

Reducción de Rezago de Avalúos Catastrales - Artículo 49 PND(2022-2026)

Modelos estadísticos para la predicción de valores de terreno

Elaborado por:

Camilo Andrés Avellaneda

Raúl Andrés Rodríguez

María Fernanda Zarate

Andrés Felipe Gómez

Carol Chicuazuque

INSTITUTO GEOGRÁFICO AGUSTÍN CODAZZI - IGAC
SUBDIRECCIÓN GENERAL

Diciembre de 2024

Abreviaturas y acrónimos

PND	- Plan Nacional de Desarrollo
IGAC	- Instituto Geográfico Agustín Codazzi
DIP	- Dirección de Investigación y Prospectiva
OIC	- Observatorio Inmobiliario Catastral
SNR	- Superintendencia de Notariado y Registro



Resumen

En el artículo 49 del plan nacional de desarrollo (PND) 2022 - 2026 se plantea que el Instituto Geográfico Agustín Codazzi (IGAC) adoptará metodologías y modelos de actualización masiva de valores catastrales rezagados, que permitan por una sola vez realizar un ajuste automático de los avalúos catastrales de todos los predios del país. Para responder a este requerimiento se plantea una propuesta a partir de información disponible en el Observatorio Inmobiliario Catastral (OIC) del IGAC, la cual se basa en transacciones, avalúos comerciales y ofertas recolectadas por diferentes fuentes de información. La propuesta es hacer uso de diferentes metodologías de modelos estadísticos que permitan predecir los valores de terreno comerciales para predios/locaciones rurales en municipios bajo gestión del IGAC. A partir de la información y del conjunto de modelos estadísticos definidos se generan resultados de valores para 26 departamentos del país. Por último, estos resultados se agregan para conformar valores por polígonos definidos por las Zonas Homogéneas Geoeconómicas (ZHG) vigentes en los municipios correspondientes, junto con reportes de información que detallan los resultados y aspectos más relevantes del proceso de modelación.

Tabla de contenido

Resumen	3
Tabla de contenido	4
1) Introducción	8
2) Justificación	9
3) Alcance	10
4) Objetivos	10
4.1 General	10
4.2 Específicos	10
5) Marco teórico	11
6) Desarrollo metodológico	15
6.1 Análisis de la información	16
6.2 Preparación y transformación de los datos	23
6.3 Desarrollo del modelo y/o metodología	27
6.3.1 Métricas de rendimiento	27
6.3.2 Determinación del Valor de la Zona Geoeconómica (ZHG)	28
7) Resultados y discusiones	29
7.1 Resultados del Proceso de Modelación	32
7.1.1 Aplicación de metodologías	33
7.1.2 Selección de variables	39
7.1.3 Generación del modelo final	39
7.1.4 Evaluación del modelo	40
7.2 Implementación y/o difusión de los resultados	42
7.2.1 Cálculo del Valor de la Zona Geoeconómica (ZHG)	42
7.3 Evaluación y/o presentación de los resultados	44
8) Conclusiones	46
9) Bibliografía	47
10) Anexos	50
10.1 Anexos A	50
10.2 Anexos B	50

Lista de figuras

Figura 1: Semivariograma a partir de datos simulados.

Figura 2: Cociente mediano de valores catastrales versus comerciales por fuente de información.

Figura 3: Cociente mediano de valores catastrales versus comerciales por fuente de información posterior al ajuste de SNR.

Figura 4: Diagrama del modelo de entidad relación para la elaboración del a base de modelamiento a partir de capas geográficas.

Figura 5: Diagrama del modelo de entidad relación para la elaboración del a base de modelamiento a partir de capas geográficas y tablas alfanuméricas.

Figura 6: Diagrama de barras correspondiente a la cantidad de predios en el marco por departamento.

Figura 7: Diagrama de barras correspondiente a la cantidad de predios en la muestra por departamento.

Figura 8: Distribución del valor de terreno por HA para los departamentos con logaritmo y sin logaritmo.

Figura 9: Resultados del pseudo-R cuadrado por rangos de área en el Departamento de Quindío.

Figura 10: Resultados del pseudo-R cuadrado a partir de diferentes metodologías en el Departamento de Quindío.

Figura 11: Resultados del RMEDSE cuadrado a partir de diferentes metodologías en el Departamento de Caquetá.

Figura 12: Pseudo R2 para los modelos finales por Departamento.

Figura 13: RMEDSE para los modelos finales por Departamento.

Figura 14: Zona Geoeconómica 24772-11-1, Suesca, Cundinamarca.

Figura 15: Predicciones de valor de terreno por hectárea generadas para el departamento del Atlántico.

Figura 16: Predicciones de valor de terreno por hectárea generadas para el departamento de Bolívar

Figura 17: Predicciones de valor de terreno por hectárea generadas para el departamento de Boyacá

Figura 18: Predicciones de valor de terreno por hectárea generadas para el departamento de Caldas

Figura 19: Predicciones de valor de terreno por hectárea generadas para el departamento de Caquetá

Figura 20: Predicciones de valor de terreno por hectárea generadas para el departamento de Cauca

Figura 21: Predicciones de valor de terreno por hectárea generadas para el departamento de Cesar

Figura 22: Predicciones de valor de terreno por hectárea generadas para el departamento de Córdoba

Figura 23: Predicciones de valor de terreno por hectárea generadas para el departamento de Cundinamarca

Figura 24: Predicciones de valor de terreno por hectárea generadas para el departamento del Chocó

Figura 25: Predicciones de valor de terreno por hectárea generadas para el departamento del Huila

Figura 26: Predicciones de valor de terreno por hectárea generadas para el departamento de La Guajira

Figura 27: Predicciones de valor de terreno por hectárea generadas para el departamento del Magdalena

Figura 28: Predicciones de valor de terreno por hectárea generadas para el departamento del Meta

Figura 29: Predicciones de valor de terreno por hectárea generadas para el departamento de Nariño

Figura 30: Predicciones de valor de terreno por hectárea generadas para el departamento de Norte de Santander

Figura 31: Predicciones de valor de terreno por hectárea generadas para el departamento del Quindío

Figura 32: Predicciones de valor de terreno por hectárea generadas para el departamento de Risaralda

Figura 33: Predicciones de valor de terreno por hectárea generadas para el departamento de Santander

Figura 34: Predicciones de valor de terreno por hectárea generadas para el departamento de Sucre

Figura 35: Predicciones de valor de terreno por hectárea generadas para el departamento del Tolima

Figura 36: Predicciones de valor de terreno por hectárea generadas para el departamento de Valle del Cauca

Figura 37: Predicciones de valor de terreno por hectárea generadas para el departamento de Arauca

Figura 38: Predicciones de valor de terreno por hectárea generadas para el departamento de Casanare

Figura 39: Predicciones de valor de terreno por hectárea generadas para el departamento del Putumayo

Figura 40: Predicciones de valor de terreno por hectárea generadas para el departamento del Vichada

Lista de tablas

Tabla 1: Cantidad de registros con asignación de código predial en la base de información consolidada.

Tabla 2: Jerarquía de las fuentes de información utilizada para eliminar registros duplicados a nivel de predio.

Tabla 3: Cantidad y participación de predios en la base de modelación resultante por Fuente de Información.

Tabla 4: Cantidad y participación de predios en la base de modelación resultante por Departamento.

Tabla 5: Cantidad de predios en el marco y en la muestra utilizada.

Tabla 6: Mediana de los resultados del pseudo-R2 por Departamento y metodología aplicada en el proceso realizado.

Tabla 7: Mediana de los resultados del pseudo-R2 por Departamento y metodología aplicada en el proceso realizado.

Tabla 8: Resultados del RMEDSE por Departamento y metodología aplicada en el proceso realizado.

Tabla 9: Resultados del RMEDSE por Departamento y metodología aplicada en el proceso realizado.

Tabla 10: Predios en la ZHG: 25772-11-1, con valores y áreas de terreno por hectárea.

Tabla 11: Valor sugerido para la ZHG calculado por la media, mediana, media y mediana ponderadas.

1) Introducción

En la actualidad existe una cantidad considerable de municipios en el país que se encuentran desactualizados desde una perspectiva catastral. Esto implica una desactualización en el enfoque económico, físico y jurídico. Esto puede conllevar a diferentes problemáticas, entre las cuales se resalta el hecho de que la tributación se realiza con valores que se pueden encontrar por debajo de los valores catastrales reales.

Por este motivo, en el artículo 49 del plan nacional de desarrollo (PND) 2022 - 2026 se plantea que el Instituto Geográfico Agustín Codazzi (IGAC), adoptará metodologías y modelos de actualización masiva de valores catastrales rezagados, que permitan por una sola vez realizar un ajuste automático de los avalúos catastrales de todos los predios del país. A pesar de que el artículo 49 especifica una reducción del rezago en predios de todo el país, incluyendo municipios bajo gestoría de catastros descentralizados, en este documento y en el trabajo realizado, únicamente se abordan municipios bajo gestoría del IGAC. Adicionalmente, debido a directrices definidas, se abordaron predios en suelo rural y solo se tiene en cuenta la actualización de valores de terreno, más no de construcción. Por otro lado, los predios en suelo urbano serán abordados en un trabajo posterior.

De esta manera, se realiza todo un proceso de verificación de insumos disponibles tanto interna como externamente, que permitan la generación de modelos estadísticos para la predicción de valores de terreno. La validación de insumos iniciales incluye la determinación de cuáles variables presentan utilidad dentro del ejercicio, cuáles requieren procesamientos adicionales y qué otros procesos se deben realizar para obtener más información relacionada con los registros dispuestos en cada una de las fuentes disponibles, tales como lo que se realiza en los procesos de georreferenciación y cruces de tipo espacial para determinar códigos identificadores. Esta información se basa en transacciones provenientes de la Superintendencia de Notariado y Registro (SNR), avalúos comerciales y ofertas recolectadas. A partir de la información procesada y consolidada con valores económicos se genera una base, la cual posteriormente se complementa con variables exógenas que apalanquen la generación de modelos estadísticos. En este caso, la variable dependiente o respuesta, $Z(\mathbf{s})$, se define como el valor de terreno por hectárea en una locación geográfica o para un predio específico. En este sentido, se asume que $\{Z(\mathbf{s})\}$ es un campo aleatorio con $\mathbf{s} \in D$ y D es el conjunto índice, el cual para este propósito se supone continuo y fijo. Este conjunto índice se delimita por la población objetivo correspondiente.

Para obtener las predicciones de valor de terreno requeridas, se generaron diferentes propuestas de modelos supervisados y a partir de métricas definidas se realizaron comparaciones para determinar cuál de los modelos propuestos es el más adecuado. De esta forma, se realizaron diferentes análisis de sensibilidad y robustez para determinar la viabilidad de los modelos

propuestos. Adicionalmente, se generaron análisis de importancia de variables para determinar cuáles son las más relevantes en la predicción de la variable respuesta. Por último, al modelo definido se adiciona un componente estocástico que contempla la autocorrelación espacial del campo aleatorio. Una vez se tienen valores de manera desagregada para cada uno de los predios, se agregan mediante medidas de tendencia central para tener un único valor por polígono definido por las Zonas Homogéneas Geoeconómicas (ZHG) vigentes en los municipios correspondientes a partir de medidas de tendencia central.

Dado el objetivo de realizar la predicción de valores para diferentes locaciones en el país y que esta labor se encuentra dentro de las actividades que se adelantan periódicamente, el ejercicio de modelación que aquí se adelanta busca ser el punto de partida de una alternativa para la generación de valores económicos en futuras vigencias. De esta manera, el planteamiento que aquí se describe debe ser escalable y por este motivo se realiza en software libre. El proyecto se desarrolla en su totalidad en R (R Core Team 2023) y Python (Python Software Foundation 2023) y se hace uso de diferentes paquetes que permiten la generación de modelos estadísticos, análisis de datos y visualización de resultados.

Este documento se compone como se describe a continuación. En la [Sección 2](#) se plantea la justificación del trabajo. En la [Sección 3](#) se define el alcance de los modelos planteados. En la [Sección 4](#) se plantean los objetivos generales y específicos. En la [Sección 5](#) se plantea el marco teórico para el desarrollo de los modelos para la predicción de valores en el contexto descrito. En la [Sección 6](#) se plantea la metodología utilizada desde la construcción de la base de modelamiento a partir de la información disponible, la elaboración de diferentes modelos para su posterior comparación y selección de un modelo definitivo. Adicionalmente, se realiza un proceso de selección de variables y posterior ajuste del componente estocástico de autocorrelación espacial. En la [Sección 7](#) se muestran los resultados obtenidos a partir de la metodología planteada. En la [Sección 8](#) se presentan las conclusiones obtenidas a partir del trabajo realizado. Por último, en la [Sección 9](#) y [Sección 10](#) se presentan las referencias bibliográficas utilizadas en el documento y los anexos que se consideraron pertinentes, respectivamente.

2) Justificación

En el artículo 49 del plan nacional de desarrollo (PND) 2022 - 2026 se plantea que el Instituto Geográfico Agustín Codazzi (IGAC), adoptará metodologías y modelos de actualización masiva de valores catastrales rezagados, que permitan por una sola vez realizar un ajuste automático de los avalúos catastrales de todos los predios del país. Como una de las posibles implicaciones de la desactualización catastral de municipios en el país, se encuentra la baja recaudación, debido a avalúos catastrales que subestiman los valores reales del mercado inmobiliario. En este sentido, se tiene la necesidad de reducir el rezago de valores implica con el fin de tener valores catastrales más cercanos con los valores reales del mercado.

Dada la necesidad establecida y contando con información económica recolectada de fuentes primarias y secundarias se plantea el uso de modelos estadísticos que permitan realizar la predicción de valores de terreno de manera masiva. Todo lo anterior se planea realizar a partir de información disponible actualmente por fuentes internas o externas.

3) Alcance

Para definir el alcance del trabajo realizado, se define el Universo de estudio como el conjunto de predios en zonas rurales en el país. Dentro de esta definición, la población objetivo se define como los predios rurales en municipios bajo gestión del IGAC, en departamentos donde se tiene información disponible en la base de datos consolidada. Esto quiere decir que en departamentos sin la información suficiente no se realizará el proceso de modelación y por ende no se tendrán valores propuestos. Adicionalmente, de esta población objetivo se excluyen predios con características específicas dadas a partir del componente temático. En adelante, en el documento la población objetivo se suele referir como marco o marco muestral.

4) Objetivos

En esta sección se plantean los objetivos referentes al proyecto que se describe en este documento. El trabajo realizado se enmarca dentro de la respuesta al requerimiento planteado por el artículo 49 del PND. Con base en lo anterior se plantean tanto los objetivos generales y específicos descritos en las secciones [Sección 4.1](#) y [Sección 4.2](#), respectivamente.

4.1 General

Generar la predicción de valor de terreno comercial para predios en suelo rural en municipios bajo gestión del IGAC utilizando modelos estadísticos, a partir de información disponible recolectada por el OIC.

4.2 Específicos

Para el cumplimiento del objetivo general se plantean los siguientes objetivos específicos:

1. Consolidar y procesar fuentes de información.
2. Realizar exploraciones para definir la cantidad óptima de modelos a desarrollar con base en la distribución espacial de la información.
3. Plantear las metodologías a utilizar dentro del proceso de modelación.

4. Seleccionar las métricas más adecuadas para evaluar y seleccionar el modelo con mejor ajuste.
5. Implementar procesos geoestadísticos para minimizar los errores de predicción de los modelos.
6. Generar la predicción del valor de terreno a nivel de predio a partir de la base de variables exógenas de la población objetivo.
7. Generar valores de referencia para las Zona Homogéneas Goeconómicas vigentes a partir de los valores de terreno.

5) Marco teórico

En esta sección se busca plantear el marco teórico que sustenta el desarrollo de los modelos estadísticos para la predicción de valores de terreno. Se incluyen los conceptos y definiciones necesarios para la comprensión de los modelos.

Como marco de referencia, se consideran estudios como el de Kontrimas and Verikas (2011), el cual compara métodos tradicionales y de inteligencia computacional para avalúos masivos de bienes raíces, destacando el uso de regresión lineal simple frente a regresiones de máquina de soporte vectorial y un perceptrón multicapa aplicados a datos oficiales del registro de Lituania. Ho, Tang, and Wong (2021) aplica algoritmos de aprendizaje automático (máquinas de soporte vectorial, bosques aleatorios y potenciación del gradiente) para valorar precios de propiedades usando 40,000 transacciones de vivienda en Hong Kong recolectadas durante 18 años. Carranza et al. (2022) utiliza regresiones a partir de bosques aleatorios, bosques aleatorios para percentiles y potenciación del gradiente para realizar la predicción de valores con base en información de Openstreetmap y valores comerciales del Mapa de Valores de América Latina. Por otro lado, Córdoba et al. (2021) propone el uso de regresión a partir de un bosque aleatorio espacial para percentiles (sQRF) para la valuación masiva de tierras rurales y se evalúa con datos de Córdoba, Argentina. Jafary et al. (2024) compara regresiones basadas en la potenciación extrema del gradiente (XGBoost), máquinas de soporte vectorial, bosques aleatorios y redes neuronales profundas para la valoración de la tierra en el área metropolitana de Melbourne, Australia, donde se resalta que la metodología de mejor rendimiento fue la Potenciación extrema del gradiente.

Por otro lado, como referentes nacionales, se resalta el trabajo elaborado por Toloza Delgado (2020), que modela de manera conjunta la media y la varianza del proceso de interés a partir de modelos semiparamétricos autorregresivos con dependencia espacial en la variable respuesta aplicado con valores recolectados en Bogotá, Colombia. En Catastro Distrital UAEDC (2023a) se describe el proceso elaborado para la actualización de valores económicos en Cartagena, Colombia, a partir de modelos aditivos generalizados y el proceso de actualización catastral en Bogotá, Colombia, descrito

en Catastro Distrital UAECD (2023b), donde se modela el valor integral de construcción (valor total dividido por el área construida) para predios en propiedad horizontal a partir de modelos lineales generalizados.

Dada la revisión realizada, en este contexto se define el campo aleatorio espacial

$$\{Z(\mathbf{s}): \mathbf{s} \in D \subset \mathbb{R}^2\},$$

donde D es un conjunto continuo y fijo. Por continuo se entiende que $Z(\mathbf{s})$ puede observarse en cualquier punto dentro del conjunto D . De esta forma, $Z(\mathbf{s})$ representa la variable aleatoria referente al valor de terreno por hectárea en la locación \mathbf{s} y esta coordenada hace referencia a la ubicación geográfica. De acuerdo con la descripción dada, el campo aleatorio anteriormente definido hace referencia a datos en el contexto de geoestadística, haciendo la diferenciación con otros tipos de datos georreferenciados como datos de área (lattice) o patrones puntuales. La diferencia de estos tres casos y una revisión detallada relacionada se puede consultar en Cressie (2015) y Schabenberger and Gotway (2017). En Pebesma and Bivand (2023) ilustra la realización de diferentes análisis en el contexto de la estadística espacial utilizando R (R Core Team 2023).

En el contexto de la geoestadística, un supuesto que se suele hacer sobre $Z(\mathbf{s})$ es que es una variable aleatoria estacionaria de segundo orden. Este supuesto implica que la media y la covarianza de $Z(\mathbf{s})$ no dependen de la ubicación \mathbf{s} . Adicionalmente, la covarianza entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$ depende únicamente de la distancia entre \mathbf{s}_i y \mathbf{s}_j . En este sentido, la covarianza entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$ se puede expresar como se muestra en la [Ecuación 1](#), donde $C(h)$ representa la función de covarianza y $h = \mathbf{s}_i - \mathbf{s}_j$ es la distancia entre los puntos \mathbf{s}_i y \mathbf{s}_j . En términos prácticos, usualmente se hace uso de la función de semivariograma, que se denota como $\gamma(h)$ y se presenta en la [Ecuación 2](#). Adicionalmente, se asume que el proceso $\{Z(\mathbf{s})\}$ es isotrópico, lo que implica que la covarianza entre dos puntos depende únicamente de la distancia entre ellos y no de la dirección (Cressie and Wikle 2011; Cressie 2015; Schabenberger and Gotway 2017).

$$\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) = C(h) \quad (1)$$

$$\gamma(h) = C(\mathbf{0}) - C(h) \quad (2)$$

La manera como se procede para realizar la predicción de $Z(\mathbf{s}_0)$, donde $\mathbf{s}_0 \in D$ es un punto no observado parte de la revisión de los supuestos de estacionariedad en media y varianza. En el caso en que la media no sea constante, en la literatura se plantean alternativas para realizar la predicción de $Z(\mathbf{s}_0)$, como el kriging universal, el kriging simple, entre otros. Este supuesto de media constante sobre todo el dominio D usualmente no es realista, motivo por el cual, tal como se plantea en Schabenberger and Gotway (2017) y Cressie and Wikle (2011), $Z(\mathbf{s})$ se puede descomponer tal como se muestra en la [Ecuación 3](#), donde $f(\mathbf{X}(\mathbf{s}), \mathbf{s})$ es una función determinista, en términos de la locación \mathbf{s} y un conjunto de una o más variables independientes georreferenciadas $\mathbf{X}(\mathbf{s})$. Este término se puede denominar un componente de regresión, mientras que $\varepsilon(\mathbf{s})$ es un componente estocástico, espacialmente dependiente y que representa la variabilidad no explicada por la función determinista. Otra forma en que se suelen describir estos

componentes es como el componente que describe la tendencia en un sentido macro y la variabilidad en un sentido micro, respectivamente. Por último $f(\mathbf{X}(\mathbf{s}), \mathbf{s})$ y $\varepsilon(\mathbf{s})$ se asumen independientes.

$$Z(\mathbf{s}) = f(\mathbf{X}(\mathbf{s}), \mathbf{s}) + \varepsilon(\mathbf{s}) \quad (3)$$

De acuerdo con lo expuesto en la [Ecuación 3](#), la predicción de $Z(\mathbf{s}_0)$ se puede realizar a partir de $f(\mathbf{X}(\mathbf{s}_0), \mathbf{s}_0)$ y $\varepsilon(\mathbf{s}_0)$, donde el primer sumando se modela como un componente de regresión y el segundo se modela con la predicción de un campo aleatorio espacialmente dependiente.

Reescribiendo $f(\mathbf{X}(\mathbf{s}), \mathbf{s})$ como se muestra en la Ecuación 4, se pueden utilizar diferentes métodos de regresión para la evaluación del mejor ajuste. En este sentido, $S(\mathbf{s})$ representa la variable respuesta y las diferentes metodologías buscan explicar la variabilidad de $S(\mathbf{s})$ a partir de un conjunto de variables independientes $\mathbf{X}(\mathbf{s})$ y la coordenada \mathbf{s} . En la literatura y en los paquetes de software existentes en la actualidad se encuentra una variedad de métodos a utilizar para esta tarea. Dentro de estas metodologías se encuentran métodos tales como los que se basan en árboles de decisión (Brownlee 2016; Bonaccorso 2018; Mohammed, Khan, and Bashier 2016), redes neuronales, métodos de regresión tradicionales (Nelder and Wedderburn 1972; Melo, López, and Melo 2007; Dobson and Barnett 2008; Montgomery, Peck, and Vining 2012; Ravishanker, Chi, and Dey 2021), entre otros.

$$S(\mathbf{s}) = f(\mathbf{X}(\mathbf{s}), \mathbf{s}) \quad (4)$$

En el caso de los modelos lineales tradicionales, $S(\mathbf{s})$ se modela a partir de una combinación lineal de las variables independientes $\mathbf{X}(\mathbf{s})$, tal como se muestra en la [Ecuación 5](#). En este caso, β_0 es el intercepto del modelo, $\beta_1, \beta_2, \dots, \beta_p$ son los coeficientes asociados a las variables independientes y $\varepsilon(\mathbf{s})$ es el término de error.

$$S(\mathbf{s}) = \beta_0 + \beta_1 X_1(\mathbf{s}) + \beta_2 X_2(\mathbf{s}) + \dots + \beta_p X_p(\mathbf{s}) + \varepsilon(\mathbf{s}) \quad (5)$$

A diferencia de los métodos basados en el potenciado o en el ensamblado de árboles de decisión, tales como el bosque aleatorio o el aumento de gradiente, los métodos de regresión tradicionales asumen una relación lineal entre la variable respuesta y las variables independientes. Por este motivo, para que estos modelos tengan la posibilidad de capturar relaciones no lineales, se pueden considerar variantes que incluyen los modelos generalizados aditivos (Bowman and Azzalini 1997; Wood 2017), en donde, de acuerdo con la relación entre la variable respuesta, una variable independiente y un conjunto de funciones base definido, se podría realizar una mejor aproximación al comportamiento que se visualice en un diagrama de dispersión. Para nuestro ejercicio se utilizaron funciones B -spline y bases de polinomios.

En este sentido, para el proceso de modelamiento de $S(\mathbf{s})$ se utilizaron diferentes metodologías, tales como las que se describen a continuación:

- Red neuronal de una sola capa,

- Árboles potenciados,
- Bosque aleatorio,
- Conjuntos de árboles de decisión,
- Árboles de decisión,
- K-vecinos más cercanos,
- Máquinas de soporte vectorial con bases de funciones polinómicas y con funciones de bases radiales,
- Splines de regresión adaptativa multivariada (MARS),
- Modelos aditivos generalizados.

Con base en el modelo definido, se procede a realizar el cálculo de los residuales $\varepsilon(\mathbf{s})$, tal y como se muestra en la [Ecuación 6](#). Con base en estos residuales se procede a realizar el ajuste del semivariograma que se muestra de la [Ecuación 2](#). En términos prácticos, en la literatura se tienen estimadores para $\gamma(h)$, los cuales se pueden ajustar a partir de la función de semivariograma empírico. En general, a partir de este último se genera la estimación de los parámetros requeridos para los modelos teóricos de semivariograma, los cuales consisten de los parámetros de rango, silla y pepita. Con base en el modelo empírico, se puede conocer la estructura de la función de covarianza del proceso $\{\varepsilon(\mathbf{s})\}$ y con base en esta se puede realizar la predicción de $\varepsilon(\mathbf{s}_0)$ utilizando el predictor dado a partir de la metodología de kriging ordinario (Cressie 2015; Schabenberger and Gotway 2017).

$$\varepsilon(\mathbf{s}) = Z(\mathbf{s}) - f(\mathbf{X}(\mathbf{s}), \mathbf{s}) \quad (6)$$

Para ilustrar el proceso de modelamiento y ajuste del semivariograma se presenta en la [Figura 1](#) el semivariograma a partir de datos simulados. En este caso, se generaron datos con base en un modelo de semivariograma esférico. En la figura se muestra la estimación del semivariograma empírico en los puntos y el modelo de semivariograma teórico ajustado en la línea punteada. Para este ejercicio y para el resto de procesamientos relacionados con el componente espacial se hizo uso del paquete gstat (Gräler, Pebesma, and Heuvelink 2016) en R.

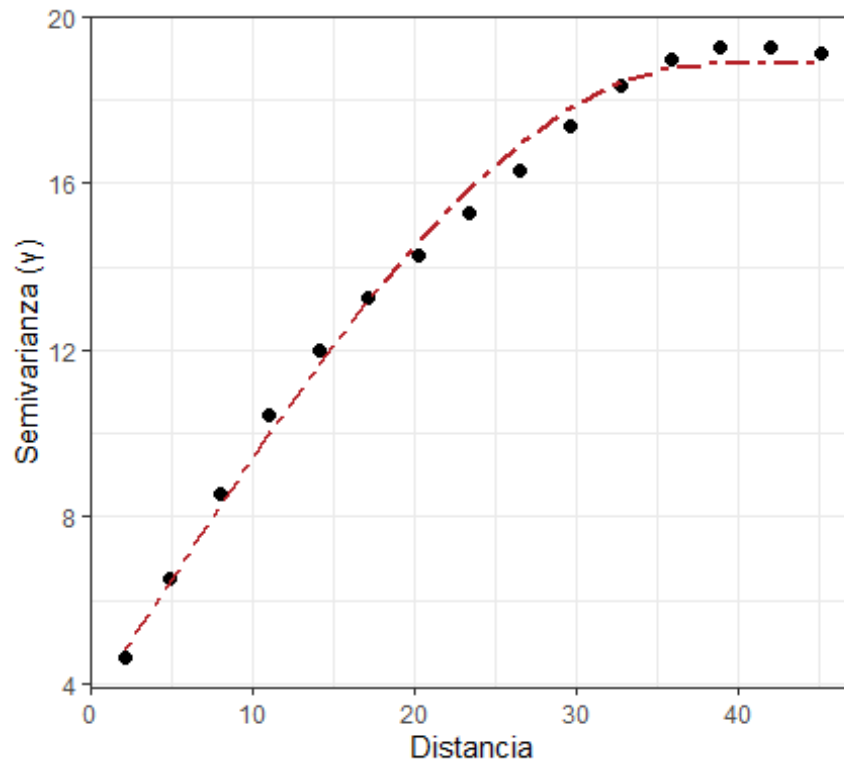


Figura 1: Semivariograma a partir de datos simulados.

6) Desarrollo metodológico

De acuerdo con los objetivos establecidos en este documento, se busca generar predicciones de valores de terreno para los predios rurales contenidos en la población objetivo.

Para este objetivo se requieren insumos, a partir de los cuales se elabora una base de modelamiento, la cual permite realizar el entrenamiento requerido por cada metodología. En este contexto, a partir de “el entrenamiento” se realizan las estimaciones requeridas para poder generar la predicción deseada.

Con el fin de generar la base de modelamiento, el OIC del IGAC ha dispuesto diferentes tablas con las que se cuenta para realizar su procesamiento y adaptación, y de esta manera obtener el conjunto de datos requerido en el proceso de modelamiento. Los insumos dispuestos en esta etapa corresponden a ofertas, avalúos comerciales realizados por entidades privadas o públicas y transacciones inmobiliarias registradas ante la Superintendencia de Notariado y Registro (SNR).

En esta sección se describe el proceso de consolidación de la base datos, su respectiva transformación y procesamiento, para luego utilizarla como insumo en el proceso de modelamiento. Posteriormente se detalla el procedimiento utilizado para la determinación de modelos óptimos para la predicción de la variable respuesta.

6.1 Análisis de la información

El OIC dispuso de una variedad de fuentes de información para este ejercicio. Como es de esperarse, cada una de ellas contiene diferentes columnas. Por este motivo, se realizó un proceso de estandarización para determinar aquellas que fueran relevantes dentro del ejercicio. Adicional a este proceso se realiza todo lo relacionado con la georreferenciación (a partir de coordenadas, direcciones o matrículas inmobiliarias) de los registros y asignación del código predial. En la [Tabla 1](#) se presenta la cantidad de registros con asignación de código predial en la base de información consolidada por fuente de información. La totalidad de registros es 2.230.637. Las fuentes con mayor participación son SNR, Entidades Bancarias¹ y Ofertas, con participaciones de 87.6%, 7.6% y 2.9%, respectivamente. Este resumen incluye información para suelo Urbano y Rural. La información de esta base consolidada está dispuesta en [este enlace](#), mientras que los códigos utilizados para el procesamiento realizado se encuentran en el repositorio de GitLab del IGAC en el siguiente [enlace](#).

Tabla 1: Cantidad de registros con asignación de código predial en la base de información consolidada.

FUENTE DE INFORMACIÓN	CANTIDAD REGISTROS	PARTICIPACIÓN
SNR	1.953.357	87.6%
ENTIDADES_BANCARIAS	168.954	7.6%
OFERTAS	65.683	2.9%
SAE	27.752	1.2%
IVP	9.573	0.4%
OFERTAS_CAPTURADAS_CAMPO	3.540	0.2%
SUB_AVALUOS_IGAC	1.046	0%
ECOPETROL	304	0%

¹ Esta categoría hace referencia a avalúos comerciales realizados por diferentes entidades públicas y privadas, incluyendo información recolectada y con convenios realizados por el IGAC.

DANE	250	0%
CAR	169	0%
OFERTAS_CUNDINAMARCA	9	0%
Total	2.230.637	100%

Una aplicación de la información consolidada es la estimación del rezago de valores económicos. Este rezago existe debido a la ausencia de actualización catastral en diferentes municipios del país. De acuerdo con las normativas, los valores catastrales deben estar entre el 60% y el 100% de los valores comerciales, donde estos últimos son los valores por los cuales un predio se tranzaría en el mercado inmobiliario y son los que se consolidan en la base de fuentes de información del OIC.

Como medida de la estimación del rezago de valores económicos, se calcula la razón entre el valor de avalúo catastral y el valor de avalúo comercial. A partir de este cociente se puede calcular el rezago de los valores catastrales en relación con los valores comerciales dispuestos en la información consolidada, que se muestra en la [Ecuación 7](#). En municipios actualizados se espera que el cociente sea cercano o superior al 60%, mientras que a medida que este valor diste del 60%, el valor catastral está siendo subestimado en mayor medida.

$$\text{Cociente} = \text{Avalúo}_\text{Catastral} / \text{Avalúo}_\text{Comercial} \quad (7)$$

En la [Figura 2](#) se presenta la mediana de los cocientes de valores de avalúo calculados a partir de la información consolidada. En este caso, se observa por fuente de información el resultado y en la línea roja el resultado general, que para este caso es de un 33.2%. Adicionalmente, la gran mayoría de fuentes tienen resultados menores a 20%, incluyendo "OFERTAS" y "ENTIDADES_BANCARIAS"², mientras que la fuente de "SNR", que es la de mayor participación, es de 37%. Lo anterior se debe a que, en esta fuente, en algunos casos se reportan transacciones por un menor valor, por temas tributarios. De esta forma, si se desea utilizar esta fuente de información tan numerosa en el proceso de modelamiento, se debe realizar un ajuste a los valores de avalúo comercial.

² Esta categoría hace referencia a avalúos comerciales realizados por diferentes entidades públicas y privadas, incluyendo información recolectada y con convenios realizados por el IGAC.

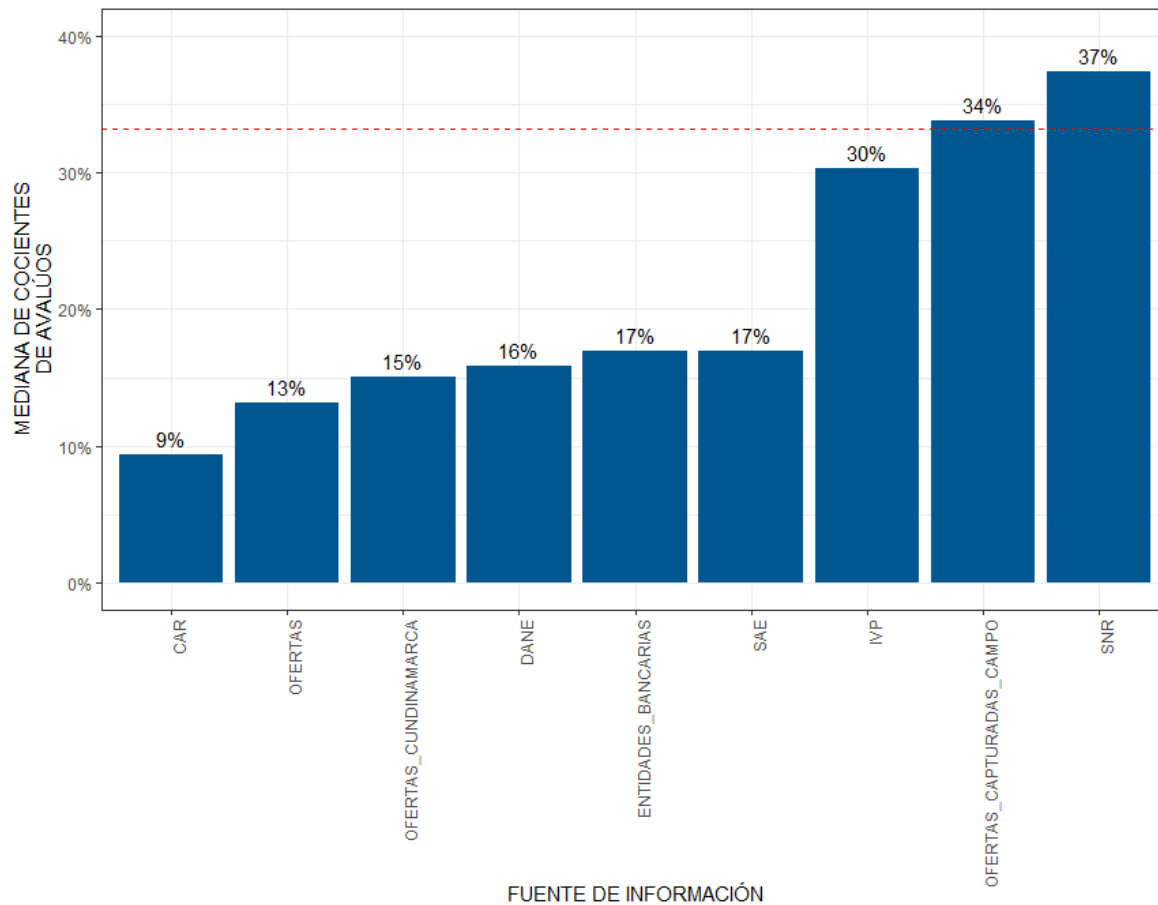


Figura 2: Cociente mediano de valores catastrales versus comerciales por fuente de información.

Para el ajuste de las transacciones de “SNR” se plantearon diferentes escenarios y posibilidades, no obstante en este documento se presenta únicamente el proceso escogido. Dentro de los planteamientos establecidos, se encuentran modelos de regresión y diferentes estimaciones a partir de cálculos del rezago por diferentes desagregaciones. Para esta última metodología, se toma la información de cada departamento, zona (urbano y rural) y año de captura de la información y se calcula por separado lo que se muestra en la [Ecuación 8](#) y la [Ecuación 9](#), donde $Avalúo_Catastral_{SNR_k}$ y $Avalúo_Comercial_{SNR_k}$ son los valores de avalúo catastral y comercial, respectivamente para el k -ésimo registro de la fuente de SNR. Por otro lado, $Avalúo_Catastral_{OF_k}$ y $Avalúo_Comercial_{OF_k}$ son los valores de avalúo catastral y comercial, respectivamente para el k -ésimo registro de las otras fuentes de información (agregando todas las fuentes menos SNR).

$$Cociente_{SNR} = Mediana(Avalúo_Catastral_{SNR_k} / Avalúo_Comercial_{SNR_k}) \quad (8)$$

$$Cociente_{OF} = Mediana(Avalúo_Catastral_{OF_k} / Avalúo_Comercial_{OF_k}) \quad (9)$$

El índice aplicado para cada departamento, zona y año de captura de la información se muestra en la [Ecuación 10](#). Para validar los resultados del ajuste, se replica la gráfica de la [Figura 2](#), pero esta vez con los valores ajustados. En la [Figura 3](#) se observa que la mediana de los cocientes de valores de avalúo calculados a partir de la información aplicando el incremento dado por el ajuste de SNR. En esta gráfica se observa que el resultado general pasó a ser de 19.5%, lo que sugiere que el ajuste realizado es adecuado y que los valores de avalúo comercial de SNR se encuentran más cercanos a los valores de avalúo comercial de las otras fuentes de información.

$$Ajuste_{SNR} = \frac{Cociente_{OF}}{Cociente_{SNR}} \quad (10)$$

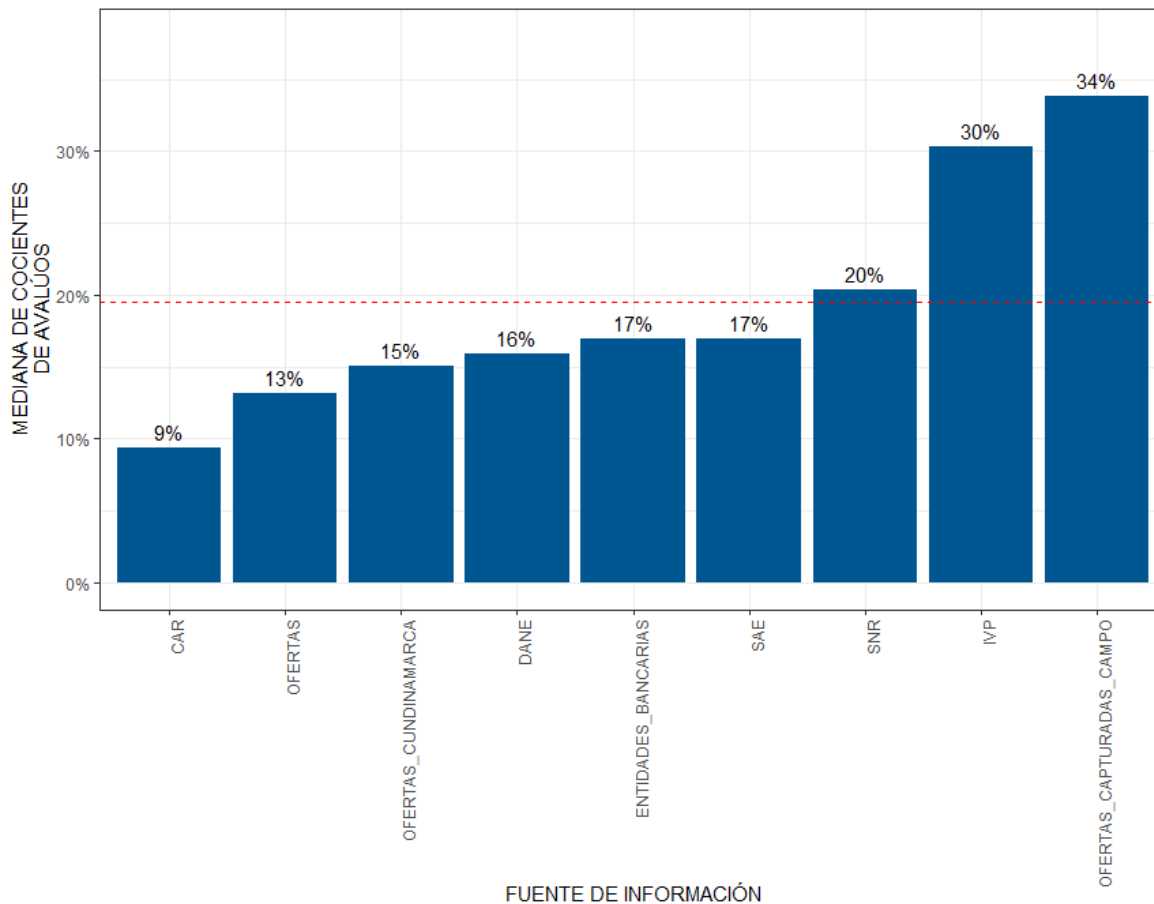


Figura 3: Cociente mediano de valores catastrales versus comerciales por fuente de información posterior al ajuste de SNR.

A partir de los 2.230.637 registros disponibles, se realiza un proceso de depuración con el fin de determinar aquellos casos que contengan información relacionada con valores de terreno. Esto se

debe a que la mayoría de las fuentes reportan los valores totales relacionados con cada predio. A manera de ejemplo, en el caso de una oferta que se publica en portales inmobiliarios, se reporta el valor total de la propiedad, más no un desagregado de valores de terreno y construcción. Para este fin, se realiza una búsqueda de aquellos casos que incluyan en su descripción o en su dirección las palabras “FINCA” o “LOTE”, puesto que, para esos casos, el valor total se puede presumir que corresponde en su totalidad o gran mayoría al terreno. Adicionalmente, para los registros en donde el cociente entre área de construcción y área de terreno es menor o igual al 1%, se asume que el valor reportado puede ser atribuible al valor del terreno. A partir de este conjunto de criterios, se determina una base de datos que incluye esta característica y que es fundamental para el proceso de modelamiento, puesto que el valor de terreno representa la variable respuesta.

En esta etapa se realizaron procesamientos adicionales referentes a controles de calidad automatizados a partir de cartas de control multivariadas (Scrucca 2004; Montgomery 2020) (en términos de los atributos de área de terreno y valor de terreno por hectárea) con el fin de excluir registros con valores atípicos, para posteriormente proceder a la eliminación de registros duplicados a nivel de predio, donde este último paso se realiza primero por fuente de información y posteriormente por fecha de captura de la información. En la [Tabla 2](#) se presenta la jerarquía de las fuentes de información utilizada para la eliminación de registros duplicados, luego, de acuerdo con este criterio, si un predio cuenta con dos registros de fuentes de información diferentes, se conserva el registro de la fuente con menor valor de jerarquía de acuerdo con la tabla indicada. Si el predio tiene más de un registro dentro de la fuente de información definida se toma el registro que tiene una fecha más reciente.

Tabla 2: Jerarquía de las fuentes de información utilizada para eliminar registros duplicados a nivel de predio.

FUENTE DE INFORMACIÓN	JERARQUÍA
AVALUOS_PUNTUALES	1
OFERTAS_CAPTURADAS_CAMPO	2
ENTIDADES_BANCARIAS	3
IVP	4
CAR	5
SAE	6
SNR	7
ECOPETROL	8

OFERTAS	9
DANE	10
SUB_AVALUOS_IGAC	11

Tabla 3: Cantidad y participación de predios en la base de modelación resultante por Fuente de Información.

FUENTE DE INFORMACIÓN	CANTIDAD REGISTROS	PARTICIPACIÓN
SNR	237.642	96.8%
ENTIDADES_BANCARIAS	5.082	2.1%
OFERTAS	2.276	0.9%
AVALUOS_PUNTUALES	353	0.1%
OFERTAS_CAPTURADAS_CAMPO	133	0.1%
CAR	79	0%
SAE	29	0%
DANE	4	0%
OFERTAS_CUNDINAMARCA	4	0%
Total	245.602	100%

En la [Tabla 3](#) se presenta la cantidad de predios en la base de modelación resultante por fuente de información. Las fuentes con mayor participación son SNR, Entidades Bancarias y Ofertas, con participaciones de 96.8%, 2.1% y 0.9%, respectivamente. En total, se cuenta con 245.602 registros en la base de modelación. Por otro lado, en la Tabla 4 se presenta la cantidad de predios y participación en la base mencionada por departamento. Los departamentos con mayor participación son Boyacá, Santander y Nariño, con participaciones de 17%, 9.1% y 8.5%, respectivamente.

Tabla 4: Cantidad y participación de predios en la base de modelación resultante por Departamento.

NOMBRE DEL DEPARTAMENTO	CANTIDAD REGISTROS	PARTICIPACIÓN
BOYACA	41.767	17%
SANTANDER	22.275	9.1%
NARIÑO	20.824	8.5%
TOLIMA	19.735	8%
CAUCA	15.289	6.2%
HUILA	11.857	4.8%
CORDOBA	10.445	4.3%
CALDAS	10.441	4.3%
CUNDINAMARCA	10.413	4.2%
META	10.225	4.2%
NORTE DE SANTANDER	9.747	4%
PUTUMAYO	8.675	3.5%
CESAR	7.742	3.2%
SUCRE	5.906	2.4%
ARAUCA	5.529	2.3%
CASANARE	4.966	2%
CAQUETA	4.529	1.8%
RISARALDA	4.206	1.7%
MAGDALENA	3.784	1.5%
VALLE DEL CAUCA	3.696	1.5%
BOLIVAR	3.486	1.4%
GUAVIARE	2.950	1.2%
QUINDIO	2.446	1%

LA GUAJIRA	1.384	0.6%
CHOCO	883	0.4%
ATLANTICO	858	0.3%
ARCHIPIELAGO DE SAN ANDRES	748	0.3%
VICHADA	380	0.2%
AMAZONAS	362	0.1%
GUAINIA	52	0%
VAUPES	2	0%
Total	245.602	100%

Con base en la información descrita en la [Tabla 3](#) y en la [Tabla 4](#) se procedió a la elaboración de los diferentes modelos. En esta base de datos se tiene principalmente la información de valores económicos y el código de identificación predial. Ahora es necesario complementar esta información con las variables explicativas que permitan realizar el modelamiento. Estas provienen de fuentes de información en formato alfanumérico y geográfico. Ese procesamiento y armado de bases se expone en la [Sección 6.2](#).

6.2 Preparación y transformación de los datos

Para la construcción de los modelos econométricos, fue necesario adaptar la información de los predios para que reflejara cómo los diferentes factores pueden influir en el valor del terreno conforme estos cambian. De esta manera, se conformó una base de datos compuesta por fuentes de información internas, externas y otras que debieron construirse, como los componentes transaccional y geográfico.

El **componente transaccional** incluyó datos recopilados de fuentes inmobiliarias, avalúos bancarios, registros de SNR, avalúos puntuales, entre otras entidades relevantes. Por su parte, el **componente geográfico** abarcó aspectos normativos relacionados con el uso del suelo, la clasificación del terreno, la capacidad de uso del suelo y otros factores físicos pertinentes. Este componente se complementó con información específica de cada municipio, obtenida tanto de fuentes externas como propias, además de la información proveniente del R1.

En esta sección se detalla cómo se conformó la base de datos a nivel de predio para el componente geográfico y cómo se integró con los demás componentes. Dentro del componente geográfico, se cruzó la información alfanumérica a nivel predial con las capas geográficas disponibles en el

IGAC. Este cruce permitió obtener características específicas del predio asociadas a cada una de las capas geográficas. Las capas utilizadas fueron:

- Corine Land Cover.
- Zona Homogénea Geoeconómica.
- Zona Homogénea Física.
- Parques Naturales.
- Uso Principal del Suelo.
- Vocación Uso del Suelo.
- Clase Agrológica.
- Frontera Agrícola.
- Conflicto Uso del Suelo.
- Terreno.
- Sistema Nacional de Áreas Protegidas.
- Área Homogénea de Tierra.
- Resguardos Indígenas.

Para obtener la marcación a nivel predial de cada una de las capas temáticas se estableció el siguiente protocolo de procesamiento para garantizar la calidad y consistencia del producto generado.

1. Preparación de los insumos: al trabajar con información generada por otras dependencias del IGAC y otras entidades, se hace necesario realizar ciertas verificaciones de esta y en algunos casos ejecutar procesamientos previos como: proyección al sistema de coordenadas CTM12 o depuración de atributos no requeridos en el análisis.
2. Marcación a nivel predial con las características temáticas: la marcación se generó con el fin de identificar la relación de cada predio con las características propias del territorio. Al ser información de carácter espacial y un análisis enmarcado en el componente geográfico, se hace uso de software especializado en el análisis SIG y se ejecuta la herramienta más adecuada para la asignación de atributos a cada predio que en este caso es Intersección.
3. Completitud de la información a nivel predial: uno de los comportamientos a tener en cuenta en la generación de marcaciones prediales es la cobertura a nivel nacional de la información temática; varias de las capas temáticas procesadas hacen referencia a información sectorizada donde el cubrimiento a nivel país no es del 100% y por ende dichas

marcaciones no contenían la totalidad de registros prediales considerados para la generación de modelos. Se evidenció la importancia de garantizar la completitud a nivel predial en cada una de las marcaciones y se generó un procesamiento intermedio en el software R para subsanar esta necesidad en cada una de las marcaciones generadas.

4. Generación del archivo final de la marcación: en el marco de la interoperabilidad, se generó un archivo plano en formato plano, donde se asigna a cada número predial nacional las características a ser analizadas de cada capa temática.

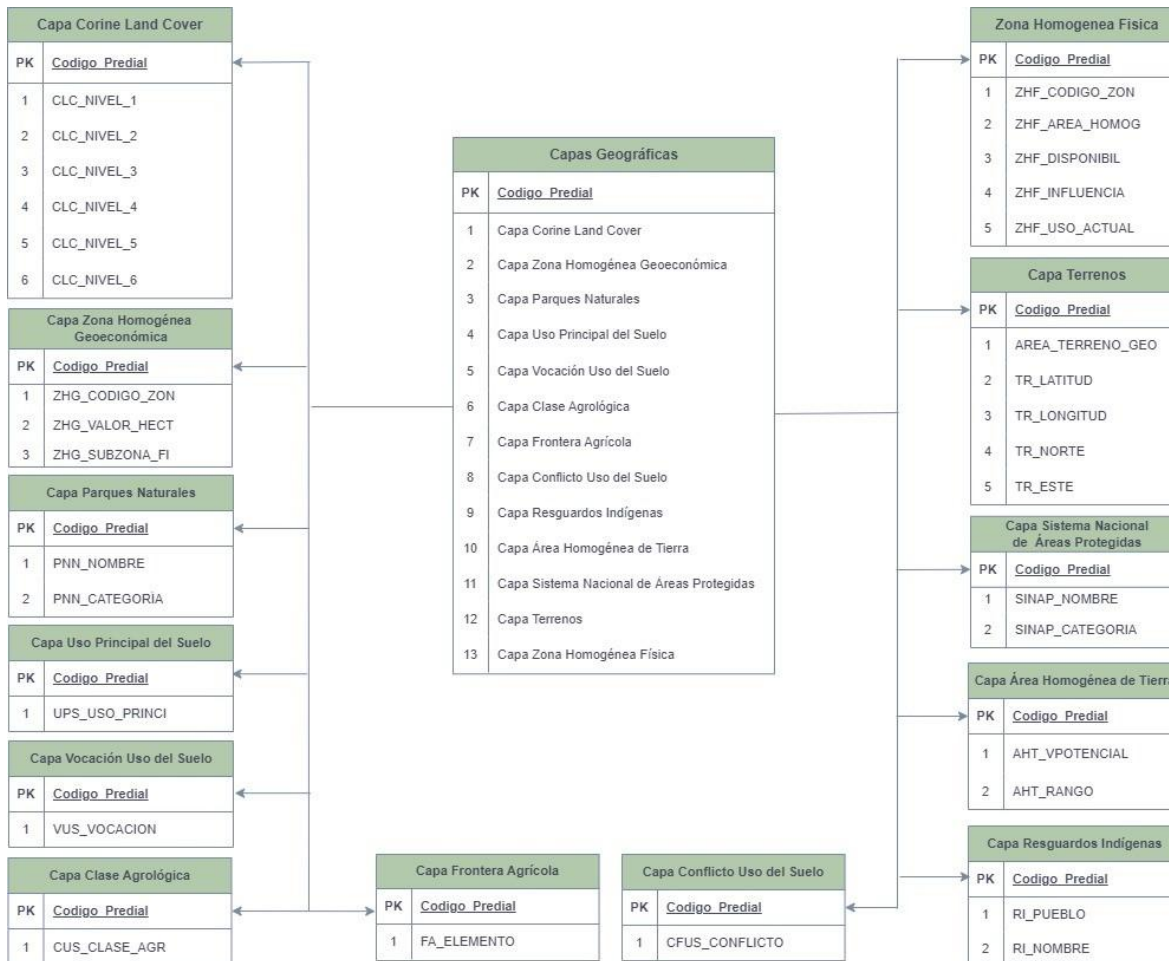


Figura 4: Diagrama del modelo de entidad relación para la elaboración de la base de modelamiento a partir de capas geográficas.

Las características de las capas se asociaron al predio considerando aquella que predominara en mayor área. Es decir, si un predio abarcaba varios tipos de uso de suelo, se tomó en cuenta únicamente el uso que ocupaba la mayor proporción del terreno. Tras realizar este proceso para

todas las capas geográficas, se consolidó la información en un solo insumo a nivel de predio, que incluyó las características de cada capa geográfica. La [Figura 4](#) presenta las características que conforman cada una de estas capas.

Una vez obtenida la información a nivel predial para el componente geográfico, el componente transaccional y el R1, se procedió a unir las bases de datos. Posteriormente, se integraron las fuentes de información externa, como el [índice de riesgo victimización](#), el [Observatorio de Ciudades Modernas](#) y la base Maestra del IGAC, a nivel de municipio.

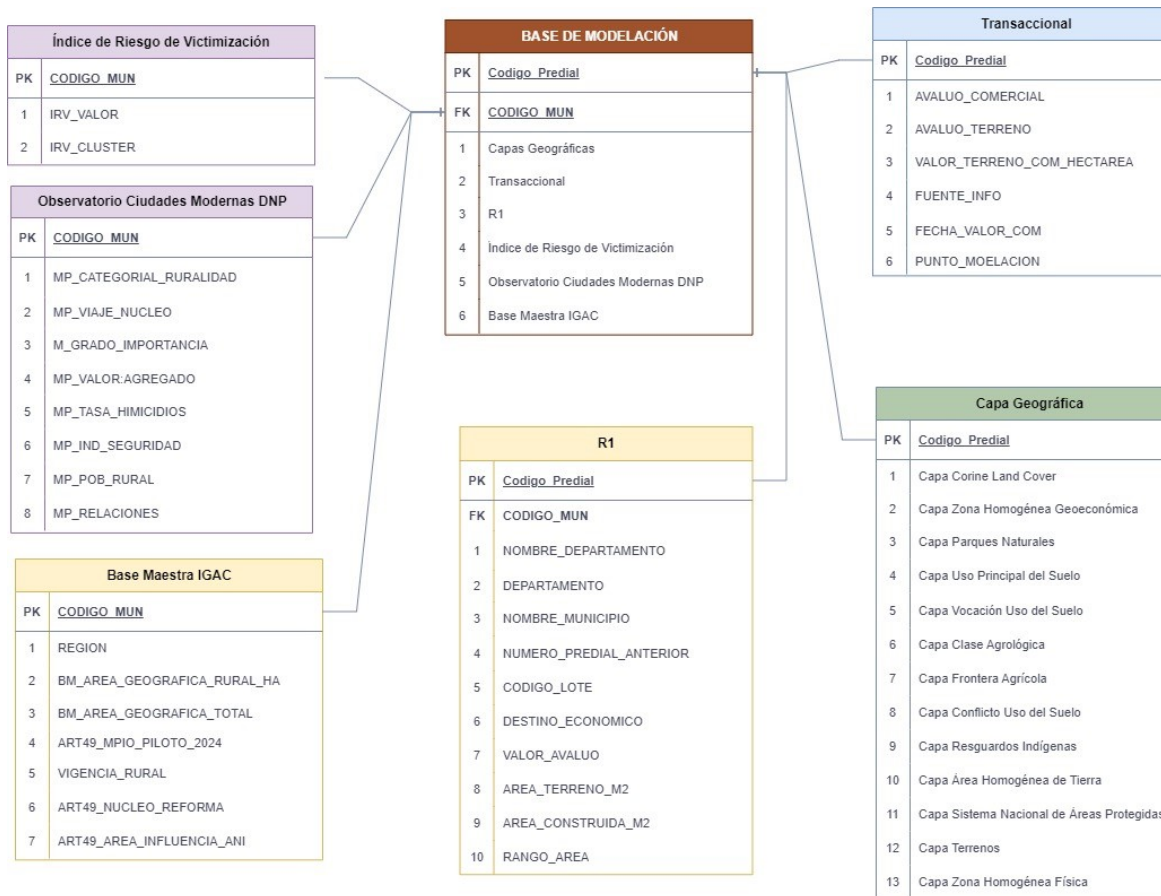


Figura 5: Diagrama del modelo de entidad relación para la elaboración de la base de modelamiento a partir de capas geográficas y tablas alfanuméricas.

Con estas uniones, se dio por conformada la base de datos final. La [Figura 5](#) presenta la estructura de la base de modelación, junto con las variables asociadas a cada una de las fuentes que la componen. En la [Sección 6.3](#) se detalla el proceso de modelamiento y los resultados obtenidos.

6.3 Desarrollo del modelo y/o metodología

Para generar los valores de terreno por hectárea, se plantearon modelos estadísticos que permiten predecir el valor de terreno en función de las características de los predios. En este sentido y dada la alta heterogeneidad en términos de valores económicos y de características físicas de los predios, se decidió no realizar un solo modelo para todo el país. De esta manera, se puede lograr un mayor ajuste para describir el comportamiento de la variable respuesta. Esto, sumado a que existen locaciones donde no se cuenta con información disponible, se decidió realizar un modelo por departamento, de forma que dentro un departamento específico se puede generar predicciones del valor de terreno en sitios sin información a partir de las características de los predios con información disponible. Lo anterior se debe a que a nivel municipal se tendría una cantidad considerable de modelos a realizar, lo cual desbordaría las capacidades de tiempo y esfuerzo disponibles, y adicionalmente no permitiría tener en cuenta la relación espacial que puede haber entre sitios contiguos que se encuentren en municipios diferentes. Por otro lado, los modelos municipales generarían la ausencia de predicciones en municipios sin registros en la base de modelación, mientras que, al hacer el trabajo por departamentos, esto se podría subsanar.

Como en cualquier planteamiento de modelos de aprendizaje supervisado se plantean métricas para determinar el rendimiento de los modelos y de esta manera conocer cuál tuvo un mejor acercamiento al problema de trabajo. En este sentido, en la [Sección 6.3.1](#) se muestran las métricas definidas para este trabajo.

6.3.1 Métricas de rendimiento

Para la evaluación de los modelos propuestos se utilizaron diferentes métricas de rendimiento. En la literatura se plantean diferentes métodos para la evaluación de modelos de regresión. Para este caso se utilizaron las métricas de la raíz del error cuadrático mediano (*RMEDSE*) y un pseudo- R^2 , los cuales se presentan en las ecuaciones [Ecuación 11](#) y [Ecuación 12](#), respectivamente. El objetivo en esta etapa es encontrar el modelo que minimice la raíz del error cuadrático mediano y maximice el pseudo- R^2 . El planteamiento del *RMEDSE* se da puesto que la variable de valor de terreno posee una distribución sesgada a la derecha, por tal motivo se considera más apropiada para cuantificar el valor del error en cada metodología. En este caso $Z(\mathbf{s}_i)$ y $\hat{Z}(\mathbf{s}_i)$ representan la variable de valor de terreno real y la predicha, respectivamente. Por otro lado, $Z(\bar{\mathbf{s}}_i)$ y $\hat{Z}(\bar{\mathbf{s}}_i)$ son las medias de las variables de valor de terreno real y predicha, respectivamente.

$$RMEDSE = \sqrt{\text{Mediana}\left(\left(Z(\mathbf{s}_i) - \hat{Z}(\mathbf{s}_i)\right)^2\right)} \quad (11)$$

$$\text{pseudo} - R^2 = \left[\frac{\sum_{i=1}^n (Z(\mathbf{s}_i) - Z(\bar{\mathbf{s}}_i)) (\hat{Z}(\mathbf{s}_i) - \hat{Z}(\bar{\mathbf{s}}_i))}{\sqrt{\sum_{i=1}^n (Z(\mathbf{s}_i) - Z(\bar{\mathbf{s}}_i))^2} \sqrt{\sum_{i=1}^n (\hat{Z}(\mathbf{s}_i) - \hat{Z}(\bar{\mathbf{s}}_i))^2}} \right]^2 \quad (12)$$

6.3.2 Determinación del Valor de la Zona Geoeconómica (ZHG).

Las zonas homogéneas geoeconómicas son espacios geográficos definidos a partir de zonas homogéneas físicas (ZHF) o por las condiciones del entorno que la componen como características del suelo, vocación del suelo, normatividad o clase de suelo. Los predios de una misma ZHG tienen valores unitarios (en terreno) similares en cuanto al precio, establecidos según las condiciones del mercado inmobiliario. Los predios que se encuentran dentro de cada una son los que permiten determinar el valor del terreno, ya que representan la unidad mínima que refleja las condiciones físicas y económicas. Ante esto, para determinar el valor de terreno, se propone utilizar una medida de tendencia central que permita agregar las predicciones prediales a nivel de polígonos de ZHG. De esta forma, se tendrán en cuenta la media, la mediana simples y ponderadas por el área de terreno para proporcionar el valor de la ZHG. A continuación, se presentan las definiciones de las medidas de tendencia central:

- **Media:** Es la suma de los valores de terreno por hectárea de todos los predios dividida por el número total de predios en una ZHG. Esta estadística se utiliza para obtener un valor representativo del conjunto de predios, asumiendo que todos los predios tienen el mismo peso.

$$\text{Media} = \frac{1}{n} \sum_{i=1}^n Z(s_i) \quad (13)$$

- **Mediana:** Es el valor central que divide la distribución de los valores de terreno de los predios en dos partes iguales. Esta medida es más robusta ante la presencia de valores extremos o atípicos, y es particularmente adecuada cuando los valores presentan una distribución sesgada o con alta varianza. Denotando $F_Z(z)$ como la función de distribución acumulada de la variable aleatoria $Z(\mathbf{s}) = Z_{s_k}$, la mediana es el valor Z_{s_k} tal que:

$$\text{Mediana} = \min(Z_{s_k}), \text{ tal que } F_Z(Z_{s_k}) \geq 0.5 \quad (14)$$

- **Media ponderada:** Es una medida en la que se asigna un peso específico a cada predio, dependiendo de su relevancia o características, como el área. Cada valor individual es multiplicado por su peso, y se calcula tomando en cuenta la suma total de los pesos.

$$\text{Media ponderada} = \frac{\sum_{i=1}^n Z_{s_i} w_i}{\sum_{i=1}^n w_i} \quad (15)$$

- **Mediana ponderada:** Es una variación de la mediana en la que se consideran los pesos asignados a cada predio. De esta forma, bajo esta ponderación, los predios con mayor área tendrán una mayor participación para el cálculo de la estadística.

$$\text{Mediana ponderada} = \min(Z_{s_k}), \text{ tal que } \sum_{i=1}^{k-1} \frac{w_i}{\sum_{i=p}^n w_p} < 0.5 \leq \sum_{i=1}^k \frac{w_i}{\sum_{p=1}^n w_p} \quad (16)$$

Donde w_i es el peso o participación en área de terreno del i -ésimo predio.

De esta forma, una vez culmina la selección del modelo y la generación de predicciones a nivel predial, se realiza el cálculo de las estadísticas descritas en esta sección para de esta forma, generar valores a nivel de ZHG que permitan tener una cifra consolidada a nivel de polígono.

7) Resultados y discusiones

Inicialmente, se presenta un resumen descriptivo de la información recolectada. En la [Tabla 5](#), se muestra el conteo de registros en la muestra y en el marco y la participación de la muestra respecto al marco, en donde se observa que 9 departamentos tienen valores comerciales para al menos el 10% de los predios, el municipio con menos valores comerciales es Atlántico con una participación del 3,9%. En la [Figura 7](#) se muestra la cantidad y porcentaje de predios en el marco por departamento, donde se observa que Boyacá es el Departamento con más registros con 564.770 predios, lo que representa el 19% del total, seguido de Nariño con 330.976 predios, lo que equivale al 11%, mientras que Cauca, con 294.208 predios, representa el 10%. En la parte inferior de la distribución se encuentran Atlántico, La Guajira y Vaupés, los cuales tienen menos de 1% de participación, con 22.258, 21.826 y 5.493 predios, respectivamente.

Tabla 5: Cantidad de predios en el marco y en la muestra utilizada.

DEPARTAMENTO	MUESTRA	MARCO	PARTICIPACIÓN DE LA MUESTRA
ATLANTICO	858	22.258	3,9%
BOLIVAR	3.486	78.262	4,5%
BOYACA	41.767	564.770	7,4%
CALDAS	10.441	88.947	11,7%
CAQUETA	4.529	44.217	10,2%
CAUCA	15.289	294.208	5,2%
CESAR	7.742	45.955	16,8%
CORDOBA	10.445	121.498	8,6%
CUNDINAMARCA	10.413	201.640	5,2%
CHOCO	883	11.045	8,0%
HUILA	11.857	142.790	8,3%
LA GUAJIRA	1.384	21.826	6,3%

MAGDALENA	3.784	49.390	7,7%
META	10.225	97.452	10,5%
NARIÑO	20.824	330.976	6,3%
NORTE DE SANTANDER	9.747	75.259	13,0%
QUINDIO	2.446	22.412	10,9%
RISARALDA	4.206	31.688	13,3%
SANTANDER	22.275	226.312	9,8%
SUCRE	5.906	59.438	9,9%
TOLIMA	19.735	215.317	9,2%
VALLE DEL CAUCA	3.696	55.778	6,6%
ARAUCA	5.529	26.182	21,1%
CASANARE	4.966	66.950	7,4%
PUTUMAYO	8.675	60.985	14,2%
VICHADA	380	5.493	6,9%

Por otro lado, la [Figura 6](#) muestra los conteos y participaciones correspondientes en la muestra utilizada. Aunque los patrones se mantienen en parte, se observan algunas diferencias interesantes. Boyacá sigue siendo el departamento con mayor participación en la muestra, con 39.271 predios 17.9%, manteniendo su predominancia, aunque con una proporción ligeramente menor en comparación con su participación en el marco. Santander ocupa el segundo lugar en la muestra con 20.925 predios con un 9.5% de participación. Departamentos como Nariño y Tolima también destacan con una participación del 8.4% cada uno, lo que corresponde a alrededor de 18.500 predios en cada caso.

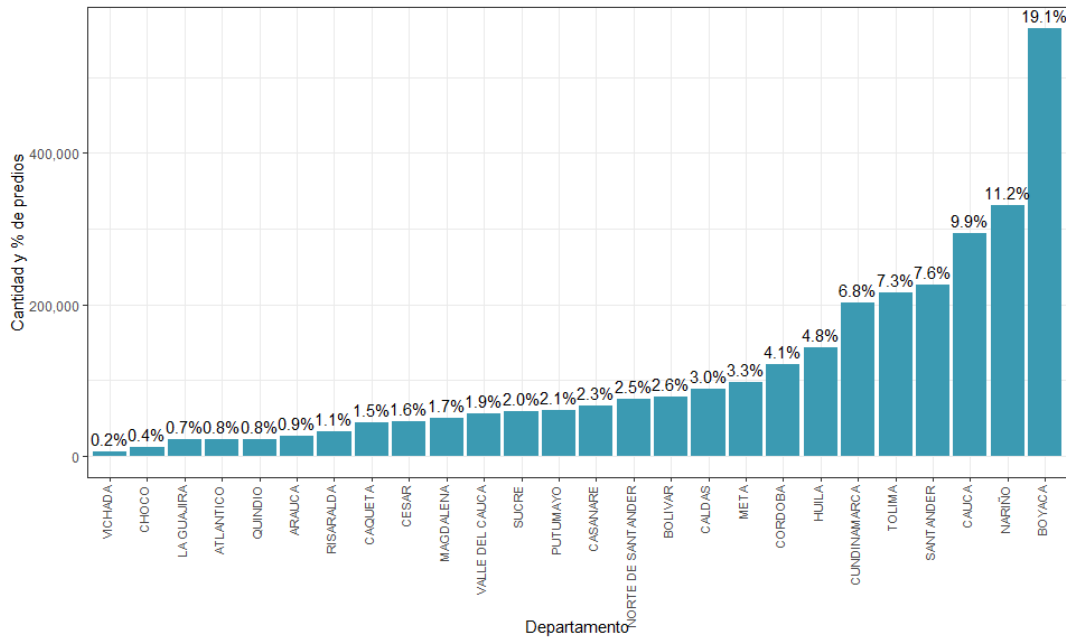


Figura 6: Diagrama de barras correspondiente a la cantidad de predios en el marco por departamento.

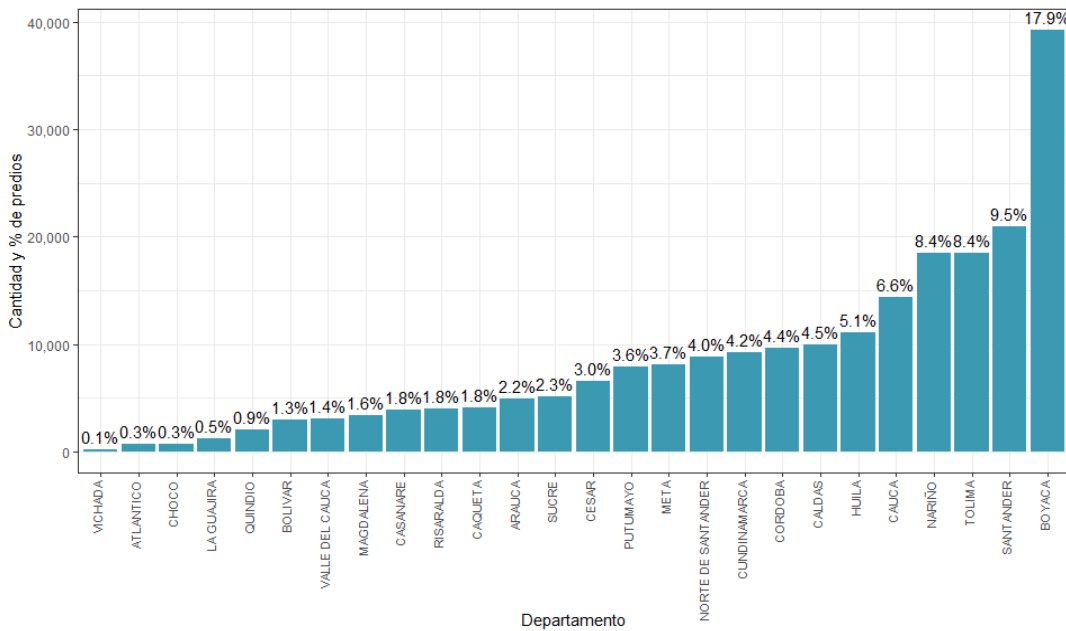


Figura 7: Diagrama de barras correspondiente a la cantidad de predios en la muestra por departamento.

Con base en la información descrita, se procedió a realizar todo el proceso de modelamiento utilizando la muestra y una vez se definió un modelo óptimo, se realizó la predicción sobre el marco, de forma que para los registros allí consignados se generó un valor predicho de terreno por hectárea. El proceso de elaboración y selección de modelos se describe en la sección [Sección 7.1](#).

7.1 Resultados del Proceso de Modelación

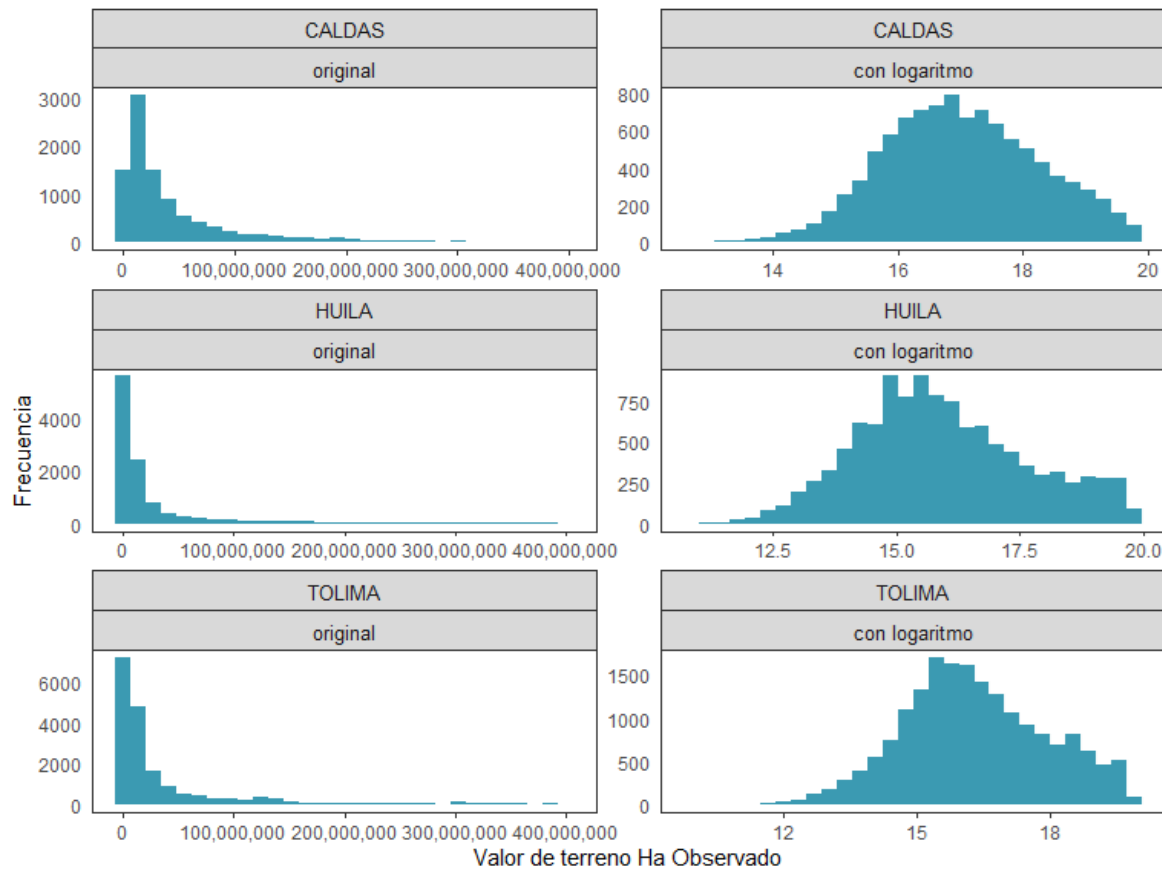


Figura 8: Distribución del valor de terreno por HA para los departamentos con logaritmo y sin logaritmo.

Para dar inicio al proceso de modelación, se analizó la distribución de la variable respuesta. El valor de terreno por hectárea (HA), debido a su naturaleza, presenta una distribución sesgada a derecha. Para corregir este sesgo y mejorar la precisión del modelo, se aplicó una transformación logarítmica a la variable respuesta, con el objetivo de normalizar su distribución. No obstante, para interpretar correctamente los resultados y cuantificar el error, se debe aplicar una transformación inversa a los valores predichos del modelo. Para ilustrar esta situación, la [Figura 8](#) muestra las distribuciones originales de los datos de valor de terreno por hectárea y el logaritmo correspondiente los departamentos de Caldas, Huila y Tolima. Allí se puede apreciar

cómo la transformación logarítmica logra una distribución más simétrica y normal, lo que facilita los procesos subsecuentes.

7.1.1 Aplicación de metodologías

Previo a realizar el trabajo de modelamiento se plantea un flujo de trabajo para la preparación de los datos. Esto permite tener una base de datos estandarizada para el entrenamiento de modelos. Esto se realizó a partir de las librerías `parSNIP`, `recipes` y `tidymodels` en R (Kuhn and Vaughan 2024; Kuhn, Wickham, and Hvitfeldt 2024; Kuhn and Wickham 2020), mientras que el proceso de manejo de tablas se hizo a partir de las librerías anidadas en el paquete de `tidyverse` (Wickham et al. 2019).

Este flujo de trabajo de preprocesamiento realiza una serie de transformaciones sobre un conjunto de datos de entrenamiento para preparar los predictores y la variable objetivo antes del modelado. Primero, se especifica que la variable de interés a predecir es el valor del terreno por hectárea, y se asigna un rol especial a una columna identificadora para excluirla del análisis. Luego, se aplican varias transformaciones a las variables numéricas y categóricas. También se imputan los valores faltantes usando la mediana para las variables numéricas y la moda para las categóricas, se agrupan las categorías poco frecuentes. Además, se omiten los predictores con poca variación y algunas transformaciones se realizan sin afectar el conjunto de validación.

Para el modelamiento se plantean diferentes metodologías, las cuales se mencionan en la [Sección 5](#). Para las metodologías de regresión lineal tales como la regresión `lasso` o `ridge` se añaden pasos al preprocesamiento que permiten modelar relaciones no lineales entre las variables explicativas y la variable respuesta. Las funciones base que se utilizan para esta tarea son bases de polinomios y B-splines.

Un planteamiento que se tuvo en cuenta fue considerar si al modelar todo el departamento se tendrían mejores resultados o si al hacer ejecuciones segmentando la base por características tales como el área de terreno se tendrían mejores resultados. De esta manera, para distintos departamentos se tomaron particiones del área de terreno que se corresponden con tipos de propiedad de la tierra, estas son: Latifundio (mayor a 200 HA), Mediana propiedad (20-200 HA), Microfundios (0-3 HA), Minifundio (3-10 HA) y Pequeña propiedad (10-20 HA).

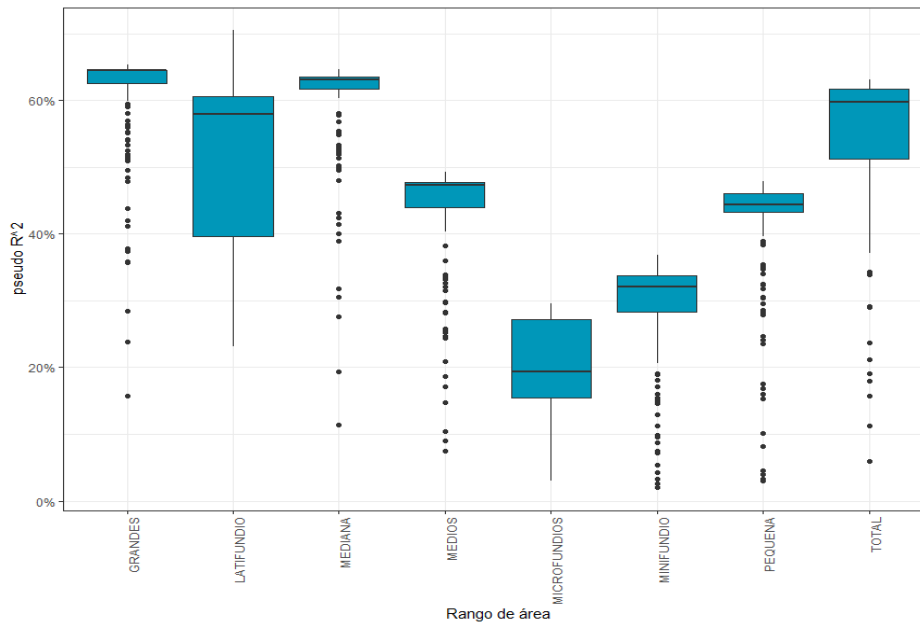


Figura 9: Resultados del pseudo-R cuadrado por rangos de área en el Departamento de Quindío.

El ejercicio se basa en tomar cada una de estas metodologías, asignar una grilla de hiperparámetros para los casos que requieran una optimización en este sentido y hacer la ejecución por departamento y para las combinaciones de departamento y rango de área. En la [Figura 9](#) se muestran los resultados del pseudo- R^2 agregados de todas las metodologías en diagramas de cajas y bigotes tanto por rangos de área como para el total del departamento. En este sentido, se observan rangos cuyos resultados están por debajo del resultado Total, por lo cual se tomó la decisión de hacer un proceso de modelado para el departamento completo. Este patrón descrito se observó para todos los departamentos.

Una vez definido que se iba a adelantar el proceso de elaboración de modelos para todos los predios y no por rangos de área, el paso siguiente es determinar la metodología que mejor explique la variabilidad de la respuesta. Para ello, en la [Figura 10](#) se muestran los resultados desagregando por cada uno de los modelos tenidos en cuenta para el Departamento de Quindío. Allí se observa que los mejores resultados en términos de la métrica de pseudo- R^2 son los modelos de bosque aleatorio (Random Forest), Splines de regresión adaptativa multivariada (MARS) y los modelos de regresión lasso y ridge. Es importante mencionar que este patrón se dio con la mayoría de los departamentos, en términos de que estas metodologías estuvieron entre las que mejores resultados generaron para el pseudo- R^2 . Por este motivo y dando prioridad al principio de parsimonia, se decidió utilizar las regresiones lasso y ridge para la construcción de los modelos, puesto que estas se basan en combinaciones lineales de las variables explicativas y en gran medida pueden ser expuestas al público y al equipo de trabajo con una mayor facilidad.

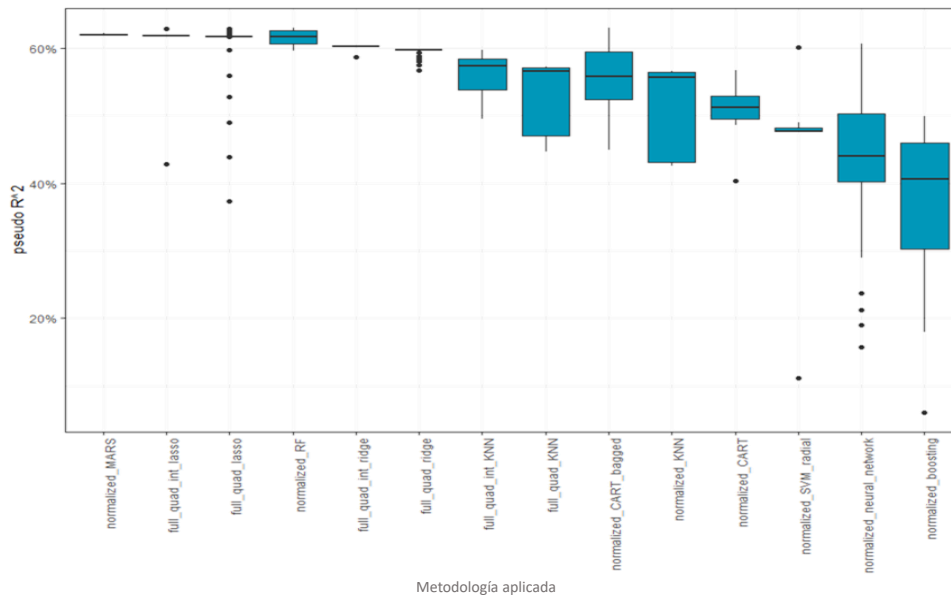


Figura 10: Resultados del pseudo-R cuadrado a partir de diferentes metodologías en el Departamento de Quindío.

En paralelo se tuvo en cuenta la medición del *RMEDSE*. En la [Figura 11](#) se ilustra el resultado de esta métrica para el Departamento de Caquetá.

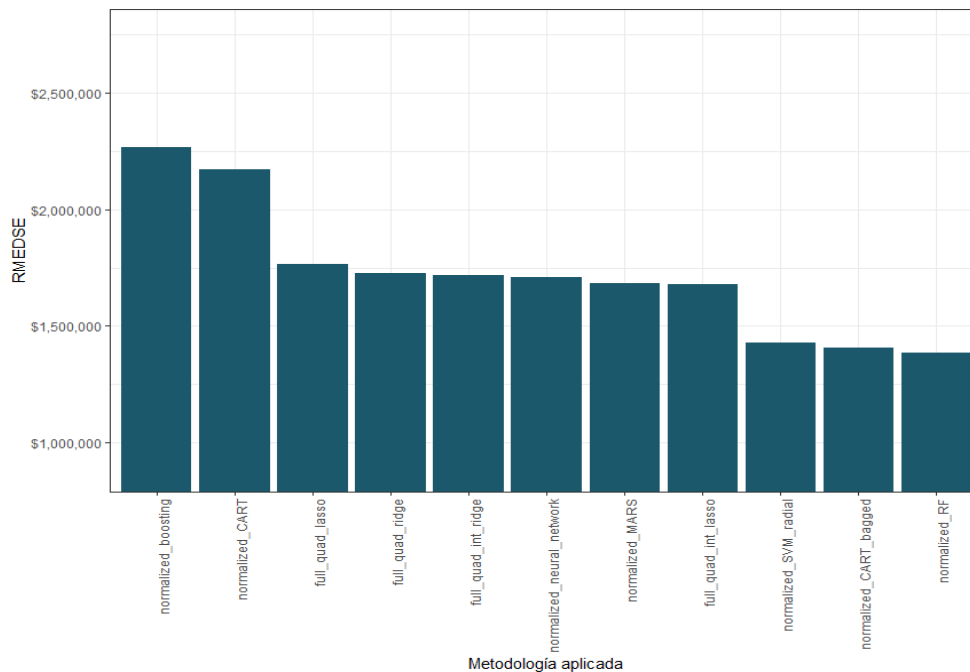


Figura 11: Resultados del RMEDSE cuadrado a partir de diferentes metodologías en el Departamento de Caquetá.

En la [Tabla 6](#) y la [Tabla 7](#) se presentan la mediana del pseudo- R^2 por metodología aplicada y Departamento. Esto se presenta con el fin de reforzar el planteamiento de por qué se seleccionaron las metodologías lasso y ridge como propuestas finales, ya que a pesar de que existen metodologías de regresión con resultados un poco mejores, la diferencia no es muy grande y se prioriza la facilidad de interpretación de los modelos. En la [Tabla 8](#) y la [Tabla 9](#) se presentan la mediana del *RMEDSE* por metodología aplicada y Departamento. En este caso, se observa que las metodologías de regresión lasso y ridge presentan resultados aceptables en términos de la métrica de error cuadrático mediano, al hacer la comparación con otras metodologías.

Tabla 6: Mediana de los resultados del pseudo-R2 por Departamento y metodología aplicada en el proceso realizado-

DEPARTAMENTO	MODELO LINEAL GENERALIZADO CON BASES DE POLINOMIOS	REGRESIÓN LASSO CON BASES DE POLINOMIOS	REGRESIÓN RIDGE CON BASES DE POLINOMIOS	K VECINOS MÁS CERCANOS CON BASES DE POLINOMIOS	MODELO LINEAL GENERALIZADO CON BASES B-SPLINES	REGRESIÓN LASSO CON BASES B-SPLINES	REGRESIÓN RIDGE CON BASES B-SPLINES	K VECINOS MÁS CERCANOS CON BASES B-SPLINES
Boyacá	58%	58%	56%		56%	58%	56%	
Caldas	44%	44%	44%		44%	44%	44%	
Caquetá		74%	72%			74%	72%	
Cauca	70%	70%	68%		70%	70%	68%	
Chocó		74%	74%			70%	74%	
Cundinamarca	64%	64%	64%		64%	64%	64%	
Nariño	66%	66%	64%		66%	66%	64%	
Putumayo		80%	78%			80%	78%	
Quindío		62%	60%	58%		62%	60%	56%
Risaralda		48%	46%			48%	46%	
Tolima	60%	60%	60%		60%	60%	60%	

Tabla 7: Mediana de los resultados del pseudo-R2 por Departamento y metodología aplicada en el proceso realizado.

DEPARTAMENTO	PERCEPTRO N MULTICAPA	ARBOL DE DECISIÓ N	ARBOLES DE DECISIÓ N AGREGADO S	BOSQUE ALEATORI O	MAQUINA DE SOPORTE VECTORIAL CON FUNCIONE S RADIALES	XGBoos t	REGRESIÓN MULTIVARIAD A ADAPTATIVA SPLINE	K VECINOS MÁS CERCANO S
BOYACÁ	58%	58%	60%	62%	40%	38%	60%	
CALDAS	38%	38%	42%	46%	30%	22%	46%	
CAQUETÁ	68%	70%	74%	78%	50%	60%	78%	
CAUCA	68%	66%	70%	72%	52%	38%	72%	
CHOCÓ	54%	64%	70%	76%	36%	64%	68%	
CUNDINAMARCA	62%	60%	66%	68%	44%	54%	66%	
NARIÑO	62%	64%	68%	70%	36%	34%	70%	
PUTUMAYO	76%	80%	82%	84%	62%	74%	82%	
QUINDIO	44%	52%	56%	62%	48%	40%	62%	56%
RISARALDA	34%	40%	42%	46%	32%	22%	46%	
TOLIMA	60%	58%	62%	64%	40%	46%	64%	

Tabla 8: Resultados del RMEDSE (en millones de pesos) por Departamento y metodología aplicada en el proceso realizado.

DEPARTAMENTO	MODELO LINEAL GENERALIZA DO CON BASES DE POLINOMIOS	REGRESIÓ N LASSO CON BASES DE POLINOMI OS	REGRESIÓ N RIDGE CON BASES DE POLINOMI OS	K VECINOS MÁS CERCANOS CON BASES DE POLINOMI OS	MODELO LINEAL GENERALIZA DO CON BASES B- SPLINES	REGRESIÓ N LASSO CON BASES B- SPLINES	REGRESIÓ N RIDGE CON BASES B- SPLINES	K VECINOS MÁS CERCANO S CON BASES B- SPLINES
BOYACÁ	18.52	18.45	18.70		18.54	18.51	18.77	

CALDAS	13.09	12.89	13.01		12.97	12.83	12.98	
CAQUETÁ		1.68	1.72			1.76	1.73	
CAUCA	13.35	12.71	12.83		13.23	12.60	12.95	
CHOCÓ		7.54	10.11			6.89	10.10	
CUNDINAMARCA		17.75	17.42			17.68	17.48	
NARIÑO	14.45	14.44	13.95		14.62	14.54	13.96	
PUTUMAYO		25.80	26.37			26.03	26.39	
QUINDIO		34.16	35.06	34.76		34.15	35.31	35.96
RISARALDA		20.99	21.51			21.06	20.85	
TOLIMA	5.75	6.00	5.93		5.81	5.97	5.90	

Tabla 9: Resultados del RMEDSE (en millones de pesos) por Departamento y metodología aplicada en el proceso realizado.

DEPARTAMENTO	PERCEPTOR MULTICAPA	ARBOL DE DECISION	ARBOLES DE DECISION AGREGADOS	BOSQUE ALEATORIO	MAQUINA DE SOPORTE VECTORIAL CON FUNCIONES RADIALES	XGBoost	REGRESIÓN MULTIVARIADA ADAPTATIVA SPLINE	K VECINOS MÁS CERCANOS
BOYACÁ	17.45	18.40	17.60	17.36	20.77	26.94	17.67	
CALDAS	13.57	13.92	12.30	12.65	11.47	17.29	12.90	
CAQUETÁ	1.71	2.17	1.40	1.38	1.43	2.27	1.68	
CAUCA	12.01	13.82	10.71	11.29	13.16	19.16	13.18	
CHOCÓ	4.60	7.94	5.00	3.02	19.17	7.29	6.62	
CUNDINAMARCA	17.23	18.55	16.93	17.17	16.45	22.95	17.63	
NARIÑO	14.33	14.74	12.61	12.63	19.01	26.36	13.28	

PUTUMAYO	16.32	13.31	10.54	10.59	19.00	13.00	25.73	
QUINDIO	37.30	37.94	33.55	32.16	35.21	39.95	35.37	35.76
RISARALDA	22.65	22.64	20.73	21.16	20.66	27.97	20.87	
TOLIMA	6.52	5.87	5.68	5.46	5.61	8.29	6.00	

Es de resaltar que para cada ejecución se extrajeron resultados de métricas, predicciones e importancia de las variables, con el objetivo de apalancar el siguiente paso que se basa en la selección de componentes y ajuste del modelo final. En la [Sección 7.1.2](#) se detalla este proceso.

7.1.2 Selección de variables

Con base en el modelo seleccionado, se procedió a realizar la selección de variables. Para ello, se utilizó el método de selección de variables incluido en el paquete vip (Variable Importance in Projection) (Greenwell and Boehmke 2020). Este método permite identificar las variables más importantes en el modelo, lo que facilita la interpretación de los resultados. Estas medidas de importancia están sujetas a cada una de las metodologías, puesto que cada una de ellas tiene un enfoque diferente en la selección de variables.

De este modo y a partir de la ejecución de modelos elaborada con base en diferentes metodologías, espacios de hiperparámetros y demás, se generó un consolidado de las variables más importantes en un sentido recurrente. Esto quiere decir que para cada departamento se tienen las variables que de forma repetitiva fueron relevantes para las diferentes metodologías. Esto dio base para el proceso de selección final de variables explicativas. Con cada una de ellas se realizaron gráficos exploratorios que permitieran evidenciar la relación entre ellas y la variable respuesta. En los casos donde se observaba una relación evidente con sentido temático y económico, se procedió a incluir la variable en el modelo final.

7.1.3 Generación del modelo final

Una vez las metodologías y el conjunto de variables explicativas fueron seleccionadas, se procedió a la generación del modelo definitivo. Para ello, se hizo una última ejecución con una grilla de hiperparámetros más amplia que la considerada anteriormente, con el objetivo de maximizar/minimizar las métricas definidas. Con base en el modelo final, se genera la predicción de $S(\mathbf{s})$, donde \mathbf{s} puede representar una locación específica o un predio particular. Con base en este ajuste, se procede a realizar el cálculo de $\varepsilon(\mathbf{s})$ para cada uno de los registros en la muestra y con estos valores calculados en la muestra de entrenamiento se procede a realizar la validación de presencia o ausencia de correlación espacial. Una vez se llevan a cabo estos ajustes, para una locación o predio no observado en la muestra, se realiza la predicción tal y como se muestra en la

[Ecuación 17](#), donde $\hat{S}(s)$ proviene de la regresión ajustada y $\hat{\varepsilon}(s)$ del predictor generado a partir de kriging ordinario ajustado con base en el semivariograma empírico y teórico determinado.

$$\hat{Z}(s) = \hat{S}(s) + \hat{\varepsilon}(s) \quad (17)$$

7.1.4 Evaluación del modelo

El rendimiento del modelo se evaluó utilizando dos métricas clave: Un pseudo R^2 y el error cuadrático mediano (RMEDSE), descritos en la [Ecuación 12](#) y la [Ecuación 11](#), respectivamente. En adelante llamaremos R^2 al pseudo R^2 . El R^2 mide qué tan bien se ajusta el modelo a los datos observados en términos de un porcentaje entre 0 y 100%, mientras que el *RMEDSE* calcula un error en la misma escala de la variable respuesta. La [Figura 12](#) y la [Figura 13](#) muestran los resultados en términos de estas métricas para cada uno de los departamentos.

Por lo anterior observamos que valores más altos de R^2 indican un mayor ajuste del modelo a los datos observados, es decir, que el modelo es capaz de capturar una mayor parte de la variación en el valor del terreno. En el gráfico correspondiente [Figura 12](#), se observa que departamentos como Casanare, Norte de Santander y Putumayo presentan valores de R^2 , de 85.5%, 82.1% y 81.9% respectivamente lo que indican un excelente desempeño en términos de ajuste del modelo en estos departamentos. Estos resultados sugieren que los modelos utilizados en estas regiones explican de manera eficiente la variabilidad del valor del terreno.

Desde otra perspectiva el *RMEDSE*, que se presenta en [Figura 13](#), es una métrica que mide la mediana de las diferencias al cuadrado entre los valores observados y los valores predichos. Un *RMEDSE* bajo indica que el modelo genera predicciones más cercanas a los valores reales, lo que refleja una mayor precisión. En este caso, departamentos como Arauca, Caquetá y Casanare con errores de \$1.568.452 \$1.527.495 y \$2.465.003 respectivamente, muestran menores errores de predicción, lo que evidencia una alta precisión en las estimaciones realizadas por el modelo en estos departamentos.

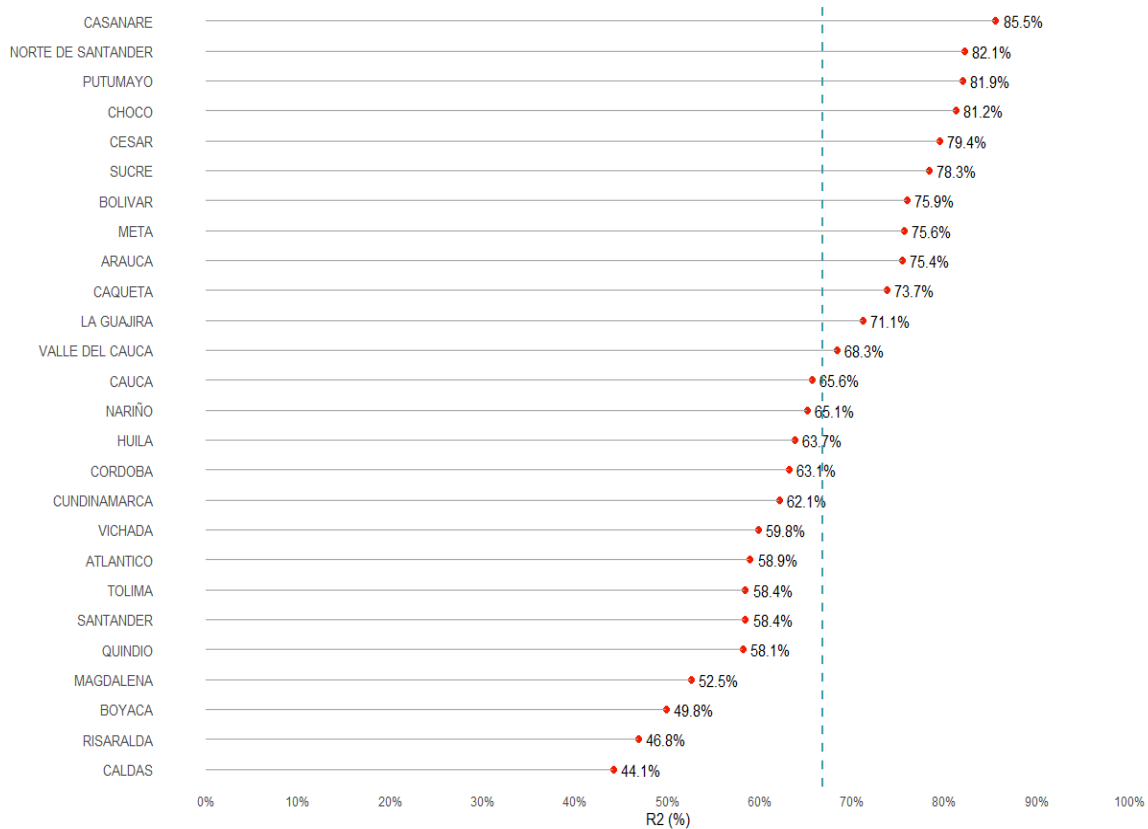


Figura 12: Pseudo R2 para los modelos finales por Departamento.

La línea punteada presente en ambos gráficos representa el valor promedio de cada métrica: 66.7% para el R^2 y \$10.759.244 para el RMEDS. Estos valores promedio sirven como referencia para evaluar el desempeño relativo de los modelos en cada departamento. Los departamentos cuyos valores de R^2 superan el promedio, como Casanare y Norte de Santander, reflejan un ajuste superior al promedio, mientras que aquellos con RMEDSE inferiores al promedio, como Arauca y Caquetá, muestran una precisión mayor en sus predicciones.

Es importante destacar que para la determinación del modelo definitivo se debe hacer una lectura conjunta para obtener una visión completa del rendimiento del modelo. Por ejemplo, Casanare presenta tanto un R^2 alto (85.5%) como un RMEDSE bajo (\$2.465.003), lo que refleja tanto un ajuste general adecuado del modelo como una precisión destacada en las predicciones. En cambio, departamentos como Risaralda (44.1% en R^2 y \$20.729.478 en RMEDS) ofrecen una lectura para mejorar el modelo, ya que presentan un ajuste moderado y un margen de mejora en la precisión de las predicciones. Estos resultados sugieren que, si bien el modelo es funcional en términos generales, su capacidad para hacer predicciones precisas en algunas áreas podría no serlo.

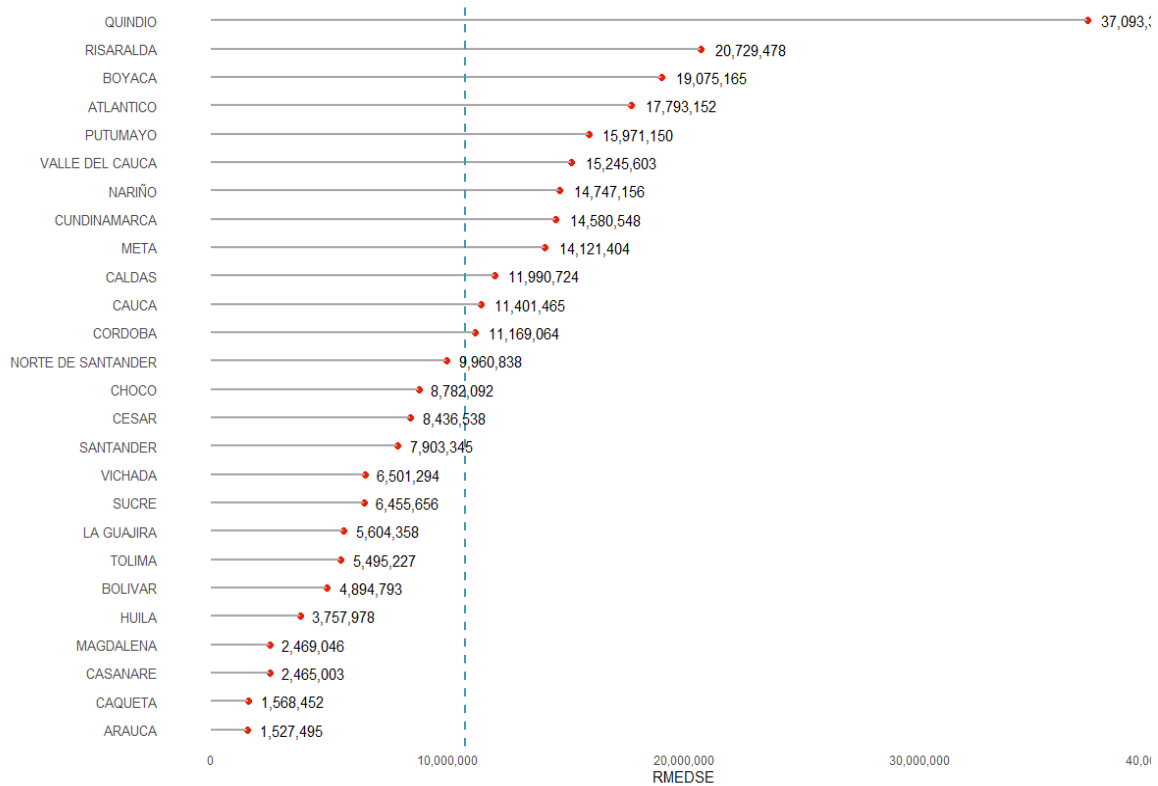


Figura 13: RMEDSE para los modelos finales por Departamento.

7.2 Implementación y/o difusión de los resultados

En esta sección se aborda el cálculo de valores sugeridos a nivel de ZHG a partir de las predicciones generadas por el modelo. Adicionalmente, se detalla el contenido incluido en el entregable generado para los modelos de cada departamento.

7.2.1 Cálculo del Valor de la Zona Geoeconómica (ZHG).

Partiendo de las definiciones dadas en la sección 6.3.2, ilustraremos por medio de un ejemplo los valores de la ZHG que se obtienen por cada una. Como ejemplo, se hará uso de una ZHG en particular la cual está compuesta por 8 polígonos con el valor y el área de terreno en hectáreas.

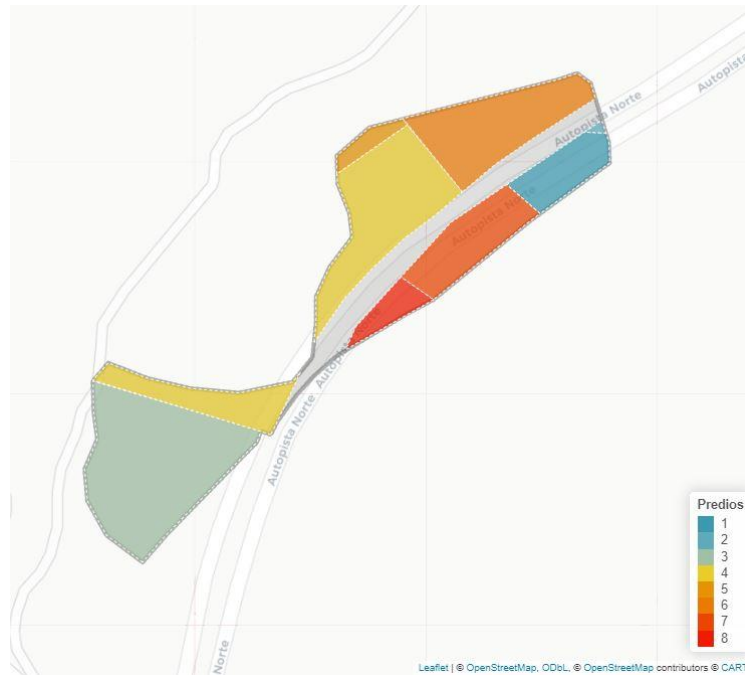


Figura 14: Zona Geoeconómica 24772-11-1, Suesca, Cundinamarca

La [Figura 14](#) es una zona homogénea ubicada en Suesca Cundinamarca compuesta por los predios que se presentan en la [Tabla 10](#). En esta tabla se presentan los valores de terreno por hectárea y el área en hectáreas de cada uno de los predios que componen la ZHG.

Tabla 10: Predios en la ZHG: 25772-11-1, con valores y áreas de terreno por hectárea

Zona Geoeconómica	Id	Área (HA)	Predicción
25772-11-1	1	0.17	\$51.502.439
25772-11-1	2	0.01	\$43.111.671
25772-11-1	3	0.77	\$19.328.707
25772-11-1	4	0.73	\$19.428.668
25772-11-1	5	0.05	\$79.957.753
25772-11-1	6	0.42	\$34.434.567
25772-11-1	7	0.27	\$24.401.699

25772-11-1	8	0.09	\$39.701.966
------------	---	------	--------------

Es así, que de acuerdo con las definiciones de las [Ecuaciones 13 a 16](#) se llega a los valores de ZHG de la [Tabla 11](#).

Tabla 11: Valor sugerido para la ZHG calculado por la media, mediana, media y mediana ponderadas.

Media	Mediana	Media Ponderada	Mediana Ponderada
\$38.983.434	\$37.068.266	\$26.752.598	\$20.620.306

7.3 Evaluación y/o presentación de los resultados

Esta sección tiene como propósito describir el reporte que contiene los resultados obtenidos de los modelos implementados. Asimismo, se detalla la configuración definitiva, desde una perspectiva técnica, para cada uno de los departamentos. El reporte se ha elaborado en formato Excel, estructurado en 10 hojas. El enlace para acceder a la ubicación de cada uno de los reportes está disponible en la sección de anexos. Cada uno de los reportes se componen de las siguientes hojas:

- **Diccionario de las variables usadas en el modelo para el departamento**

Esta hoja del reporte tiene como propósito presentar las variables explicativas incluidas en el modelo final, el tipo de variable, su definición y el rol que cumple.

- **Distribución de los valores comerciales por rangos de valor para el departamento**

En esta hoja se presenta un resumen de la distribución de la cantidad de registros en la base de modelación según los diferentes rangos de valor de terreno por hectárea (HA) para el departamento en cuestión. En esta se incluye una tabla que detalla los rangos de valor, la cantidad de registros por rango y su participación respecto al total. Además de mostrarse de forma tabular, se presenta un mapa del departamento con la distribución espacial de la información.

- **Distribución de los valores comerciales por municipio**

En esta hoja se presenta un resumen de la distribución de la cantidad de registros en la base de modelación según los diferentes rangos de valor por hectárea (HA) desagregando por municipios. Como en la anterior, en esta se incluye una tabla que detalla la cantidad de registros por municipio y su participación respecto al total.

- **Configuración y especificaciones del modelo**

Esta hoja tiene por objetivo mostrar el modelo que se configuró y las métricas de rendimiento. En esta se encuentra el R^2 y el $RMEDSE$ obtenido a partir de la comparación del $\hat{Z}(s)$, que se presenta en la [Ecuación 17](#) y el $Z(s)$ recolectado. Las métricas se presentan tanto para la muestra completa como para la muestra de entrenamiento que es la utilizada para evaluar el rendimiento del modelo.

- **Adecuación y transformaciones de las variables del modelo**

En esta hoja se detalla la tabla que describe las variables utilizadas en el modelo, junto con su rol y las transformaciones aplicadas. La columna “VARIABLE” enumera las distintas variables que forman parte del modelo, la columna “ROLE” especifica la función de cada variable en el modelo y la columna “TRANSFORMACION” describe las modificaciones realizadas a las variables. La lectura de la tabla se debe hacer de la siguiente manera, a modo de ejemplo se tomaron las variables y transformaciones para el departamento del César. La variable objetivo, identificada como “outcome”, **VALOR_TERRENO_COM_HECTAREA**, se sometió a una transformación logarítmica. A las variables predictoras, como **CODIGO_MUN**, **UPS_USO_PRINCI**, **MP_GRADO_IMPORTANCIA**, y **MP_CATEGORIAL_RURALIDAD**, se transformaron a variables dummy y a otras variables, como **AREA_TERRENO_M2** y **MP_VALOR_AGREGADO**, se les aplicó una transformación logarítmica y un ajuste utilizando base splines de grado 5.

- **Distribución de las variaciones en la muestra de prueba entre la predicción y valor observado**

En esta hoja se presenta una comparación detallada de las variaciones entre el valor observado y el predicho a nivel de predio.

La tabla en la parte superior ofrece un resumen de las variaciones porcentuales en los valores de predicción con respecto a los valores observados usando los percentiles de la distribución para medir las diferencias en la muestra. El gráfico de dispersión y el boxplot proporcionan una comparación visual entre las distribuciones de los valores observados (en rojo) y las predicciones (en azul).

- **Importancia de las variables**

En esta hoja se presenta la relevancia de las variables utilizadas en el modelo. La tabla describe la importancia de cada variable en una escala de 1 a 100. Al tratarse de un modelo de regresión Lasso, la importancia se calcula a partir del score derivado de la estadística t (ver Greenwell and Boehmke 2020). De esta manera, se cuantifica la importancia de cada variable en el modelo. La tabla se complementa con un gráfico de barras que visualiza la importancia de las variables, facilitando la interpretación de los resultados.

- **Efectos marginales entre las covariables y la predicción**

En esta se hoja se detalla el efecto individual o marginal de las covariables del modelo respecto a la predicción. Para las variables que son categóricas se describe ese efecto por medio un boxplot, en

cuanto a las variables continuas el efecto individual se presenta por medio de un gráfico de dispersión de puntos o de tendencia suavizado.

- **Ajuste de la predicción por rangos de valor por hectárea**

Esta hoja tiene como propósito mostrar el ajuste del modelo en función de distintos rangos de valor por hectárea expresados en millones. La visualización está dividida en múltiples gráficos, cada uno representando un rango específico de valores, ubicado en el eje horizontal (valor real). El eje vertical indica el valor predicho por el modelo.

- **Predicción sobre el marco del Departamento**

En esta hoja se presenta un resumen de la distribución de la predicción del modelo según los diferentes rangos de valor por hectárea (HA) para el departamento. Se presenta un mapa con la distribución espacial de la información por departamento. Los puntos en el mapa permiten visualizar la concentración transaccional por rango de valor.

8) Conclusiones

En este documento se realiza la descripción del proceso realizado para predecir el valor del terreno por hectárea para los predios rurales en municipios bajo gestión del IGAC. Para llevar a cabo este objetivo se hizo uso de registros con información económica proveniente del OIC del IGAC, tales como avalúos comerciales, ofertas y transacciones de SNR. A partir de estos datos, se generaron modelos de regresión que permitieron predecir el valor del terreno por hectárea en cada uno de los departamentos. A continuación, se resaltan los aspectos más relevantes del trabajo realizado:

- **Consolidación de información disponible:** Para este proyecto se utilizó únicamente la información disponible en el IGAC. De esta manera, no fue requerido hacer visitas a campo, lo cual puede ser desgastante en términos de tiempo y económicos. Para el procesamiento de esta información se generaron códigos que permiten hacer la consolidación y georreferenciación de forma masiva y automática. De esta forma, se propone la generación de una base con actualización periódica que suministre datos valiosos tanto a este proyecto como a procesos de actualización catastral o a otros proyectos requeridos en la entidad.
- **Relevancia de las variables:** Las variables relacionadas con la ubicación geográfica y las características físicas del terreno resultaron ser los principales determinantes en la predicción del valor de terreno por hectárea, destacando la importancia de estos factores en el proceso de modelación.
- **Desempeño del modelo:** los modelos demostraron un ajuste satisfactorio en función de su pseudo R^2 para la mayoría de los departamentos. De los 26 modelos, 15 lograron superar el 60% en esta métrica, mientras que 6 se ubicaron entre el 50% y el 60%, lo que refleja un

buen ajuste general. En cuanto al RMEDSE, 20 de los 26 modelos presentaron un error inferior a 15 millones de pesos, y 14 de ellos lograron un error menor a 10 millones de pesos, evidenciando una precisión adecuada en las predicciones realizadas.

- **Impacto de la metodología:** la metodología adoptada no se limitó a un único enfoque para modelar la variable de valor de terreno, sino que incorporó una combinación de técnicas avanzadas de regresión. Entre estas se incluyeron regresión Lasso, árboles de decisión, bosques aleatorios, máquinas de soporte vectorial y modelos basados en métodos de ensamble, utilizando la paquetería tidymodels en R. Esta estrategia facilitó la ejecución masiva de diversos modelos, permitiendo explorar metodologías ajustadas a las características específicas de cada departamento. Además, se implementaron diversas transformaciones y ajustes en las variables, destacando las transformaciones logarítmicas para variables continuas y el uso de bs-splines para capturar relaciones no lineales en métodos de regresión tradicionales.
- **Inclusión del término espacialmente dependiente:** adicionalmente al componente de regresión generado, como se ilustra en la Ecuación 17, se incorporó un término derivado del kriging ordinario para capturar la correlación espacial presente en los datos. Este término permitió ajustar el modelo a la variabilidad y dependencia espacial, lo que resultó en una mejora significativa en la precisión de las predicciones.
- **Determinación del valor de la ZHG:** al ofrecer alternativas para la determinación del valor del terreno de la ZHG a partir de las predicciones a nivel de predio, el uso de la media, mediana, media y mediana ponderadas proporciona una evaluación a los valores estimados para cada zona. Estas medidas brindan al equipo evaluador un conjunto de opciones para comparar y analizar los datos obtenidos, permitiendo así la selección de la medida más adecuada según las características específicas de la ZHG.
- **Automatización del proceso:** la metodología propuesta permite la automatización de la generación de los modelos y la predicción del valor de terreno por hectárea para los predios rurales en los municipios bajo gestión del IGAC. Esto facilita la actualización periódica de los modelos y la generación de los valores de la ZHG, lo que permite una mayor eficiencia en la toma de decisiones y la planificación de políticas públicas.

9) Bibliografía

Bonaccorso, Giuseppe. 2018. *Machine Learning Algorithms: Popular Algorithms for Data Science and Machine Learning*. Packt Publishing Ltd.

Bowman, Adrian W, and Adelchi Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with s-Plus Illustrations*. Vol. 18. OUP Oxford.

Brownlee, Jason. 2016. *Machine Learning Algorithms from Scratch with Python*. Machine Learning Mastery.

Carranza, Juan Pablo, Mario Andrés Piumetto, Carlos María Lucca, and Everton Da Silva. 2022. "Mass Appraisal as Affordable Public Policy: Open Data and Machine Learning for Mapping Urban Land Values." *Land Use Policy* 119: 106211.

Catastro Distrital UAEC, Unidad Administrativa Especial de. 2023a. "ENTREGABLE: DOCUMENTO METODOLÓGICO DE MODELOS ECONÓMICOS DISTRITO DE CARTAGENA."
https://www.catastrobogota.gov.co/sites/default/files/archivos/normas/20230106_MODELOS_ECONOMICOS_CARTAGENA.pdf.

———. 2023b. "RESULTADOS - CENSO INMOBILIARIO CATASTRAL - 2023."
https://antiguportal.shd.gov.co/shd/sites/default/files/documentos/Presentacion_procesos_UAEC_CD.pdf.

Córdoba, Mariano, Juan Pablo Carranza, Mario Piumetto, Federico Monzani, and Mónica Balzarini. 2021. "A Spatially Based Quantile Regression Forest Model for Mapping Rural Land Values." *Journal of Environmental Management* 289: 112509.

Cressie, Noel. 2015. *Statistics for Spatial Data*. John Wiley & Sons.

Cressie, Noel, and Christopher K Wikle. 2011. *Statistics for Spatio-Temporal Data*. John Wiley & Sons.

Dobson, Annette J, and Adrian G Barnett. 2008. *An Introduction to Generalized Linear Models*. Chapman; Hall/CRC.

Gräler, Benedikt, Edzer Pebesma, and Gerard Heuvelink. 2016. "Spatio-Temporal Interpolation Using Gstat." *The R Journal* 8: 204–18. <https://journal.r-project.org/archive/2016/RJ-2016-014/index.html>.

Greenwell, Brandon M., and Bradley C. Boehmke. 2020. "Variable Importance Plots—an Introduction to the Vip Package." *The R Journal* 12 (1): 343–66. <https://doi.org/10.32614/RJ-2020-013>.

Ho, Winky KO, Bo-Sin Tang, and Siu Wai Wong. 2021. "Predicting Property Prices with Machine Learning Algorithms." *Journal of Property Research* 38 (1): 48–70.

Jafary, Peyman, Davood Shojaei, Abbas Rajabifard, and Tuan Ngo. 2024. "Automated Land Valuation Models: A Comparative Study of Four Machine Learning and Deep Learning Methods Based on a Comprehensive Range of Influential Factors." *Cities* 151: 105115.

Kontrimas, Vilius, and Antanas Verikas. 2011. "The Mass Appraisal of the Real Estate by Computational Intelligence." *Applied Soft Computing* 11 (1): 443–48.

Kuhn, Max, and Davis Vaughan. 2024. *Parsnip: A Common API to Modeling and Analysis Functions*. <https://CRAN.R-project.org/package=parsnip>.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles*. <https://www.tidymodels.org>.

Kuhn, Max, Hadley Wickham, and Emil Hvitfeldt. 2024. *Recipes: Preprocessing and Feature Engineering Steps for Modeling*. <https://CRAN.R-project.org/package=recipes>.

Melo, O, L López, and S Melo. 2007. "Diseño de Experimentos: Métodos y Aplicaciones." *Editorial Universidad Nacional de Colombia*. Bogotá.

Mohammed, Mohssen, Muhammad Badruddin Khan, and Eihab Bashier Mohammed Bashier. 2016. *Machine Learning: Algorithms and Applications*. Crc Press.

Montgomery, Douglas C. 2020. *Introduction to Statistical Quality Control*. John Wiley & Sons.

Montgomery, Douglas C, Elizabeth A Peck, and G Geoffrey Vining. 2012. *Introduction to Linear Regression Analysis*. Vol. 821. John Wiley; Sons, New York.

Nelder, John Ashworth, and Robert WM Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)* 135 (3): 370–84.

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With Applications in r*. Chapman; Hall/CRC.

Python Software Foundation. 2023. *Python 3.11: A Dynamic, Open Source Programming Language*. Python Software Foundation.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Ravishanker, Nalini, Zhiyi Chi, and Dipak K Dey. 2021. *A First Course in Linear Model Theory*. Chapman; Hall/CRC.

Schabenberger, Oliver, and Carol A Gotway. 2017. *Statistical Methods for Spatial Data Analysis*. Chapman; Hall/CRC.

Scrucca, Luca. 2004. "Qcc: An r Package for Quality Control Charting and Statistical Process Control." *R News* 4/1: 11–17. <https://cran.r-project.org/doc/Rnews/>.

Tolosa Delgado, Jurgen Daniel. 2020. "Modelación Conjunta de Media y Varianza En Modelos Semiparamétricos Autorregresivos Espaciales."

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Golemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with r*. Chapman; Hall/CRC.

10) Anexos

10.1 Anexos A

Se presenta en este anexo las rutas donde se encuentran los reportes generados para los modelos por departamento, así como también los valores de ZHG como resultado del proceso de modelación por cada una de las medidas de tendencia central.

[Reportes Modelos Art. 49](#)

10.2 Anexos B

En las figuras 15 a la 40 se presentan los resultados de las predicciones generadas para los predios rurales que conforman el departamento.

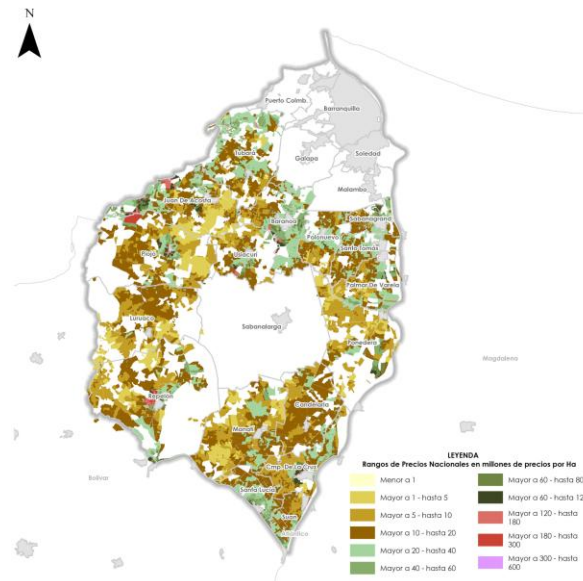


Figura 15: Predicciones de valor de terreno por hectárea generadas para el departamento del Atlántico

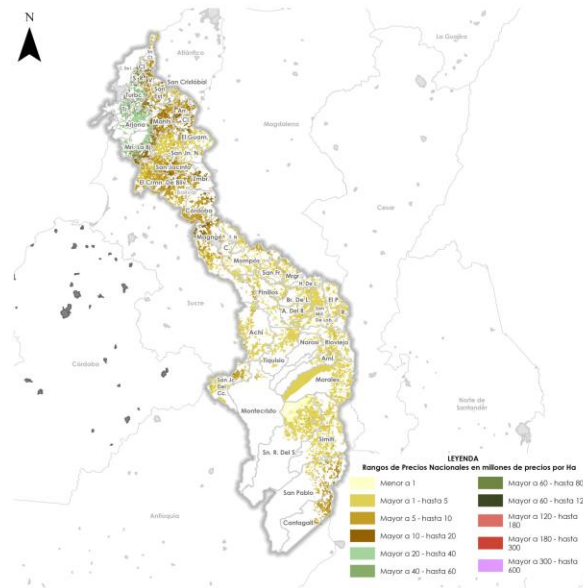


Figura 16: Predicciones de valor de terreno por hectárea generadas para el departamento de Bolívar

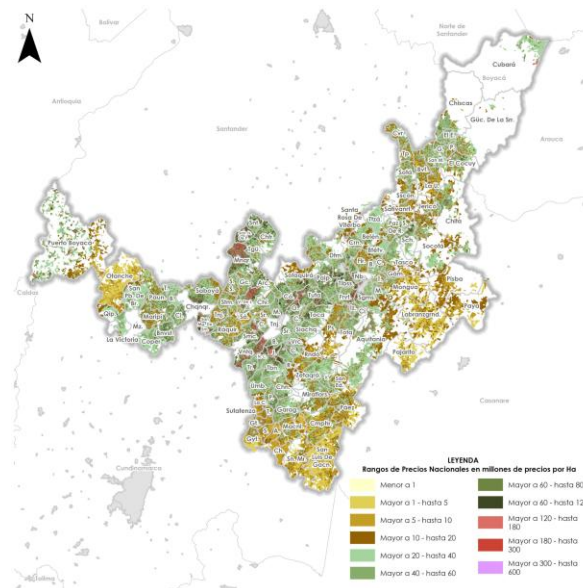


Figura 17: Predicciones de valor de terreno por hectárea generadas para el departamento de Boyacá

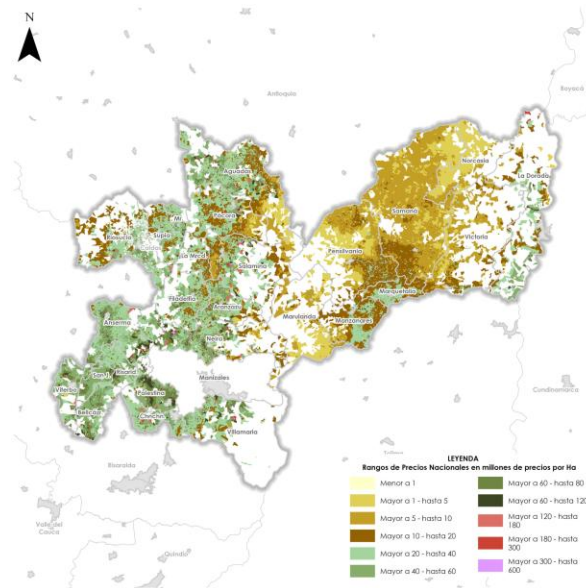


Figura 18: Predicciones de valor de terreno por hectárea generadas para el departamento de Caldas

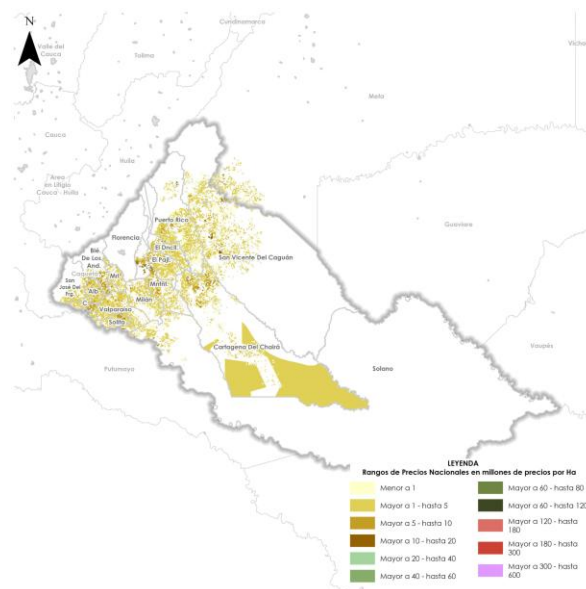


Figura 19: Predicciones de valor de terreno por hectárea generadas para el departamento de Caquetá

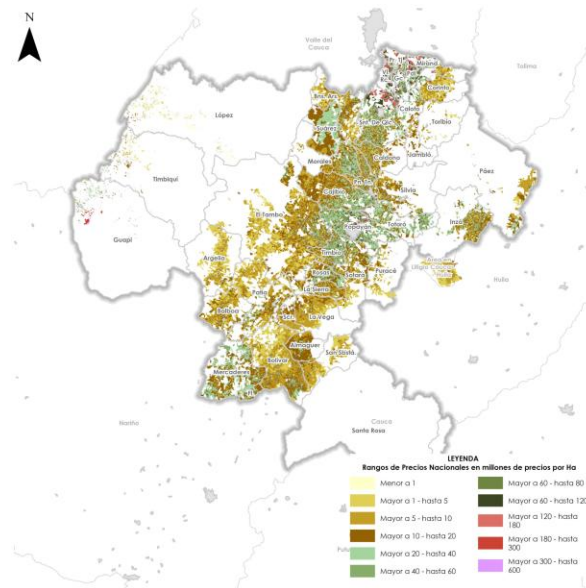


Figura 20: Predicciones de valor de terreno por hectárea generadas para el departamento de Cauca

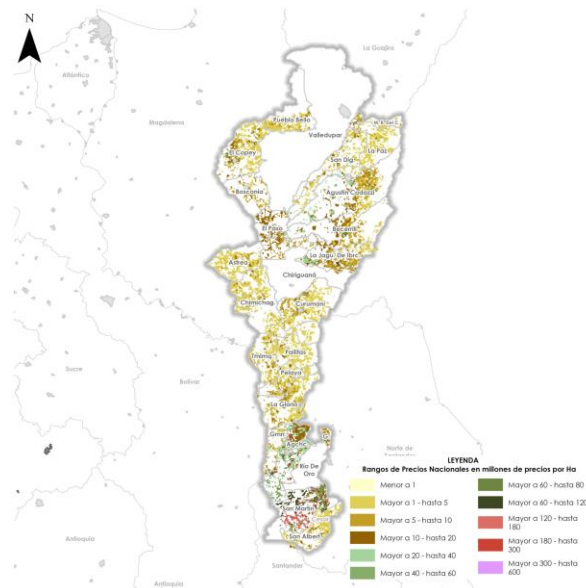


Figura 21: Predicciones de valor de terreno por hectárea generadas para el departamento de Cesar

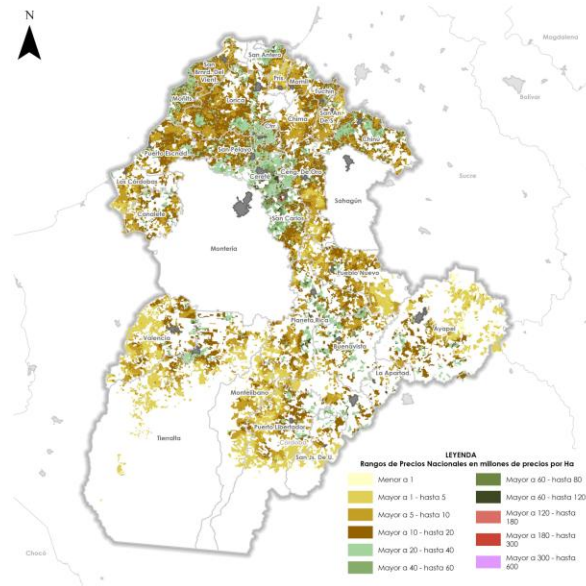


Figura 22: Predicciones de valor de terreno por hectárea generadas para el departamento de Córdoba

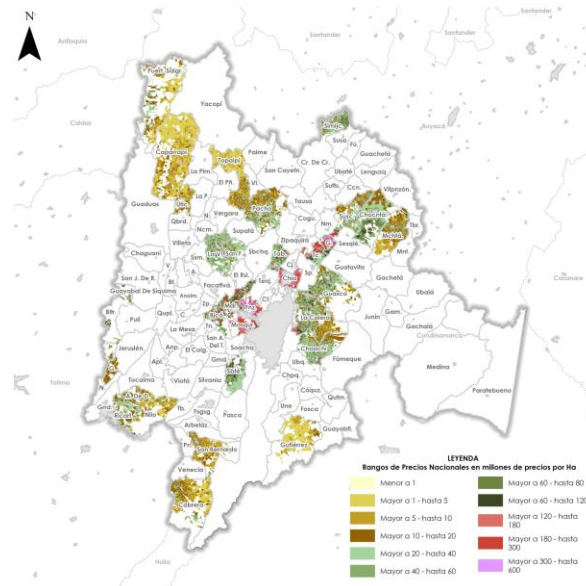


Figura 23: Predicciones de valor de terreno por hectárea generadas para el departamento de Cundinamarca

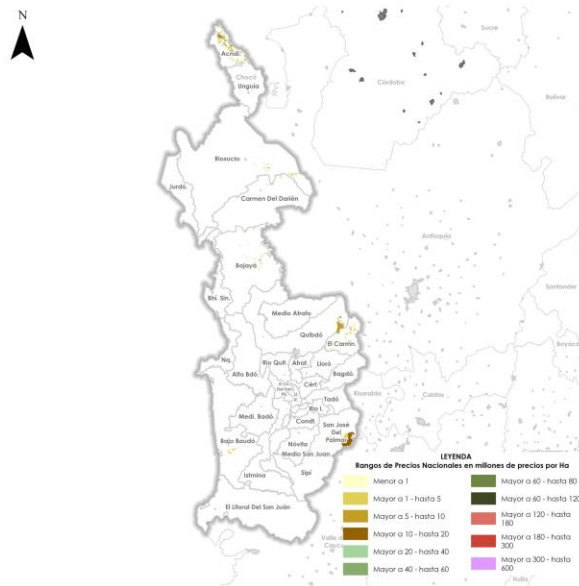


Figura 24: Predicciones de valor de terreno por hectárea generadas para el departamento del Chocó

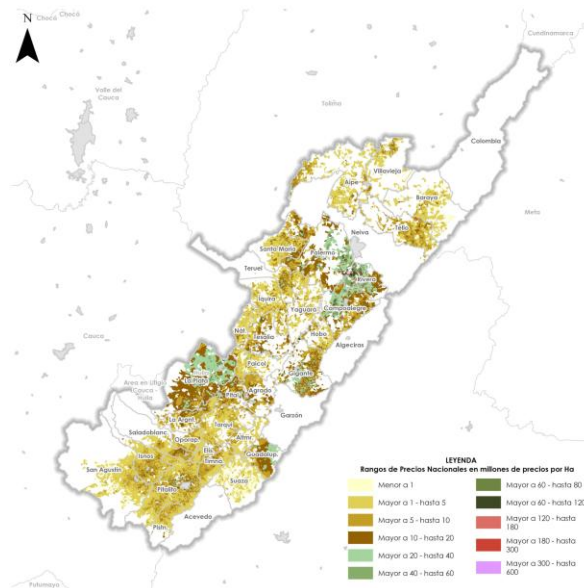


Figura 25: Predicciones de valor de terreno por hectárea generadas para el departamento del Huila

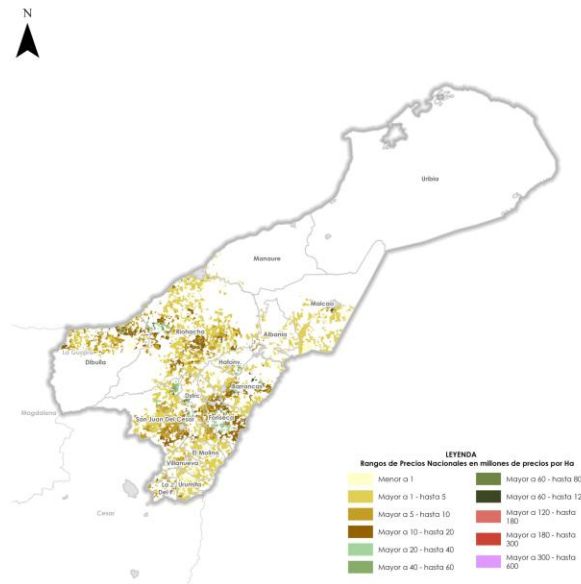


Figura 26: Predicciones de valor de terreno por hectárea generadas para el departamento de La Guajira

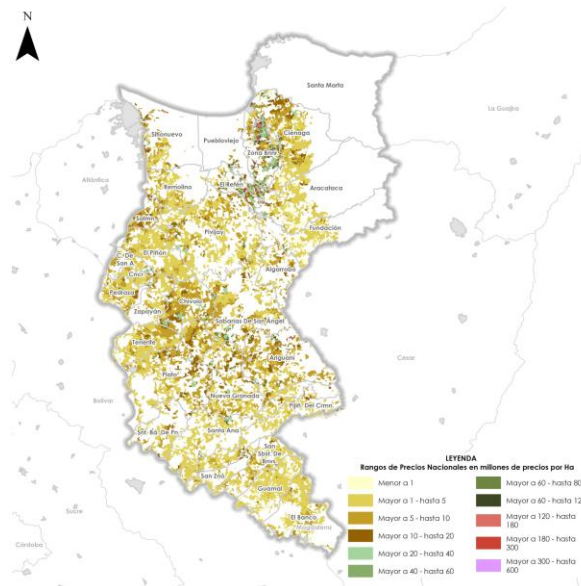


Figura 27: Predicciones de valor de terreno por hectárea generadas para el departamento del Magdalena

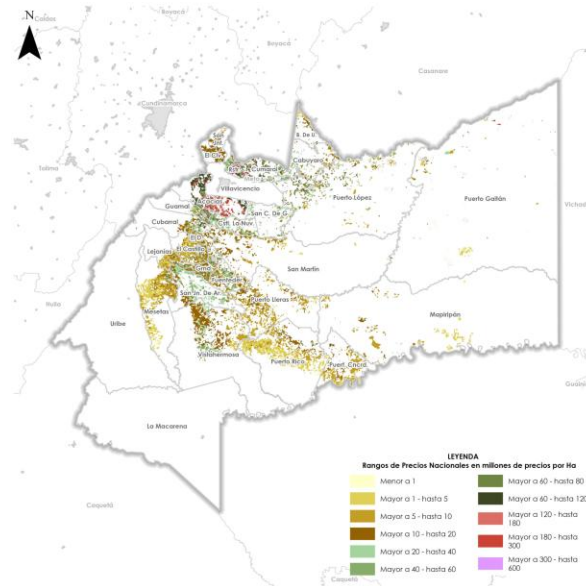


Figura 28: Predicciones de valor de terreno por hectárea generadas para el departamento del Meta

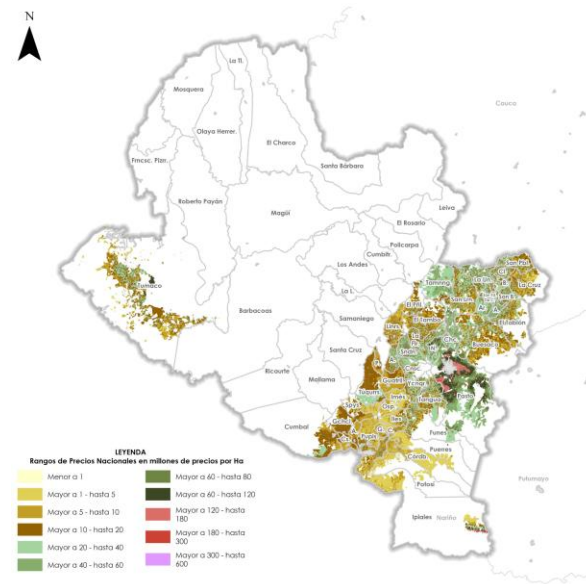


Figura 29: Predicciones de valor de terreno por hectárea generadas para el departamento de Nariño

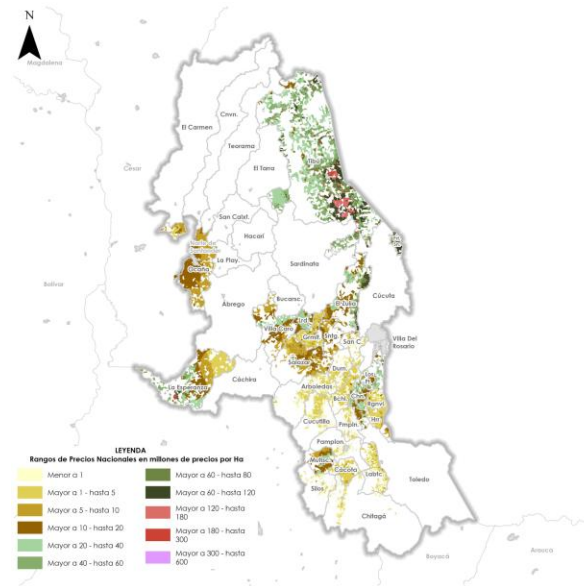


Figura 30: Predicciones de valor de terreno por hectárea generadas para el departamento de Norte de Santander

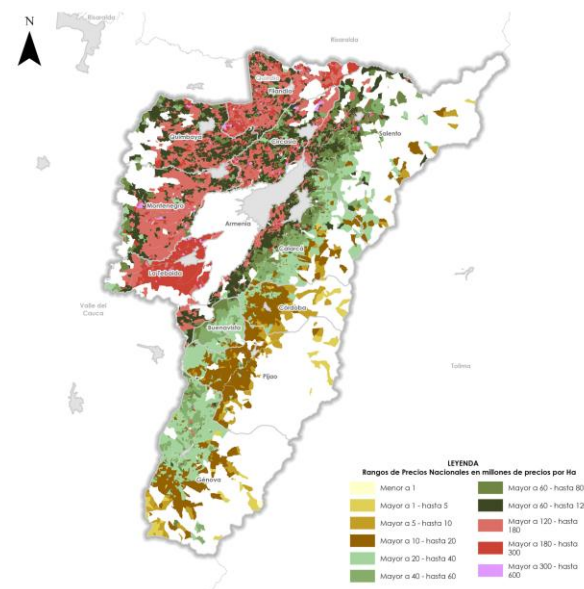


Figura 31: Predicciones de valor de terreno por hectárea generadas para el departamento del Quindío

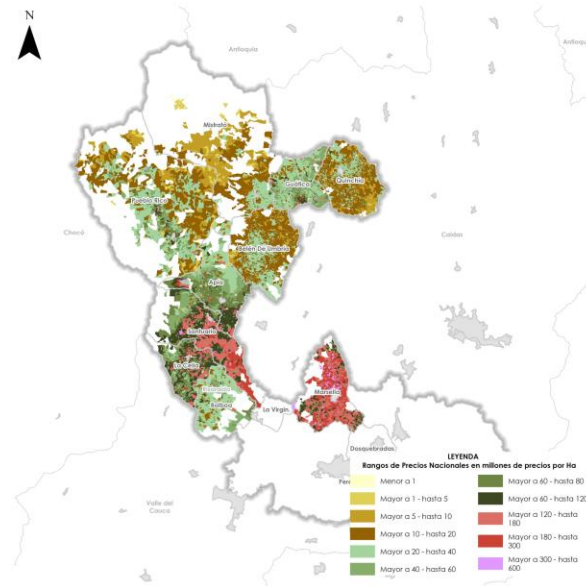


Figura 32: Predicciones de valor de terreno por hectárea generadas para el departamento de Risaralda

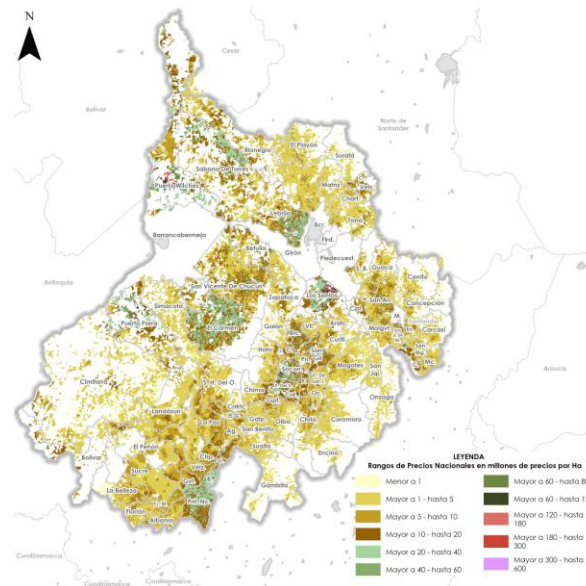


Figura 33: Predicciones de valor de terreno por hectárea generadas para el departamento de Santander

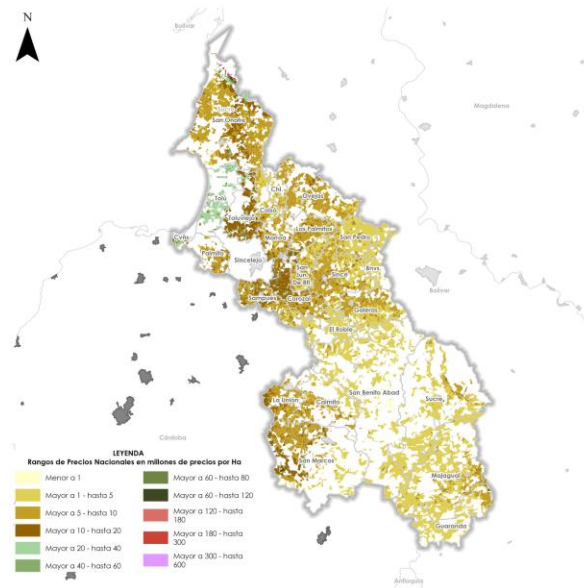


Figura 34: Predicciones de valor de terreno por hectárea generadas para el departamento de Sucre

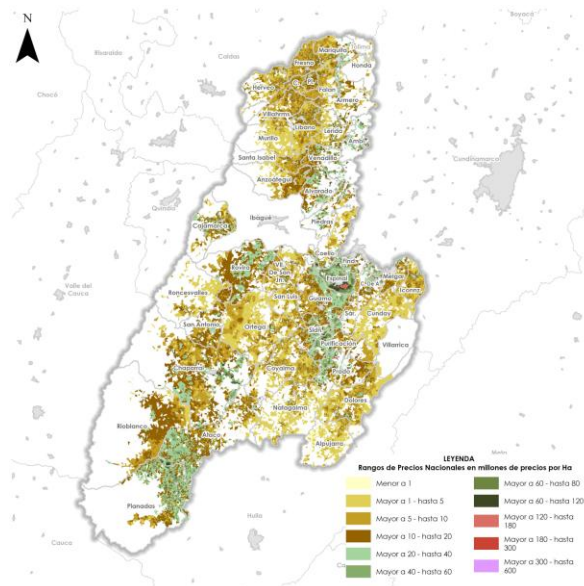


Figura 35: Predicciones de valor de terreno por hectárea generadas para el departamento del Tolima

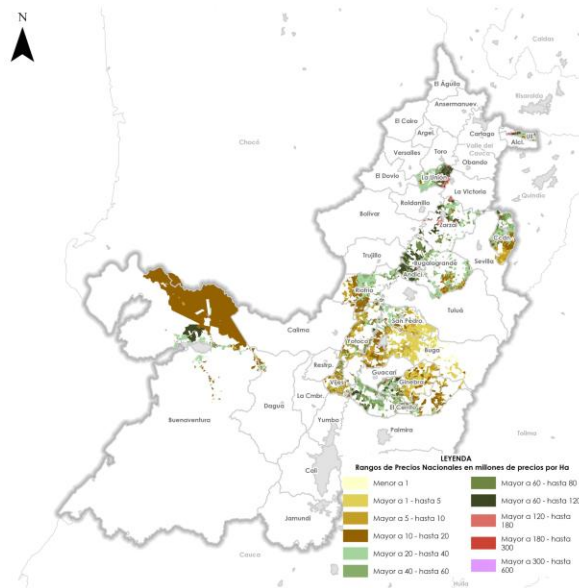


Figura 36: Predicciones de valor de terreno por hectárea generadas para el departamento de Valle del Cauca

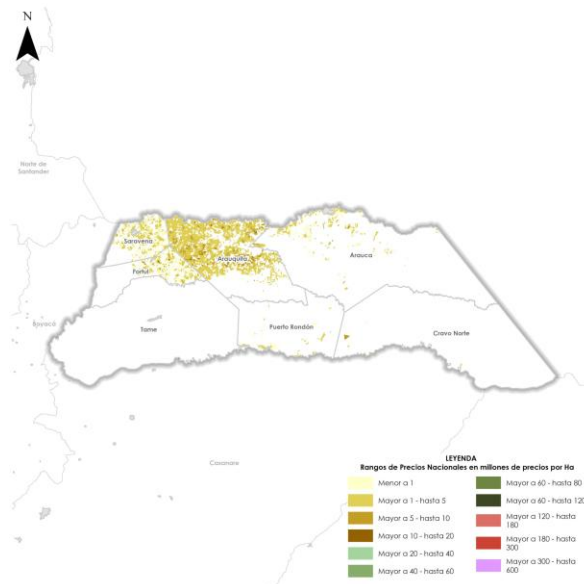


Figura 37: Predicciones de valor de terreno por hectárea generadas para el departamento de Arauca

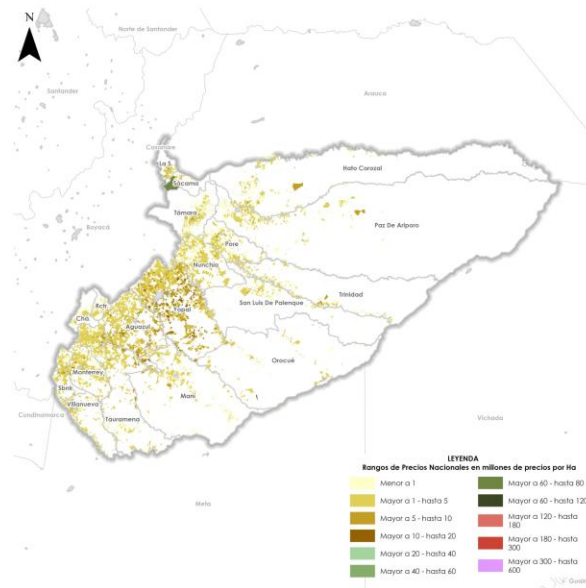


Figura 38: Predicciones de valor de terreno por hectárea generadas para el departamento de Casanare

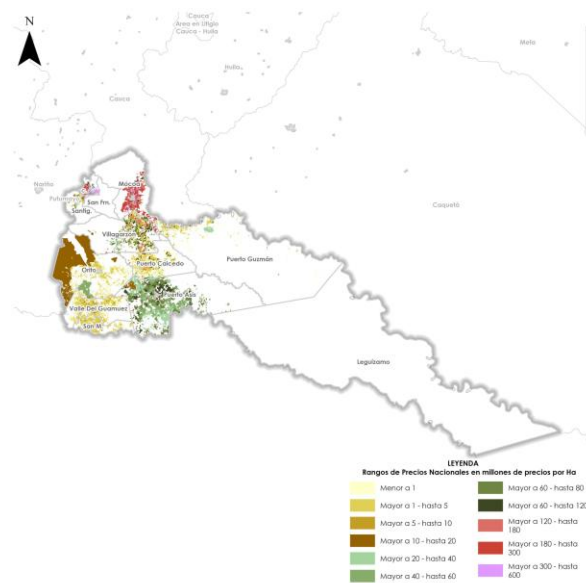


Figura 39: Predicciones de valor de terreno por hectárea generadas para el departamento del Putumayo

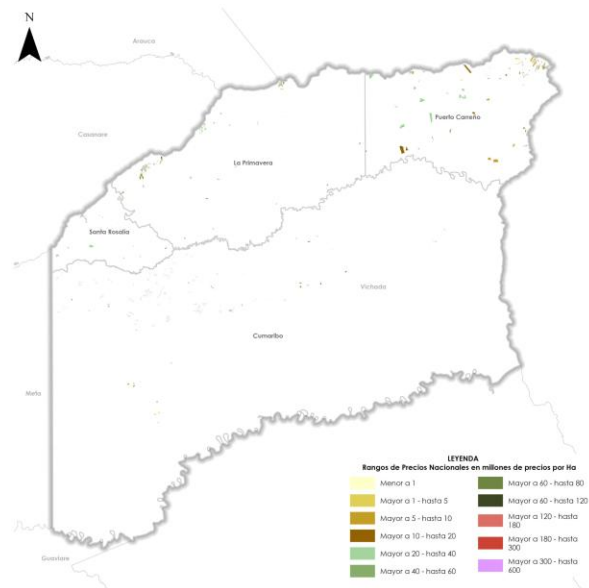


Figura 40: Predicciones de valor de terreno por hectárea generadas para el departamento del Vichada