_primaria', '*inscripcion_secundaria*', '*inscripcion_secundaria*' y '*art_científicos'*, de acuerdo a los resultados de los modelos de Efectos Fijos y el SHAP del Random Forest aplicado al residuo.

El código para la implementación del modelo para Colombia se encuentra disponible en GitHub:

- https://github.com/rortizgeo/Maestria_CD_Proyecto-Aplicado/blob/main/3_PA_Workflow_Paneldata.ipynb

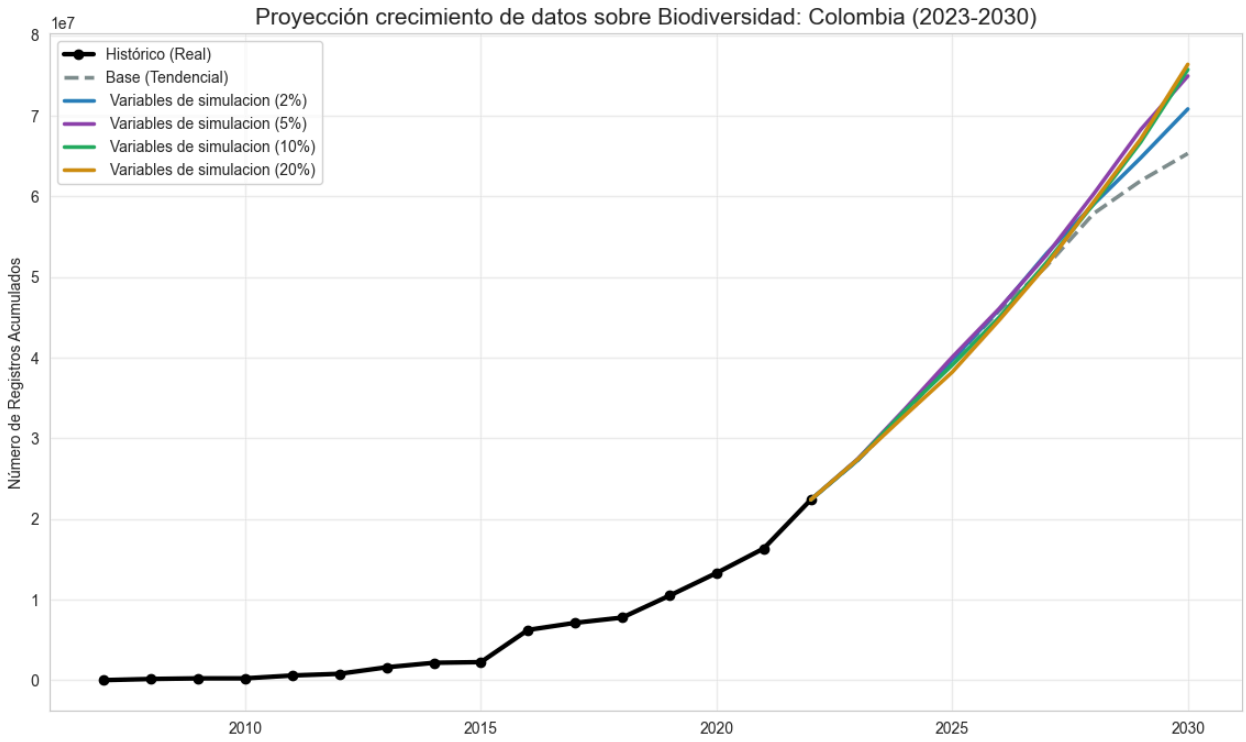


Figura 22. Simulación de escenario para el Nodo GBIF Colombia para hasta el año 2030 con cambios incrementales en las políticas del 2% hasta el 20%.

La ejecución de la simulación a partir del modelo híbrido muestra en general una dinámica de rendimientos crecientes sobre un escenario base (Tendencial) que proyecta un crecimiento inercial que alcanzaría aproximadamente los 65 millones de registros para 2030.

El modelo parece indicar que incrementos del 2% y 5% logran despegar la curva de la base, demostrando que incluso mejoras marginales en conectividad y educación evitan el estancamiento y aportan a la publicación de datos y crecimiento de la red. Los escenarios de 10% y 20% generan un salto más marcado, superando los 75 millones de registros para 2030. Sin embargo el comportamiento es bastante similar en escenarios del 5%, 10% y 20%.

Por último, se observa que a partir de 2027 las curvas se separan drásticamente. Esto se debe a la naturaleza recursiva del modelo (Bloque 9 de código en GitHub), pues las mejoras en las variables base hoy se acumulan y potencian a través de sus rezagos (l*ags*) y promedios móviles (*roll windows*) en los años finales.

9 CONCLUSIONES Y TRABAJOS FUTUROS

9.1. CONCLUSIONES

El presente trabajo permitió abordar, desde una perspectiva de ciencia de datos y a través del uso de metodologías como el CRISP-DP, el desarrollo de un modelo híbrido para la predicción del crecimiento de la publicación de datos sobre biodiversidad en el nodo GBIF Colombia (SiB Colombia) mediante el análisis de variables socioeconómicas y decisiones gubernamentales.

A partir de las fuentes de datos seleccionadas que cubrían los datos generados a través de la red global GBIF, los datos socioeconómicos disponibles en el Banco Mundial y otras fuentes de información como la participación de los países en las Alianzas Gubernamentales de datos abiertos, así como de la implementación de técnicas de modelado como Efectos Fijos (FE), regularización LASSO y análisis de interpretabilidad SHAP sobre modelos de Machine Learning (Random Forest), se determinó que las variables socioeconómicas logran explicar aproximadamente el 50% de la variabilidad residual en el crecimiento de los datos al interior de los países. Este hallazgo identifica a la educación terciaria (superior) y la efectividad gubernamental como los catalizadores fundamentales que permiten a una nación superar su tendencia inercial histórica.

Asimismo, los análisis mediante regresión LASSO destacaron la infraestructura digital, específicamente el uso de internet, como un predictor de alta magnitud cuyo impacto triplica al de otras variables estructurales en la movilización de datos científicos. Finalmente, la implementación del modelo híbrido permitió evidenciar que variables asociadas al acceso a la educación e investigación operan con rezagos temporales de entre 3 y 5 años. Lo anterior indica que las políticas de ciencia y educación básica ejecutadas en el presente consolidan la capacidad de producción y publicación de datos en el mediano plazo, garantizando una evolución técnica coherente con la trayectoria institucional del país.

Por otro lado, en cuanto a la implementación y evaluación de modelos estadísticos y de Machine Learning para la predicción del crecimiento en publicación de datos en los nodos de GBIF, el análisis comparativo de desempeño evidencia que el modelo híbrido (*Ridge + Random Forest*) constituye la arquitectura más robusta y precisa para la predicción del crecimiento en la publicación de datos de biodiversidad, específicamente en el nodo SiB Colombia, en comparación de enfoques como *Random Forest*, XGBoost que incluye los rezagos y ventanas móviles de la variable objetivo, y que eclipsaron el aporte de las variables socioeconómicas

(sobre todo en modelos tabulares). El resultado del modelo híbrido se debe principalmente a que este logra una descomposición analítica del fenómeno en dos etapas fundamentales que implicaron separar la tendencia inercial del crecimiento de los datos a través de un modelo como *Ridge* que penalizaba el crecimiento exponencial y lo ajustaba a la realidad, a la vez que se implementaba un modelo basado en *Machine Learning* que toma los residuos no explicados en la etapa 1 del modelo híbrido y los convierte en un aporte a la predicción del modelo, llegando a explicar cerca del 50% de la varianza que la inercia histórica no puede justificar el modelo *Ridge*. Con resultados de R2 de 0.97 se puede decir que el modelo híbrido es robusto para la predicción.

El modelo híbrido no solo predice el volumen de datos con un margen de error reducido, sino que actúa como una herramienta de diagnóstico capaz de aislar las variables socioeconómicas que más impactan el crecimiento de dicho volumen, de una forma fácilmente interpretable. Esto permite proyectar escenarios al 2030 que pueden alcanzar hasta los 75 millones de registros publicados bajo escenarios optimistas de crecimiento, contra los cerca de 65 millones de registros que muestra el escenario tendencial. Adicionalmente, las simulaciones mostraron un punto de inflexión marcado a partir de 2027, momento en el que los efectos rezagados de las mejoras en educación y gobernanza comienzan a potenciar drásticamente la curva de publicación.

El modelo desarrollado constituye una herramienta técnica de planificación que permite al país fundamentar sus metas ante el Marco Mundial Kunming-Montreal, demostrando que el cumplimiento de la Meta 21 del marco, depende de una sinergia entre inversión en educación ciencia y fortalecimiento institucional para, cómo mínimo, mantener el ritmo de publicación actual.

9.2. TRABAJOS FUTUROS

A partir de los resultados obtenidos, se recomienda avanzar hacia una exploración más exhaustiva y diversificada de variables explicativas que permitan fortalecer la capacidad predictiva del modelo y capturar con mayor precisión los determinantes del crecimiento en la publicación de datos de biodiversidad. En particular, sería valioso incorporar indicadores directamente vinculados con la infraestructura de información biológica y con los procesos de gestión institucional de datos, tales como: el número de colecciones biológicas registradas y activas por país, el volumen estimado de especímenes aún no digitalizados o sin publicar, organizados por año (Para su replicabilidad en series de tiempo), la existencia de proyectos de

digitalización o movilización de datos en curso, y los niveles de inversión provenientes de fuentes multilaterales (p. ej., Banco Mundial, BID o GEF) destinados específicamente a la gestión de datos e información ambiental y sobre biodiversidad. Estos indicadores permitirían complementar las variables macroeconómicas y sociales utilizadas en este estudio, ofreciendo una caracterización más directa del ecosistema de publicación de datos y de las capacidades reales de los países para sostener procesos de movilización a largo plazo.

El modelo y las decisiones alrededor de los resultados también podrían beneficiarse de hacer pruebas con proyecciones de publicación de datos provenientes de diferentes fuentes como: ciencia ciudadana, colecciones biológicas, metabarcoding, o incluso generar diferenciaciones institucionales (centros/institutos de Investigación, academia, sector empresarial, autoridades ambientales, etc. Esto permitiría diseñar estrategias diferenciadas según el tipo de publicador.

Otro aspecto a trabajar es la integración de un ciclo de retroalimentación de expertos para que los escenarios de simulación sean más realistas y en lugar de usar factores fijos (2%, 5% o 20%), se definen umbrales de cambio factibles por variable a partir de las recomendaciones de paneles de expertos, al menos para el contexto colombiano.

A partir de los resultados en el presente trabajo se podría continuar con la optimización del modelo híbrido, específicamente con el modelo *Ridge* de la etapa 1 y ajustar el pipeline para una posible automatización, que permita escalar el modelo hacia una herramienta interactiva de proyección bajo escenarios de crecimiento, donde los parámetros de entrada, como la inversión en ciencia y tecnología, porcentaje de escolaridad, efectividad del gobierno, el fortalecimiento institucional, entre otras, puedan ajustarse en función de escenarios realistas construidos con el conocimiento de expertos y responsables de política pública. Este tipo de aplicación tendría un alto valor estratégico, al ofrecer una base cuantitativa para la toma de decisiones y la planificación prospectiva de iniciativas como el SiB Colombia y otros nodos nacionales de GBIF.

10 REFERENCIAS BIBLIOGRÁFICAS

[1] IPBES, “Summary for policymakers of the global assessment report on biodiversity and ecosystem services”, Zenodo, nov. 2019. doi: 10.5281/zenodo.3553579.

[2] Conference of the Parties to the Convention on Biological Diversity, “Kunming-Montreal Global Biodiversity Framework”. [En línea]. Disponible en: https://www.cbd.int/doc/c/2c37/244c/133052cdb1ff4d5556ffac94/cop-15-l-25-es.pdf

[3] “GBIF”, Global Biodiversity Information Facility. Free and Open Access to Biodiversity Data. [En línea]. Disponible en: https://www.gbif.org/

[4] “SiB Colombia”, Sistema de información sobre biodiversidad de Colombia - SiB Colombia. Consultado: el 18 de noviembre de 2024. [En línea]. Disponible en: https://biodiversidad.co/

[5] E. Ortegón, J. F. Pacheco, y A. Prieto, *Metodología del marco lógico para la planificación, el seguimiento y la evaluación de proyectos y programas*. en Manuales, no. 42. Santiago de Chile: Naciones Unidas. CEPAL, 2005.

[6] N. L. Cole, E. Kormann, T. Klebel, S. Apartis, y T. Ross-Hellauer, “The societal impact of Open Science: a scoping review”, *R. Soc. Open Sci.*, vol. 11, núm. 6, p. 240286, jun. 2024, doi: 10.1098/rsos.240286.

[7] S. Beck *et al.*, “The Open Innovation in Science research field: a collaborative conceptualisation approach”, *Ind. Innov.*, vol. 29, núm. 2, pp. 136–185, feb. 2022, doi: 10.1080/13662716.2020.1792274.

[8] GBIF Secretariat, “GBIF Science Review No. 11”. Consultado: el 18 de noviembre de 2024. [En línea]. Disponible en: https://www.gbif.org/document/5N9YVBkTP3y7kqhttps://doi.org/10.35035/d9pk-1162

[9] D. Escobar, “Publicación de datos e información sobre biodiversidad, un modelo del SiB Colombia”.

[10] D. E. Bowler *et al.*, “Temporal trends in the spatial bias of species occurrence records”, *Ecography*, vol. 2022, núm. 8, p. e06219, 2022, doi: 10.1111/ecog.06219.

[11] J. M. Heberling, J. T. Miller, D. Noesgaard, S. B. Weingart, y D. Schigel, “Data integration enables global biodiversity synthesis”, *Proc. Natl. Acad. Sci.*, vol. 118, núm. 6, p. e2018093118, feb. 2021, doi: 10.1073/pnas.2018093118.

[12] J. Troudet, P. Grandcolas, A. Blin, R. Vignes-Lebbe, y F. Legendre, “Taxonomic bias in biodiversity data and societal preferences”, *Sci. Rep.*, vol. 7, núm. 1, p. 9132, ago. 2017, doi: 10.1038/s41598-017-09084-6.

[13] A. Fairbrass, G. Mace, P. Ekins, y B. Milligan, “The natural capital indicator framework (NCIF) for improved national natural capital reporting”, *Ecosyst. Serv.*, vol. 46, p. 101198, dic. 2020, doi: 10.1016/j.ecoser.2020.101198.

[14] S. L. Stevenson, K. Watermeyer, G. Caggiano, E. A. Fulton, S. Ferrier, y E. Nicholson, “Matching biodiversity indicators to policy needs”, *Conserv. Biol.*, vol. 35, núm. 2, pp. 522–532, 2021, doi: 10.1111/cobi.13575.

[15] P. Leadley *et al.*, “Achieving global biodiversity goals by 2050 requires urgent and integrated actions”, *One Earth*, vol. 5, núm. 6, pp. 597–603, jun. 2022, doi: 10.1016/j.oneear.2022.05.009.

[16] IPBES, “Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services”, Zenodo, may 2019. doi: 10.5281/ZENODO.3831673.

[17] “Estimation of resources needed for implementing the post-2020 global biodiversity framework. Preliminary second report of the Panel of Experts on Resource Mobilization: Supplementary information”.

[18] S. S. Farley, A. Dawson, S. J. Goring, y J. W. Williams, “Situating Ecology as a Big-Data Science: Current Advances, Challenges, and Solutions”, *BioScience*, vol. 68, núm. 8, pp. 563–576, ago. 2018, doi: 10.1093/biosci/biy068.

[19] D. Steinke, B. Gemeinholzer, E. Martínez-Meyer, D. Noesgaard, A. Young, y D. Schigel, “Globally aggregated biodiversity data impact predictive and descriptive research”, *Proc. Natl. Acad. Sci.*, vol. 122, núm. 50, p. e2519119122, 2025, doi: 10.1073/pnas.2519119122.

[20] C. Mandeville, W. Koch, E. Nilsen, y A. Finstad, “Open Data Practices among Users of Primary Biodiversity Data”, *BioScience*, vol. 71, ago. 2021, doi: 10.1093/biosci/biab072.

[21] R. Mac Nally, “Multiple regression and inference in ecology and conservation biology: further comments on identifying important predictor variables”, *Biodivers. Conserv.*, vol. 11, núm. 8, pp. 1397–1401, ago. 2002, doi: 10.1023/A:1016250716679.

[22] R. Mac Nally, “Regression and model-building in conservation biology, biogeography and ecology: The distinction between – and reconciliation of – ‘predictive’ and ‘explanatory’ models”, *Biodivers. Conserv.*, vol. 9, núm. 5, pp. 655–671, may 2000, doi: 10.1023/A:1008985925162.

[23] L. Breiman, “Random Forests”, *Mach. Learn.*, vol. 45, núm. 1, pp. 5–32, oct. 2001, doi: 10.1023/A:1010933404324.

[24] T. Chen y C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, en KDD ’16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[25] S. N. Wood, *Generalized Additive Models: An Introduction with R, Second Edition*, 2a ed. New York: Chapman and Hall/CRC, 2017. doi: 10.1201/9781315370279.

[26] M. Camargo, M. Dumas, y O. González-Rojas, “Learning Accurate LSTM Models of Business Processes”, en *Business Process Management*, T. Hildebrandt, B. F. van Dongen, M. Röglinger, y J. Mendling, Eds., Cham: Springer International Publishing, 2019, pp. 286–302. doi: 10.1007/978-3-030-26619-6_19.

[27] M. Zuccotto, A. Castellini, D. L. Torre, L. Mola, y A. Farinelli, “Reinforcement learning applications in environmental sustainability: a review”, *Artif. Intell. Rev.*, vol. 57, núm. 4, p. 88, mar. 2024, doi: 10.1007/s10462-024-10706-5.

[28] S. Fujimoto, H. Hoof, y D. Meger, “Addressing Function Approximation Error in Actor-Critic Methods”, en *Proceedings of the 35th International Conference on Machine Learning*, PMLR, jul. 2018, pp. 1587–1596. Consultado: el 18 de noviembre de 2024. [En línea]. Disponible en: https://proceedings.mlr.press/v80/fujimoto18a.html

[29] M. Lapeyrolerie, M. S. Chapman, K. E. A. Norman, y C. Boettiger, “Deep reinforcement learning for conservation decisions”, *Methods Ecol. Evol.*, vol. 13, núm. 11, pp. 2649–2662, 2022, doi: 10.1111/2041-210X.13954.

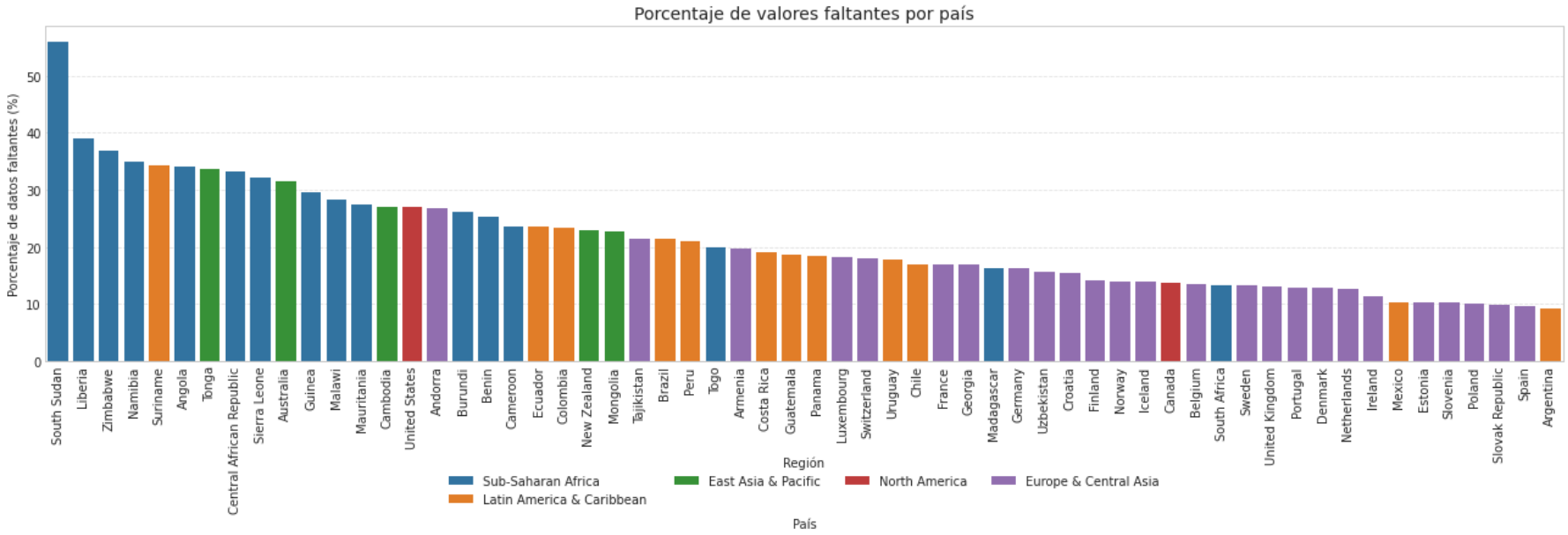
[30] C. Schröer, F. Kruse, y J. M. Gómez, “A Systematic Literature Review on Applying CRISP-DM Process Model”, *Procedia Comput. Sci.*, vol. 181, pp. 526–534, ene. 2021, doi: 10.1016/j.procs.2021.01.199.

[31] P. Chapman, "CRISP-DM 1.0: Step-by-step data mining guide", 2000. Consultado: el 17 de enero de 2026. [En línea]. Disponible en: https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72
[32] *wbgapi: wbgapi provides a comprehensive interface to the World Bank's data and metadata APIs*. Python. Consultado: el 2 de noviembre de 2025. [OS Independent]. Disponible en: https://github.com/tgherzog/wbgapi
[33] "GBIF API". Consultado: el 2 de noviembre de 2025. [En línea]. Disponible en: https://api.gbif.org/
[34] "GBIF". Consultado: el 2 de noviembre de 2025. [En línea]. Disponible en: https://www.gbif.org/
[35] D. Oldoni *et al.*, "Occurrence cubes: a new paradigm for aggregating species occurrence data", el 25 de marzo de 2020, *bioRxiv*. doi: 10.1101/2020.03.23.983601.
[36] J. Otegui, A. H. Ariño, M. A. Encinas, y F. Pando, "Assessing the Primary Data Hosted by the Spanish Node of the Global Biodiversity Information Facility (GBIF)", *PLOS ONE*, vol. 8, núm. 1, p. e55144, ene. 2013, doi: 10.1371/journal.pone.0055144.
[37] "Index of /". Consultado: el 2 de noviembre de 2025. [En línea]. Disponible en: https://analytics-files.gbif.org/
[38] *requests: Python HTTP for Humans.* Python. Consultado: el 2 de noviembre de 2025. [OS Independent]. Disponible en: https://requests.readthedocs.io
[39] *pandas: Powerful data structures for data analysis, time series, and statistics*. Cython, Python.
[40] J. C. Correa Morales y J. C. Salazar Uribe, *Introducción a los modelos mixtos*. Universidad Nacional de Colombia, 2016. Consultado: el 13 de diciembre de 2025. [En línea]. Disponible en: https://www.researchgate.net/publication/314536942_Introduccion_a_los_modelos_mixtos_Introduction_to_mixed_models
[41] B. H. Baltagi, *Econometric Analysis of Panel Data*. en Springer Texts in Business and Economics. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3-030-53953-5.
[42] Y. Qin y M. Al Amin, "Panel Data Using R: Fixed-effects and Random-effects". Consultado: el 13 de diciembre de 2025. [En línea]. Disponible en: https://libguides.princeton.edu/c.php?g=1258919&p=9227112
[43] "What are the most effective methods for testing heteroscedasticity in panel data?", What are the most effective methods for testing heteroscedasticity in panel data? [En línea]. Disponible en: https://www.linkedin.com/advice/0/what-most-effective-methods-testing-heteroscedasticity?lang=es&originalSubdomain=es
[44] "StandardScaler", scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
[45] "IterativeImputer", scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.impute.IterativeImputer.html
[46] "7.4. Imputation of missing values", scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/impute.html
[47] "Precisión mixta | TensorFlow Core", TensorFlow. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://www.tensorflow.org/guide/mixed_precision?hl=es-419
[48] "TimeSeriesSplit", scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

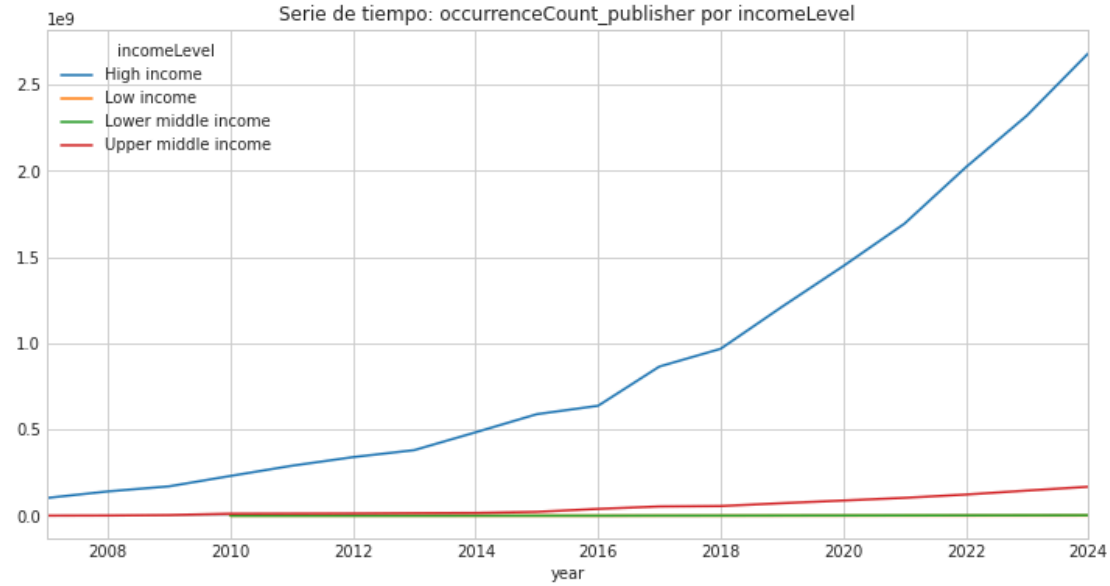
[49] “10. Fallas comunes y prácticas recomendadas — documentación de scikit-learn - 0.24.1”. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://qu4nt.github.io/sklearn-doc-es/common_pitfalls.html
[50] “Nested versus non-nested cross-validation”, scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/auto_examples/model_selection/plot_nested_cross_validation_iris.html
[51] “RandomizedSearchCV”, scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html
[52] “tf.keras.callbacks.EarlyStopping | TensorFlow v2.16.1”, TensorFlow. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://www.tensorflow.org/api_docs/python/tf/keras/callbacks/EarlyStopping
[53] “Welcome to the SHAP documentation — SHAP latest documentation”. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://shap.readthedocs.io/en/latest/
[54] “LinearRegression”, scikit-learn. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.linear_model.LinearRegression.html
[55] J. Murel y E. Kavlakoglu, “¿Qué es la regresión Ridge? | IBM”. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://www.ibm.com/es-es/think/topics/ridge-regression
[56] “Ridge”, scikit-learn. Consultado: el 16 de enero de 2026. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.linear_model.Ridge.html
[57] “numpy.polyfit — NumPy v2.4 Manual”. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://numpy.org/doc/stable/reference/generated/numpy.polyfit.html
[58] “numpy.poly1d — NumPy v2.4 Manual”. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://numpy.org/doc/stable/reference/generated/numpy.poly1d.html
[59] “numpy.clip — NumPy v2.4 Manual”. Consultado: el 11 de enero de 2026. [En línea]. Disponible en: https://numpy.org/doc/stable/reference/generated/numpy.clip.html

11 ANEXOS

Anexo 1. Datos faltantes por país y región.



Anexo 2. Evolución anual de la publicación de datos de biodiversidad por nivel de ingreso del país.



Anexo 3. Prueba de comportamiento del modelo predictivo sin la aproximación del modelo híbrido y usando lags y roll windows en el modelo ganador RF discutido en el capítulo 6.