Estimación del precio de renta en predios rurales mediante modelación espacial en Colombia

Carlos Andrés Salgado Ramírez
Código 00020438133

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

Director
David Arango Londoño PhD©

Codirectora
Martha Patricia Bohorquez Castañeda PhD

Facultad de Ingeniería y Ciencias
Maestría en Ciencia de Datos
Santiago de Cali, Abril 21 de 2024

Ficha Resumen

- **TÍTULO:** Estimación del precio de renta en predios rurales mediante modelación espacial en Colombia.

1. **ÁREA DE TRABAJO:** Sector gubernamental
2. **TIPO DE PROYECTO (Aplicado, Innovación, Investigación):** Aplicado
3. **ESTUDIANTE:** Carlos Andrés Salgado
4. **CORREO ELECTRÓNICO:** c.salgado@javerianacali.edu.co
5. **DIRECCIÓN Y TELÉFONO:** CALLE 10 A # 3S-02 | (+57) 315 5563061
6. **DIRECTOR:** David Arango Londoño
7. **VINCULACIÓN DEL DIRECTOR:** Docente Catedrático
8. **CORREO ELECTRÓNICO DEL DIRECTOR:** david.arango@javerianacali.edu.co
9. **CO-DIRECTORA:** Martha Patricia Bohorquez
10. **GRUPO O EMPRESA QUE LO AVALA:** Sociedad de activos especiales (SAE).
11. **PALABRAS CLAVE (al menos 5):** Ciencia de datos, Modelos espaciales, Estimación de Renta, Predios Rurales, Colombia, Machine Learning.
12. **FECHA DE INICIO:** 01-jun-23
13. **FECHA DE FINALIZACIÓN:** 10-oct-24
14. **RESUMEN:** La Sociedad de Activos Especiales (SAE), que administra predios en extinción de dominio, actualmente está en el proceso de otorgar sus predios a diversas instituciones y comunidades para contribuir tanto a la soberanía alimentaria, como a otros sectores sociales del país. Para lograr esto, se deben establecer precios de renta justos y transparentes. Los métodos existentes para estimar la renta en predios rurales no han sido implementados hasta ahora en el país. Este proyecto desarrolla un modelo basado en técnicas de aprendizaje supervisado usando machine learning para estimar la renta de predios rurales en Colombia. El modelo incorpora la dependencia espacial, lo que permite una comprensión más profunda de las variaciones en los precios de renta. Este proyecto representa una oportunidad de innovación para la estimación de la renta en predios rurales y contribuye significativamente a la seguridad y soberanía alimentaria del país.

Tabla de Contenido

Índice de figuras

Índice de tablas

1. Introducción

La Sociedad de Activos Especiales (SAE) desempeña una tarea crucial en la gestión de predios en extinción de dominio y su otorgamiento a diversas entidades y comunidades rurales. Este esfuerzo es fundamental para contribuir a la seguridad y a la soberanía alimentaria del país, un objetivo de gran relevancia social con profundas implicaciones que se enmarcan tanto en el acuerdo de paz firmado en La Habana como en el Objetivo 2: Hambre cero de los Objetivos de Desarrollo Sostenible de la ONU para 2030. Sin embargo, para lograr esto de manera efectiva, es esencial establecer precios de renta justos y sostenibles en el tiempo para los predios otorgados. Estos precios no deben representar una carga para las comunidades que desarrollarán proyectos productivos en favor de la seguridad y soberanía alimentaria del país.

Los modelos espaciales son una valiosa herramienta para abordar este desafío, ya que permiten incorporar una variedad de factores que pueden influir en los precios de renta, incluyendo la influencia diferenciada de estos por su localización geográfica. Esto, junto a un buen conjunto de covariables, permite una comprensión más profunda de la variable de respuesta (dependiente), en este caso, los precios de renta.

Un estudio de 2013 de J. C. Muñoz Mora y H. Cardona Jaramillo ofrece una visión detallada de las metodologías de valoración de predios rurales en Colombia, incluyendo los métodos tradicionales y alternativos de estimación de dicha valorización, considerando tanto características mercables como no mercables. Sin embargo, este estudio también destaca la falta de investigaciones que expliquen los mecanismos de generación de precios para los predios rurales en Colombia, lo que se presenta actualmente como un obstáculo para el desarrollo rural.

Ante este escenario, se desarrolla un modelo basado en técnicas de aprendizaje supervisado utilizando *machine learning*. Este modelo involucra la dependencia espacial y permite la incorporación de una amplia gama de covariables que pueden influir en los precios de renta. La elección del modelo más adecuado se realiza teniendo en cuenta su validación y la evaluación de sus métricas de rendimiento. Con este enfoque, se construye una herramienta que sirva para establecer precios de renta justos que no comprometan la rentabilidad de los proyectos productivos en los predios rurales otorgados (rentados) a las diferentes comunidades por la SAE en el país.

2. Definición del Problema

En esta sección, se desarrolla el planteamiento del problema. Se identifica el problema en términos concretos y explícitos, formulando las variables que lo constituyen y sobre las cuales se fundamenta la formulación de los objetivos e hipótesis de la investigación.

2.1. Planteamiento del Problema

La Sociedad de Activos Especiales (SAE) de Colombia, que administra predios en extinción de dominio, se enfrenta a un desafío crucial. Su objetivo es otorgar estos predios a diversas instituciones y comunidades para contribuir a la soberanía alimentaria del país. Sin embargo, establecer precios de renta justos y transparentes que sean sostenibles en el tiempo y que no representen una carga para las comunidades es un reto significativo.

Los métodos existentes para estimar la renta en predios rurales, aunque se han aplicado en otros contextos globales, no se han implementado específicamente en Colombia [1]. Estos métodos (modelos) suelen tener un enfoque económico y determinista, lo que limita la evaluación de covariables adicionales que podrían ser relevantes para la estimación de la renta, como la ubicación, el tamaño del terreno, la calidad del suelo, la proximidad a los mercados, la infraestructura, entre otras [1].

Esta situación ha dejado un vacío en la literatura científica y plantea un desafío en el contexto colombiano. La falta de un método adecuado para estimar la renta de predios rurales en Colombia podría llevar al cobro de precios de renta injustos o insostenibles. Esto podría comprometer el acuerdo Támesis que lidera el gobierno y que hace parte de la reforma agraria, el cual enmarca la ruta para el otorgamiento de predios a las comunidades por parte de la SAE. Además, podría poner en riesgo la viabilidad de los proyectos productivos en los predios otorgados.

Finalmente, este es un momento decisivo para el país. Tras el acuerdo de paz, el gobierno está enfocando sus esfuerzos en el sector rural como eje principal para resolver los diversos conflictos que aún persisten en ciertos territorios. Además, se está prestando atención al cambio climático mediante la implementación de regulaciones sólidas en todo lo relacionado con el desarrollo rural y el ordenamiento territorial. Por otro lado, este proyecto es de gran utilidad para abordar el desafío de garantizar una comprensión más profunda de las variaciones regionales en los precios de la renta rural, algo crucial para generar beneficios significativos tanto en la industria agraria, como para mejorar las condiciones de vida del campesinado y demás comunidades que sean acogidos por el acuerdo Támesis.

2.2. Formulación del Problema

El objetivo de esta investigación es responder a la pregunta general:

¿Cómo se puede estimar la renta de predios rurales en Colombia mediante un modelo basado en la ciencia de datos?

Para lograrlo, se plantean las siguientes preguntas:

- ¿Cómo se pueden caracterizar las variables relevantes para la estimación de la renta en predios rurales?
- ¿Cómo se puede desarrollar un modelo que incorpore el componente espacial para determinar la renta de predios rurales en Colombia?
- ¿Cómo se puede validar la precisión de las estimaciones de renta generadas por el modelo por medio de validación cruzada?

3. Objetivos del Proyecto

3.1. Objetivo General

Estimar la renta de predios rurales en Colombia mediante un modelo basado en técnicas de aprendizaje supervisado usando machine learning.

3.2. Objetivos Específicos

- Caracterizar las variables para la estimación de la renta en predios rurales.
- Desarrollar un modelo de aprendizaje supervisado con dependencia espacial (sar - spatial lag error) para la renta de predios rurales en Colombia.
- Validar la precisión de las estimaciones de renta generadas por el modelo mediante validación cruzada.

4. Marco Teórico y Antecedentes

4.1. Antecedentes

El artículo de 2013 de J. C. Muñoz Mora y H. Cardona Jaramillo, **El precio de la tierra: estado del arte de las metodologías de valoración de predios rurales y su aplicación en Colombia**, ofrece una visión detallada de las metodologías de valoración de predios rurales, incluyendo los métodos tradicionales de retornos esperados y alternativos que consideran características no mercables [1]. El artículo destaca la metodología de precios hedónicos, que integra diversas variables para asignar un valor a la tierra. Sin embargo, la revisión de la literatura para Colombia revela una falta de estudios que expliquen los mecanismos de generación de precios para los predios rurales en el país, lo que podría ser un obstáculo para el desarrollo rural [1]. Este trabajo es un antecedente útil para modelar la renta de predios rurales, ya que proporciona una comprensión de las metodologías existentes, aunque deterministas, y el potencial de la metodología de precios hedónicos para considerar una amplia gama de factores que pueden influir en la renta.

El estudio de 2019 realizado por Hyunwoo Lim y Minyoung Park, titulado **Modeling the Spatial Dimensions of Warehouse Rent Determinants: A Case Study of Seoul Metropolitan Area, South Korea**, se enfoca en la dimensión espacial de los determinantes del alquiler de almacenes [2]. Utilizando el Área Metropolitana de Seúl en Corea del Sur como caso de estudio, los autores desarrollan modelos de regresión autorregresiva espacial (SAR) y regresión ponderada geográficamente mixta (MGWR) para explicar las relaciones espaciales entre el alquiler de almacenes y las variables explicativas [2].

Los resultados del estudio ofrecen una comprensión profunda de la interacción entre las actividades logísticas, la infraestructura de transporte y el uso de la tierra, proporcionando información valiosa para el desarrollo sostenible de las instalaciones de logística urbana. Este estudio es particularmente útil para modelar la renta en predios rurales, ya que proporciona un marco para entender cómo los factores espaciales y las características de la propiedad pueden influir en los precios de alquiler. Además, las técnicas de modelado espacial utilizadas en el estudio podrían ser aplicables al modelado de la renta en predios rurales.

El artículo de 2021 de Andrius Grybauskas, Vaida Pilinkienė y Alina Stundžienė, **Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic**, explora qué atributos de un apartamento son más propensos a influir en una revisión de precio durante la pandemia de COVID-19 [3]. El estudio recopiló casi 19,000 listados de propiedades en Vilnius durante la primera ola de la pandemia y aplicó 15 modelos de aprendizaje automático para pronosticar las revisiones de apartamentos, utilizando los valores SHAP para la interpretabilidad [3].

Los hallazgos del estudio indican que el mercado inmobiliario es bastante resistente a las pandemias, con caídas de precios no tan dramáticas como se creía inicialmente [3]. De los 15 modelos probados, el impulso de gradiente extremo fue el más preciso, aunque la diferencia fue insignificante. Los valores SHAP concluyen que el tiempo en el mercado fue la variable más dominante para la predicción de la revisión de precios [3]. Este trabajo es útil para estimar la renta en predios rurales, ya que proporciona un marco para entender cómo los factores espaciales y las características de la propiedad pueden influir en los precios de alquiler.

El estudio de 2021 de Scott W. Hegerty, **Are Rents Excessive in the Central City? A Geospatial Analysis**, analiza la relación entre los valores de las propiedades y las rentas en las ciudades centrales de los Estados Unidos. El estudio calcula medidas a nivel de tracto del censo de la relación renta-valor de la propiedad (RPV) para 30 grandes ciudades y sus áreas metropolitanas circundantes. Utiliza los puntajes Z y los cuantiles para identificar valores RPV extremos en todo el país [4].

Se emplea una estimación de regresión de rezago espacial que muestra que, controlando por ingresos, valores de propiedad y tasas de vacancia, las características raciales a menudo tienen los signos opuestos a lo que se podría esperar y que hay poca evidencia de "explotación" de inquilinos basada puramente en la raza [4]. Este trabajo es útil para modelar la renta en predios rurales, ya que proporciona un marco para entender cómo los factores socioeconómicos y las características de la propiedad pueden influir en los precios de alquiler. Además, las técnicas de modelado espacial utilizadas en el estudio podrían ser aplicables al modelado de la renta en predios rurales.

4.2. Marco Teórico

4.2.1. Apartado Contextual

Sociedad de Activos Especiales (SAE) La Sociedad de Activos Especiales S.A.S. (SAE) es una sociedad por acciones simplificada de economía mixta. Está conformada con capital estatal y privado, es de orden nacional y está vinculada al Ministerio de Hacienda y Crédito Público [5]. La SAE tiene por objeto administrar bienes especiales que se encuentran en proceso de extinción o se les haya decretado extinción de dominio [5].

Seguridad Alimentaria La seguridad alimentaria existe cuando todas las personas tienen en todo momento acceso físico y económico a suficientes alimentos inocuos y nutritivos para satisfacer sus necesidades alimenticias y sus preferencias en cuanto a los alimentos a fin de llevar una vida activa y sana [6]. Este concepto se centra en garantizar que todas las personas tengan acceso regular a alimentos suficientes, seguros y nutritivos para llevar una vida activa y saludable. Los componentes clave de la seguridad

alimentaria incluyen la disponibilidad de alimentos, el acceso a estos, la utilización y la estabilidad [6].

Soberanía Alimentaria La soberanía alimentaria es un enfoque conceptual que se refiere al derecho de las comunidades a definir sus propias políticas de alimentos y agricultura [7]. Este concepto se centra en el poder sobre los sistemas alimentarios, en lugar de simplemente la presencia o ausencia de alimentos [7]. En otras palabras, se trata de quién controla los recursos y procesos de producción de alimentos, y cómo se distribuyen y consumen estos alimentos. En el contexto de Colombia, la soberanía alimentaria aún no se ha logrado plenamente, pero hay esfuerzos en curso para trabajar hacia este objetivo. Durante la emergencia sanitaria causada por COVID-19 en Colombia, se han generado estrategias para fortalecer la seguridad alimentaria y la soberanía alimentaria [8]. Entre los mecanismos adoptados por las comunidades urbanas y rurales y los gobiernos locales se encuentran formas alternativas de obtener alimentos, una de las cuales es el trueque, que permite el intercambio de alimentos producidos por campesinos entre territorios [8]. En Bogotá, se han fortalecido los mercados campesinos, que han permitido llevar alimentos producidos por campesinos en Tolima, Meta y Boyacá a los hogares urbanos, promoviendo circuitos cortos de comercialización y un pago justo a los productores [8].

Acuerdo de Támesis El Acuerdo de Támesis es un pacto firmado por el gobierno y la Sociedad de Activos Especiales (SAE) con el objetivo de ofrecer garantías para entregar predios de la mafia a poblaciones afectadas por el despojo y la violencia [9]. Este acuerdo se firmó en el marco de la entrega de las primeras 600 hectáreas de tierras a campesinos en Montería [9]. Se trata de la finca Támesis, que perteneció al exjefe paramilitar asesinado Carlos Castaño, y cuya propiedad pasó a ser de dominio del Estado colombiano [9].

Definición de Predio Predio, es una porción de tierra con límites definidos, que puede ser tanto urbana como rural [10]. En términos legales, un predio es una unidad de propiedad que puede ser poseída, comprada, vendida, arrendada, hipotecada, entre otros [10]. Un predio rural, por otro lado, es un predio ubicado en el campo o fuera de los límites de la ciudad cuyo destino en el ordenamiento municipal es al desempeño de una actividad económica o a su protección ambiental [11]. Se considera predio rural a aquella porción de tierra ubicada en área rural o en área de expansión urbana declarada zona intangible, dedicada a uso agrícola, pecuario o forestal [11].

Concepto de Renta La renta en términos económicos se refiere a los ingresos derivados de la propiedad de la tierra y otros regalos gratuitos de la naturaleza [12]. Este concepto fue elegido por razones técnicas por el economista neoclásico Alfred Marshall y otros después de él, aunque es un poco más restrictivo que el significado dado al término en el uso popular [12]. En términos más generales, la renta también puede referirse

a los pagos periódicos realizados por el uso temporal de un bien, como una casa, un automóvil, un televisor, entre otros, con el entendimiento de que el bien alquilado debe ser devuelto a su propietario en esencialmente las mismas condiciones físicas [13].

4.2.2. Apartado Técnico

Análisis Exploratorio de Datos (EDA) El análisis exploratorio de datos (EDA) como se menciona en [14], [15], [16], [17] es un enfoque para analizar conjuntos de datos para resumir sus características principales. Durante el EDA, se emplean herramientas gráficas como histogramas, diagramas de dispersión, gráficos de cajas y diagramas de barras para representar visualmente los datos y analizar su distribución, variabilidad y relaciones. Además, se pueden calcular medidas descriptivas como la media, la mediana, la desviación estándar y los cuartiles para resumir las características numéricas de los datos.

El EDA también puede incluir la identificación de valores atípicos o datos anómalos que se desvían significativamente de la tendencia general del conjunto de datos. Estos valores atípicos pueden ser el resultado de errores de medición, errores de entrada de datos o pueden indicar la presencia de eventos inusuales o anómalos en el fenómeno estudiado.

Al explorar las relaciones entre variables, se pueden detectar patrones, correlaciones o dependencias que ayuden a comprender mejor los datos y proporcionen pistas sobre posibles asociaciones o influencias entre diferentes variables. Esto puede ser especialmente útil en la identificación de variables relevantes para futuros análisis o en la formulación de hipótesis iniciales.

Es importante destacar que el EDA no es un proceso lineal y se realiza de manera iterativa a medida que se van adquiriendo nuevos conocimientos y surgen nuevas preguntas. Además, la selección de técnicas y herramientas específicas puede variar dependiendo del tipo de datos y el objetivo del análisis.

Modelos Lineales Generalizados (GLM) Los modelos lineales generalizados (GLM) fueron pensados para estimar la máxima verosimilitud en modelos de la familia exponencial [18]. Luego, estos modelos se consideran una extensión de los modelos de regresión lineal que permiten manejar variables dependientes con distribuciones diferentes a la normal y enlazar la media de la variable respuesta con el predictor lineal mediante funciones de enlace [19].

Una variable aleatoria Y tiene una distribución en la familia exponencial, si su función de densidad $f_{GLM}(y_i \mid \theta_i, \phi)$ se puede representar como:

$$f_{GLM}(y_i|\theta_i, \phi) = \exp\left\{\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}$$

si ϕ (parámetro de dispersión) es conocido, este es un modelo de la familia exponencial con parámetro canónico o natural θ_i. $b(.)$ satisface que $E(Y_i) = b'(\theta_i)$, con $b'(\theta_i)$ la primera derivada de $b(.)$. $\text{Var}(Y_i) = a(\phi)b''(\theta_i)$, con $b''(\theta_i)$ la segunda derivada de $b(.)$, $c(y_i, \phi)$ es una constante normalizadora.

De lo anterior se sabe también que los GLM se componen de tres elementos principales [19]:

- **Componente Aleatorio:** Especifica la variable respuesta y su función de distribución, perteneciente a la familia exponencial. Las observaciones $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$ provienen de variables aleatorias independientes $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)^\top$.
- **Componente Sistemático:** También llamado predictor lineal, es la combinación lineal del vector de parámetros $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^t$ y la matriz $\mathbf{X}$ de tamaño $n \times p$ que contiene n observaciones de las p variables explicativas: $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.
- **Función de Enlace:** Permite enlazar el componente aleatorio a partir de la media, $\lambda_i = E(Y_i)$ con el componente sistemático así, $g(E(Y_i)) = \eta_i$. $g(.)$ se llama la función de enlace. En esta formulación, los modelos lineales clásicos tienen una distribución normal (o gaussiana) en el primer componente y la función de enlace es la identidad. Los modelos lineales generalizados permiten dos extensiones: primero, la distribución en el primer componente puede provenir de una familia exponencial distinta de la normal y, en segundo lugar, la función de enlace puede ser cualquier función diferenciable monótona [19].

Estructura de autocorrelación espacial En este contexto, el supuesto de independencia entre observaciones (para este caso entre predios rurales) no se cumple. El precio de renta del predio i junto a las características de cada lugar, pueden influir en el precio de renta del predio j, y viceversa. Se debe entender entonces esta dependencia espacial para ser incorporada en el modelo a construir.

La covarianza entre dos variables aleatorias espaciales se define de manera usual como [20]:

$$\text{Cov}(Y(x_i), Y(x_j)) = E[Y(x_i)Y(x_j)] - E[Y(x_i)]E[Y(x_j)] = 0, \quad \text{para } i \neq j$$

Donde $Y(x_i)$ y $Y(x_j)$ son observaciones de una variable aleatoria en las localizaciones de los centroides x_i y x_j, con i y $j = 1, 2, \ldots, N$, $x_i, x_j \in D_x$, siendo D_x la región de interés.

Ahora, es necesario definir para cada unidad de área un conjunto de vecinos. Para este fin existen muchas opciones tales como contigüidad, distancias, kernel, entre otros [20]. Estas relaciones de vecindad se cuantifican a través de la matriz $\mathbf{W}$ de pesos (o ponderaciones) espaciales.

W es una matriz cuadrada, no estocástica, con entradas no negativas y finitas que representan la interdependencia espacial entre las unidades geográficas. Es una matriz de pesos (ponderaciones) espaciales que define la magnitud de las interacciones posibles dentro de un sistema espacial. Se denotará por **W**, así:

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & w_{13} & \dots & w_{1n} \\ w_{21} & 0 & w_{23} & \dots & w_{2n} \\ w_{31} & w_{32} & 0 & \dots & w_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & w_{n3} & \dots & 0 \end{bmatrix}$$

para $i, j = 1, \dots, n$.

Las entradas no nulas corresponden a unidades espaciales relacionadas. Otra de las funciones de la matriz de pesos espaciales es incorporar la multidireccionalidad de las relaciones geográficas a partir del concepto del cambio espacial. Nótese que no se habla de vecinos sino de vecindad. Usualmente, la matriz de pesos espaciales es estandarizada por fila:

$$w'_{ij} = \frac{w_{ij}}{\sum_{j=1}^{n} w_{ij}}, \quad \sum_{j=1}^{n} w'_{ij} = 1, \quad i = 1, 2, \dots, n$$

Índice de Moran El Índice de Moran denotado de ahora en adelante como I [21] toma la forma

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(Y(x_i) - \bar{Y})(Y(x_j) - \bar{Y})}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \sum_{i=1}^{n} (Y(x_i) - \bar{Y})^2},$$

donde n es el número de regiones, $Y(x_i)$ es el valor observado de la variable de interés en la región i, y $\bar{Y}$ es la media de todos los valores. w_{ij} son pesos espaciales que denotan la proximidad espacial entre las regiones i y j, con $w_{ii} = 0$ y $i, j = 1, \dots, n$. La definición de los pesos espaciales depende de la variable de estudio y del contexto específico.

Se puede probar la presencia de autocorrelación espacial usando el Índice de Moran I, que cuantifica cuán similar es cada región con sus vecinas y promedia todas estas evaluaciones. Bajo la hipótesis nula de no autocorrelación espacial, las observaciones Y_i son independientes e idénticamente distribuidas, y I se distribuye asintóticamente de manera normal con media y varianza iguales a

$$E[I] = -\frac{1}{n-1} \quad \text{y} \quad \text{Var}[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 + 2S_0^2}{(n+1)(n-1)^2 S_0^2},$$

donde

$$S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}, \quad S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij} + w_{ji})^2, \quad S_2 = \sum_{k=1}^{n}\left(\sum_{j=1}^{n} w_{kj} + \sum_{i=1}^{n} w_{ik}\right)^2.$$

Los valores del Índice de Moran I usualmente varían de -1 a 1. Valores de I significativamente por encima de $E[I] = -\frac{1}{n-1}$ indican autocorrelación espacial positiva o agrupamiento. Esto ocurre cuando las regiones vecinas tienden a tener valores similares. Valores de I significativamente por debajo de $E[I]$ indican autocorrelación espacial negativa o dispersión. Esto sucede cuando las regiones cercanas tienden a tener valores diferentes. Finalmente, valores de I alrededor de $E[I]$ indican aleatoriedad, es decir, ausencia de patrón espacial.

Cuando el número de regiones es suficientemente grande, I tiene una distribución normal y se puede evaluar si algún patrón dado se desvía significativamente de un patrón aleatorio comparando el puntaje z con la distribución normal estándar.

$$z = \frac{I - E(I)}{\sqrt{\mathrm{Var}(I)}}$$

Un enfoque alternativo para juzgar la significancia es la aleatorización de Monte Carlo. Este método crea patrones aleatorios reasignando los valores observados entre las áreas y calcula el Índice de Moran I para cada uno de los patrones, proporcionando una distribución de aleatorización para el Índice de Moran I. Si el valor observado del Índice de Moran I se encuentra en las colas de esta distribución, se rechaza la suposición de independencia entre las observaciones [22]. Así, podemos probar la autocorrelación espacial siguiendo estos pasos:

1. Establecer las hipótesis nula y alternativa:

$$H_0 : I = E[I] \quad \text{(sin autocorrelación espacial)},$$

$$H_1 : I \neq E[I] \quad \text{(autocorrelación espacial)}.$$

2. Elegir el nivel de significancia α que estamos dispuestos a tolerar, que representa el valor máximo para la probabilidad de rechazar incorrectamente la hipótesis nula cuando es verdadera (usualmente $\alpha = 0{,}05$).

3. Calcular el estadístico de prueba:

$$z = \frac{I - E(I)}{\sqrt{\mathrm{Var}(I)}}.$$

4. Encontrar el valor p para los datos observados comparando el puntaje z con la distribución normal estándar o mediante la aleatorización de Monte Carlo. El valor p es la probabilidad de obtener un estadístico de prueba tan extremo como o más extremo que el estadístico de prueba observado en la dirección de la hipótesis alternativa, asumiendo que la hipótesis nula es verdadera.

5. Tomar una de estas dos decisiones y establecer una conclusión:

   - Si el valor p $< \alpha$, rechazamos la hipótesis nula. Concluimos que los datos proporcionan evidencia para la hipótesis alternativa.
   - Si el valor p $\geq \alpha$, no rechazamos la hipótesis nula. Los datos no proporcionan evidencia para la hipótesis alternativa.

Modelación espacial Es un enfoque sistemático para comprender la configuración espacial de la actividad económica desde una escala local hasta una escala global [23]. En el campo de la estadística, la modelación espacial se refiere a un conjunto de procedimientos analíticos utilizados para derivar información sobre las relaciones espaciales entre los fenómenos geográficos [24]. Involucra el uso de un modelo de datos geográficos, que es una estructura matemática y digital para representar fenómenos sobre la Tierra. Los modelos espaciales son lenguajes formales para expresar mecanismos de procesos geográficos y diseñar flujos de trabajo analíticos para entender estos procesos [24].

En el contexto de la ciencia de datos, la modelación espacial se utiliza para analizar y predecir para una variedad de modelos estadísticos espaciales aplicados a datos referenciados por puntos o areales (rejilla) [25]. Los parámetros se estiman utilizando varios métodos, incluyendo la optimización basada en la verosimilitud y los mínimos cuadrados ponderados basados en variogramas [25].

Modelos de regresión espacial Estos modelos, como los Modelos Autorregresivos Espaciales (SAR) y los Modelos de Errores Espaciales (SEM), pueden ser útiles para capturar la dependencia espacial en los datos [26], [27]. Estos modelos asumen que el valor de la variable de interés en una ubicación dada está influenciado por los valores en las ubicaciones cercanas. Por otro lado, para evaluar el rendimiento de estos modelos es posible usar métricas como el Error Cuadrático Medio (MSE), entre otros. La fórmula que describe los modelos SAR es la siguiente:

$$Y(\mathbf{x}) = \rho \mathbf{W} Y(\mathbf{x}) + \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

Modelos de aprendizaje automático (Machine Learning) Según el libro *"The Hundred-Page Machine Learning Book"*, es un proceso que resuelve problemas prácticos mediante la recopilación de un conjunto de datos y la construcción algorítmica de un modelo estadístico basado en ese conjunto de datos [28]. Este proceso se divide en tres partes principales [28]:

1. Un proceso de decisión: Los algoritmos de aprendizaje automático se utilizan para realizar predicciones o clasificaciones. Utilizando datos de entrada, ya sean etiquetados o no, el algoritmo genera una estimación acerca de un patrón presente en los datos [28].

2. Una función de error: La función de error tiene como propósito evaluar la precisión de las predicciones realizadas por el modelo. Si se dispone de ejemplos conocidos, la función de error permite realizar una comparación y determinar qué tan precisa es la estimación realizada por el modelo [28].

3. Un proceso de optimización del modelo: Si el modelo puede ajustarse de manera más precisa a los puntos de datos del conjunto de entrenamiento, los pesos se modifican para reducir la discrepancia entre los ejemplos conocidos y las estimaciones realizadas por el modelo. El algoritmo repite este proceso de evaluación y optimización, actualizando los pesos de manera autónoma hasta alcanzar un nivel de precisión deseado [28].

Algunos modelos de aprendizaje automático que se pueden adaptar para tener en cuenta la estructura espacial de los datos son las máquinas de soporte vectorial (SVM) y los bosques aleatorios (Random Forest) [29]. Estos métodos pueden capturar relaciones no lineales y complejas entre las variables.

Bosque Aleatorio (Random Forest) Es un método de ensamble basado en árboles de decisión, diseñado tanto para problemas de clasificación como de regresión. En este estudio, su aplicación está enfocada exclusivamente en regresión, donde el modelo aprende a predecir valores continuos en función de múltiples predictores. Se compone de una combinación de predictores de árboles $\{h(x, \Theta_k), k = 1, \ldots\}$, donde cada árbol es entrenado con una muestra aleatoria del conjunto de datos y genera una estimación numérica que luego es promediada entre todos los árboles del bosque para obtener la predicción final [30]. Es una modificación sustancial del *bagging* [31], que crea una colección de árboles no correlacionados, donde para hacer crecer cada árbol se hace una selección aleatoria (sin reemplazo) de los ejemplos en el conjunto de entrenamiento y luego los promedia. En muchos problemas, el rendimiento del bosque aleatorio es muy similar al *boosting*, y son más sencillos de entrenar y ajustar.

Para problemas de regresión, la predicción final del modelo es el valor promedio de las predicciones individuales de cada árbol en el bosque, lo que se expresa como:

$$\hat{f}_{\mathrm{RF}}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(\mathbf{x})$$

donde cada $\hat{f}_b(x)$ representa la predicción generada por un árbol individual en la muestra de bootstrap b. Este esquema de agregación reduce la varianza del modelo y mejora

su capacidad de generalización en comparación con modelos de árbol único [30].

En el caso de regresión, el error de generalización cuadrático medio para cualquier predictor numérico $h(x)$ es

$$E_{\mathbf{X},\mathbf{Y}}(Y - h(\mathbf{X}))^2$$

El predictor de bosque aleatorio se forma tomando el promedio sobre k de los árboles $h(x, \Theta_k)$. Similar al caso de clasificación, se tiene que a medida que aumenta el número de árboles, casi seguramente converge a

$$E_{\mathbf{X},\mathbf{Y}}\left(Y - \text{av}_k h(\mathbf{X}, \Theta_k)\right)^2 \to E_{\mathbf{X},\mathbf{Y}}\left(Y - E_\Theta h(\mathbf{X}, \Theta)\right)^2$$

En el caso de regresión, cada árbol dentro del bosque aleatorio divide el espacio de características en regiones homogéneas con respecto a la variable objetivo. Esto se realiza mediante la selección de una variable de partición j y un umbral s que minimizan el error cuadrático medio dentro de cada región. Formalmente, el criterio de división es:

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y-c_1)^2 + \min_{c_2}\sum_{x_i\in R_2(j,s)}(y-c_2)^2\right]$$

donde cada c_m representa el valor promedio de la variable objetivo en la región correspondiente. La predicción final del árbol de decisión es simplemente la media de los valores en cada región. Como el bosque aleatorio es una combinación de múltiples árboles de regresión, la predicción final del modelo es el promedio de todas las predicciones generadas por los árboles en el ensamble.

Para cada variable de partición y punto de partición, la minimización interna es la media de cada región.

$$\hat{c}_m = \frac{1}{n_m}\sum_{y_i|x_i\in R_m} y_i$$

Y si el árbol final tiene, supongamos M regiones en total, entonces la predicción es

$$\hat{f}(x) = \sum_{m=1}^{M}\hat{c}_m I(x \in R_m)$$

Luego, para mejorar el rendimiento, se extraen repetidamente muestras de bootstrap $(X_i^b, Y_i^b)_{i=1}^N$ de la muestra observada y para cada muestra se ajusta un árbol de regresión $\hat{f}_b(x)$; el bosque aleatorio lo que hace es reducir la correlación entre los árboles en el bootstrap. Como hay $p = 5$ predictores, entonces en cada partición se usa solamente $m < p$ predictores, escogidos aleatoriamente. Por último, se realiza el promedio entre muestras de bootstrap para obtener el predictor [29]:

$$\hat{f}_{\mathrm{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(\mathbf{x})$$

Máquinas de soporte vectorial (SVM) Es una técnica de clasificación, aproximación de funciones y regresión desarrollada por Vladimir Vapnik y su grupo en ATT Bell Labs [32].

En un problema de regresión se tiene una base de datos dada por $G = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ y alguna función desconocida $g(\mathbf{X})$ y se requiere determinar una función f que se aproxime a $g(\mathbf{X})$, con base en el conocimiento de G. La SVM considera funciones de aproximación de la forma:

$$f(\mathbf{x}, \mathbf{c}) = \sum_{i=1}^{d} c_i\varphi_i(\mathbf{x}) + b$$

Las funciones $\{\varphi_i(\mathbf{X})\}_{i=1}^{D}$ son características, y b y $\{c_i\}_{i=1}^{D}$ son coeficientes. Esta forma de aproximación se puede considerar como un hiperplano en el espacio de características D-dimensional definido por las funciones $\varphi_i(\mathbf{X})$. Los coeficientes desconocidos se estiman minimizando la siguiente función:

$$R(c) = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(\mathbf{x}_i, c)|_\epsilon + \lambda||c||^2$$

Donde λ es una constante y se ha definido la siguiente función de error robusta:

$$|y_i - f(\mathbf{x}_i, c)|_\epsilon = \begin{cases} 0 & \text{si } |y_i - f(\mathbf{x}_i, c)| < \epsilon \\ |y_i - f(\mathbf{x}_i, c)| - \epsilon & \text{en otro caso} \end{cases}$$

Se ha demostrado que la función que minimiza la ecuación anterior depende de un número finito de parámetros y tiene la siguiente forma [33]:

$$f(\mathbf{X}, \alpha, \alpha^*) = \sum_{i=1}^{N}(\alpha_i^* - \alpha_i)K(\mathbf{X}, \mathbf{x}_i) + b$$

Donde $\alpha_i^*\alpha_i = 0$, $\alpha_i, \alpha_i^* \geq 0$, $i = 1, \ldots, N$, y $K(\mathbf{X}, y)$ es la función Kernel y describe el producto interno en el espacio de características D-dimensional:

$$K(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{d}\varphi_i(\mathbf{x})\varphi_i(\mathbf{x}')$$

Hay disponibles varias opciones para el kernel, incluidas gaussianas, B-splines de productos tensoriales y polinomios trigonométricos. Los vectores de coeficientes α y α^* se obtienen maximizando la siguiente forma cuadrática:

$$R(\alpha, \alpha^*) = -\epsilon \sum_{i=1}^{N}(\alpha_i^* + \alpha_i) + \sum_{i=1}^{N} y_i(\alpha_i^* - \alpha_i) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_i^* - \alpha_i)(\alpha_i^* - \alpha_i)K(\mathbf{x}_i, \mathbf{x}_j)$$

Sujeto a las restricciones $0 \leq \alpha_i^*, \alpha_i \leq c$ y $\sum_{i=1}^{N}(\alpha_i^* - \alpha_i) = 0$. Debido a la naturaleza del problema de programación cuadrática, solo una pequeña cantidad de coeficientes $\alpha_i^* - \alpha_i$ serán diferentes de cero, y los puntos de datos asociados con ellos se denominan vectores de soporte.

Al usar una función de costo para medir el riesgo empírico, se minimiza el error de regresión entre los valores previstos y reales. La base principal de la SVM es la función insensible a ϵ y la función Kernel. Algunos ejemplos de la función de pérdida insensible a ϵ son:

- Lineal ϵ-insensible:
$$L(y, f(\Lambda, \mathbf{X})) = |y - f(\Lambda, \mathbf{X})|_\epsilon$$
- Cuadrática ϵ-insensible:
$$L(y, f(\Lambda, \mathbf{X})) = |y - f(\Lambda, \mathbf{X})|_\epsilon^2$$
- Conteo de errores ϵ-insensible:
$$L(y, f(\Lambda, \mathbf{X})) = \begin{cases} 0 & \text{si } |y - f(\Lambda, \mathbf{X})| \leq \epsilon \\ 1 & \text{en otro caso} \end{cases}$$

Luego, las funciones de kernel más utilizadas son [34]:

1. Función de base radial (FKBR):
$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}_j||^2), \gamma > 0$$
2. Multicuadrático inverso:
$$K(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{\sqrt{||\mathbf{x}_i - \mathbf{x}_j|| + r}}$$
3. Polinomial:
$$K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma(\mathbf{x}_j^\top \cdot \mathbf{x}_i) + r)^d$$
4. Sigmoide:
$$K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma(\mathbf{x}_j^\top \cdot \mathbf{x}_i) + r), \gamma > 0$$
5. Lineal:
$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_j^\top \cdot \mathbf{x}_i$$

Para el estudio se deberá elegir tanto una función de pérdida insensible a ϵ, como un kernel que permita minimizar el error de estimación del modelo construido.

Inclusión de dependencia espacial en modelos de aprendizaje automatico Diversos métodos de aprendizaje automático se les puede incorporar propiedades espaciales en el algoritmo de aprendizaje. Los métodos basados en árboles de decisión pueden abordar el componente espacial mediante clusters o jerarquías de grupos de datos similares y se asocia un modelo predictivo con cada grupo, de esa manera se obtiene una función objetivo evaluada en cada nodo del árbol, basado en los grupos creados previamente. Cuando se considera dividir un grupo en un nodo, se ejecuta una prueba que maximiza la reducción de la varianza dentro del grupo. Para tener en cuenta la ausencia de estacionariedad espacial en la variable objetivo, se agregó a esta prueba un término basado en medidas globales de autocorrelación espacial [35].

También se han desarrollado versiones ponderadas geográficamente de árboles de decisión para tener en cuenta la heterogeneidad espacial. Para ello, la matriz de ponderación espacial y el bosque aleatorio se integran en un marco de análisis de regresión local. Se define la matriz de pesos espaciales con algún criterio conocido y se especifican los vecinos de i como: si j $(j \in (1, 2, \ldots, p) \wedge i \neq j)$ es un vecino de la unidad i, el valor de la matriz es 1, $w_{ij} = 1$, y 0 si no es vecino. La unidad espacial i y sus vecinos son las entradas para construir un bosque aleatorio local para la unidad i (BA(i)). Al ejecutar BA(i), se puede calcular la importancia de variables para la unidad espacial i. Por último, se construye un bosque aleatorio local para cada unidad espacial en el área de estudio y se estima la importancia de variables local para cada unidad espacial [36], [37].

Con respecto a SVM, existe una extensión llamada Campos Aleatorios de Vectores de Soporte (CAVS) que modela explícitamente las correlaciones espaciales en datos multidimensionales mediante el uso de campos aleatorios de Markov (CAM) y, más recientemente, campos aleatorios condicionales (CAC) para datos espaciales [38].

Validación cruzada La validación cruzada es una técnica estadística para evaluar la capacidad de generalización de un modelo. Se utiliza principalmente en el aprendizaje automático y en la estadística para estimar el rendimiento de los modelos de aprendizaje cuando se aplican a datos independientes [29].

Los paquetes en R que implementan diferentes estrategias de partición o validación cruzada espacial son blockCV, CAST, kmeans, sperrorest y mlr3spatiotempcv [39]. El interés es predecir la respuesta Y de un objeto o instancia usando un vector de características $X = (x^{(1)}, x^{(2)}, \ldots, x^{(p)})^t \in R^p$ y un modelo $\hat{f}_L$ que ha sido entrenado en una muestra $L = \{(y_i, x_i), i = 1, 2, \ldots, n\}$. El objetivo es estimar el valor esperado del rendimiento (performance) de $\hat{f}_L$:

$$\text{perf}(\hat{f}_L) = E(\, l(Y, \hat{f}_L(X)))$$

donde l es una función de pérdida de valor real y el valor esperado es con respecto a la distribución de probabilidad de X, las características de una instancia (Y, X) extraída

al azar de la población subyacente. La función de pérdida puede tomar la forma del error cuadrático $(Y - \hat{f}_L(X))^2$.
Seleccionando una muestra T de datos de prueba extraídos de la población, se puede estimar el rendimiento condicional de $\hat{f}_L$:

$$\hat{\text{perf}}_T(\hat{f}_L) = \frac{1}{|T|} \sum_{(Y,X)\in T} l(Y, \hat{f}_L(X))$$

Esta representación como un estimador puntual de $\text{perf}(\hat{f}_L)$ subraya la importancia de utilizar una muestra aleatoria para la evaluación del modelo para evitar el sesgo de estimación. Dado que la reutilización de la muestra de aprendizaje L para la prueba, es decir, $T = L$, produciría una resustitución muy optimista o un rendimiento aparente, la validación cruzada (VC) divide la muestra L en conjuntos de prueba y entrenamiento disjuntos. Específicamente, L se divide en K particiones,

$$L = L_1 \cup L_2 \cup \cdots \cup L_K, \quad L_i \cap L_j = \emptyset \, \forall \, i \neq j$$

y se ajusta un modelo $\hat{f}_{(i)}$ en $L_{(i)} = L \setminus L_i$. Esto se repite para $i = 1, 2, \ldots, K$ con el fin de utilizar de manera efectiva toda la muestra para la prueba, manteniendo los conjuntos de entrenamiento y prueba separados en todo momento. Por lo tanto, el estimador K-fold VC se puede escribir como

$$\hat{\text{perf}}_{L,\text{VC}}(f) = \frac{1}{K} \sum_{i=1}^{K} \hat{\text{perf}}_{L_i}(\hat{f}_{L_{(i)}})$$

donde f es un algoritmo de aprendizaje automático, es decir, un mapeo que entrena un modelo $\hat{f}_S$ utilizando alguna muestra de entrenamiento adecuada S. El estimador VC K-fold es un estimador casi insesgado de la medida de rendimiento condicional cuando las observaciones se extrajeron de forma independiente. Los métodos de VC espacial más utilizados son Spatial leave-one-out, Leave-one-block-out VC, VC a nivel de bloque y VC para datos espacio temporales [39].

Mediante los paquetes CARET (Entrenamiento de Clasificación y Regresión) y CAST (Aplicaciones de CARET para Modelos Espacio-Temporales) de R se puede llevar a cabo la estimación de algoritmos de Aprendizaje automático usando datos con componentes espacial y espacio-temporal. El desarrollo de este paquete ha permitido implementar métodos de selección de características hacia adelante y validación orientada al objetivo [40].

El rendimiento de una validación cruzada (VC) espacial de K-fold se puede comparar a su vez con Leave-Location-Out (LLO), Leave-Time-Out (LTO) y Leave-Location-and-Time-Out (LLTO) VC. Para disminuir el sobreajuste y mejorar el rendimiento del modelo, el paquete implementa una selección de características directa que selecciona variables predictoras adecuadas en vista de su contribución al rendimiento orientado al

objetivo: Eliminación de características recursivas (ECR) y selección de características hacia adelante (SCA) [40], [41].

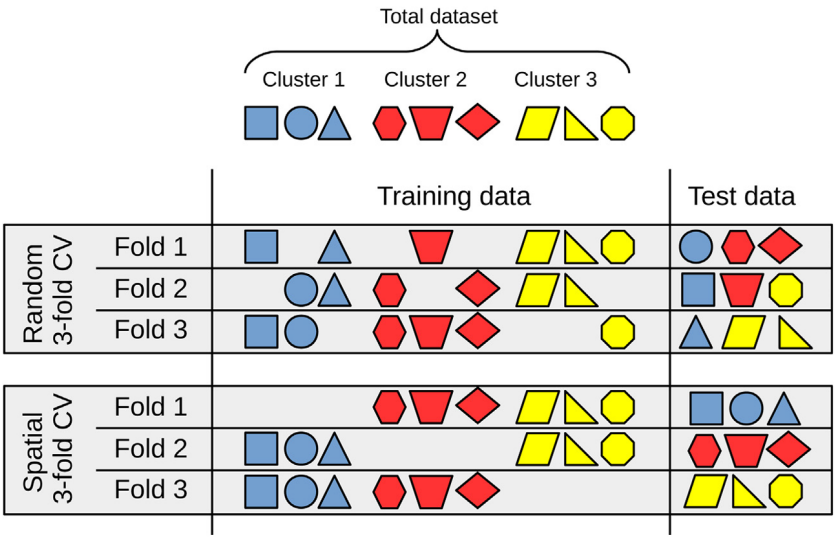

Figura 1: Validación cruzada

Una representación gráfica de la VC espacial implementado por Meyer y colaboradores, se encuentra en la figura 1. Esta muestra el concepto de validación cruzada, aleatoria y espacial. VC espacial significa que los datos se dividen en pliegues según la ubicación espacial (por ejemplo, un grupo espacial o un bloque espacial, en este caso representado por un color único).

Finalmente, mencionar que mediante la validación cruzada es posible calcular y estimar los siguientes indicadores/métricas para evaluar el rendimiento tanto explicativo como predictivo del modelo:

Coeficiente de Determinación (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde y_i son los valores observados, $\hat{y}_i$ son los valores predichos y $\bar{y}$ es la media de los valores observados.

Error Cuadrático Medio (MSE)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Error Absoluto Medio (MAE)

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

5. Preparación y Análisis de Datos para el Modelado Supervisado

En este capítulo, se describe la gestión y análisis de los datos antes de la construcción del modelo supervisado. A través de varios subcapítulos, se documenta todo el proceso necesario para la creación del conjunto de datos que será utilizado en los procesos posteriores. Se muestra cómo el uso de fuentes externas a los datos originales puede ser útil para la creación de nuevas variables. Estas variables pueden ser fundamentales para el modelado de la renta en predios rurales. La idea es proporcionar una visión completa y detallada de cada paso en la preparación de los datos para el modelado.

5.1. Recopilación de datos

La Sociedad de Activos Especiales (SAE) consolida sus datos en dos formatos. El primero es un archivo plano (.csv) que contiene información de todos los predios que administra actualmente. El segundo es un archivo (.shp) que contiene los polígonos correspondientes a la forma del área de cada predio en su inventario a nivel rural, con corte al 12/12/2022. Estos predios poseen algunas variables asociadas, siendo la más relevante el área del predio en hectáreas. Además, el archivo (.shp) incluye una tabla con atributos similares a los que se encuentran en el archivo plano.

Las variables que contiene la tabla de atributos en el archivo (.shp) son las siguientes:

Variable	**Descripción**
DPTOMPIO	Código del municipio según DANE
FMI	Código del folio de matrícula del predio
TIPO	Rural o Urbano
CLASIFI_AC	La edificación que hay en el predio
DIRECCION	La dirección del predio
REGIONAL	La regional en la SAE que administra el predio
ACTIVO_SOC	Social o no social
ESTADO_FIS	Estado físico del predio
ESTADO_LEG	Estado legal del predio
FECHA_RECE	Fecha en que el predio empezó a ser administrado por la SAE
ESTADO_INV	Estado administrativo actual del predio
AVALUO_CAT	Valor catastral del predio
VIGENCIA	Es un año
ESTADO_OC	Estado de ocupación del predio
AVALUO_COM	Valor comercial del predio
ESTIMATIVO	Valor aproximado de renta del predio

Tabla 1: Descripción de las variables generales otorgadas por la SAE

Para realizar una buena estimación de la renta de los predios rurales, resulta esencial consolidar un conjunto de datos mucho más amplio que el anterior. Esto con el objetivo de proporcionar una perspectiva multidimensional y una mayor variabilidad en cuanto a las implicaciones de la ubicación específica de un predio.

Inicialmente, se filtran los predios para que el conjunto de datos corresponda exclusivamente a predios rurales, utilizando la variable **TIPO**. Posteriormente, se realiza la primera adición al conjunto de datos, que corresponde a los centroides de los predios rurales mediante su longitud X y su latitud Y. Se realiza una depuración de datos duplicados y se consolidan los predios resultantes en un mapa:

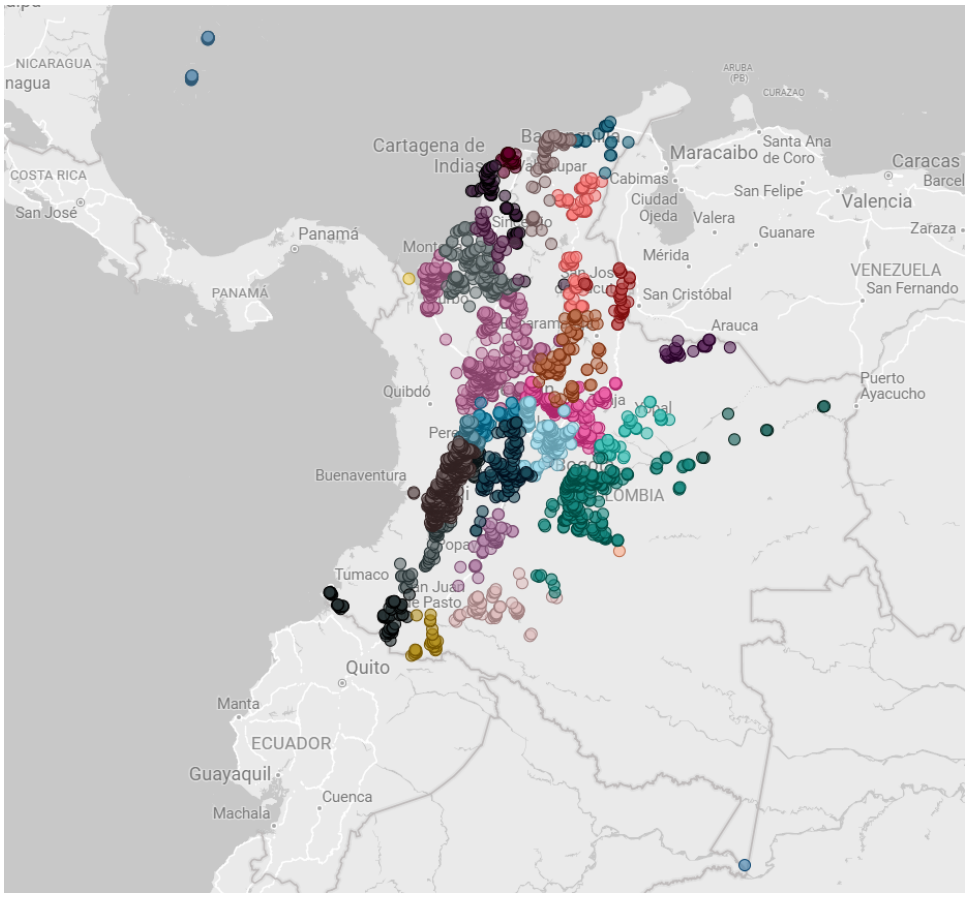

Figura 2: Mapa centroides predios SAE

En la figura 2, los colores indican los diferentes departamentos a los que pertenecen los predios. Es notable cómo hay una zona del país, principalmente la selva del Amazonas, donde no se registra ningún predio rural. También se observan predios en San Andrés y Providencia.

En total se cuenta con 6459 predios rurales, administrados por la SAE a Diciembre del año 2022.

Luego de la revisión de duplicados, se filtra el set de datos para obtener un total de 5872 predios, de los cuales algunos no tenían valor en la variable **ESTIMATIVO**, que es la variable de interés.

En un segundo momento se inicia con la consolidación de variables adicionales que puedan servir tanto para la modelación como para una caracterización detallada de la situación de cada predio. Los grupos de variables que se usan, se ejemplifican en el esquema que se presenta en la figura 3.

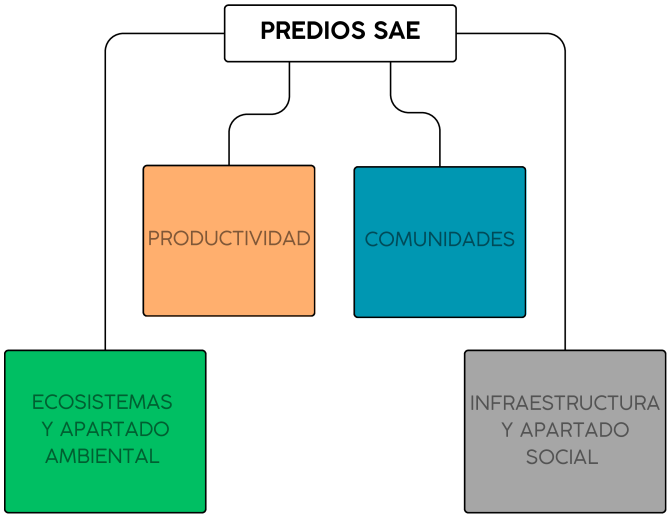

Figura 3: Esquema grupos de variables agregados

Estos grupos de variables se han diseñado teniendo en cuenta los factores que podrían afectar el precio de renta de un predio, discutido previamente con la SAE. Por ejemplo, se considera la productividad de la tierra de cada predio, lo cual se puede determinar mediante indicadores como *PASTOS_CON_PROBABLE_ACTVIDAD_GANERA*, *MEDIA_PONDERADA*, *DIVERSIDAD_DE_CULTIVOS*, *CLASIFICACION_IGAC* y *AREAPREDIOHA*. También se incluye el *precio_entorno*, que refleja el valor promedio en las áreas cercanas.

En lo que respecta al apartado ambiental, variables como *ECOSISTEMAS*, *TRASLAPE_RUNAP*, *AMENAZA_AMBIENTAL*, y *TRASLAPE_ZRF* permiten identificar restricciones y condiciones ambientales que pueden limitar la producción o el acceso a la tierra. Estas son particularmente relevantes en los análisis de uso de suelo.

Por otro lado, el grupo de variables relacionado con comunidades incluye *CLASIFI_AC*, *DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM*, *DIST_ZRC_KM*, *DIST_TITULOSCOLECTIVOS_KM* y *DIST_RESGUARDOS_KM*, que sirven para identificar la proximidad de los predios a comunidades campesinas y étnicas constituidas en diferentes zonas de Colombia.

Finalmente, en el apartado de infraestructura y factores sociales, variables como *AREA_MINERIA*, *IPM_SECCION_RURAL*, *TRASLAPE_EXTRACCION_COMBUSTIBLE*, *DISTANCIA_A_VIAS_PRINCIPALES*, *DIST_VIAS_PRIMARIAS_KM*, *DIST_VIAS_SECUNDARIAS_KM* y *DIST_VIAS_TERCIARIAS_KM* permiten analizar la accesibilidad y conectividad de los predios,además de condiciones frente a la calidad de vida que podrían tener.

Con toda esta información, se realizan diferentes cruces y cálculos que permiten obtener de forma detallada un conjunto de datos, resultado del cruce de diversas capas geográficas con el archivo (.shp) proporcionado por la SAE.

La tabla 11 en el apartado de anexos, contiene el nombre de las variables consolidadas para cada grupo de datos.

5.2. Análisis de datos para el modelado

Una vez terminado este proceso de preparación de los datos, los pasos a seguir son los siguientes:

Se realiza el análisis exploratorio de las variables. Lo anterior mediante algunas estadísticas descriptivas de las variables de interés con un fuerte enfoque en la variable dependiente identificada como 'ESTIMATIVO'. Adicional a esto, se inician los cruces con otras variables para explorar posibles relaciones y patrones.

A continuación, se llevó a cabo una etapa de selección de variables, que comenzó con la identificación de aquellas que, por su contexto o falta de datos, no podían incluirse en el modelo. Posteriormente, se implementaron tres estrategias principales de selección de variables: *backward selection*, *forward selection* y *stepwise selection*, utilizando el criterio de información de Akaike (AIC) como métrica de evaluación. Estas estrategias permitieron refinar el conjunto de variables seleccionadas, optimizando el balance entre la complejidad del modelo y su ajuste a los datos. Además, se revisaron las correlaciones entre las variables para identificar posibles redundancias o relaciones significativas, lo que contribuyó a mejorar la interpretabilidad y desempeño del modelo final.

Antes de proceder con el modelado, todas las variables predictoras fueron estandarizadas para asegurar comparabilidad y estabilidad en los algoritmos utilizados. Se utilizó la estandarización por *z-score*, donde cada variable X_i fue transformada según la ecuación:

$$X_i = \frac{x_i - \mu_i}{\sigma_i} \tag{1}$$

donde μ_i y σ_i corresponden a la media y desviación estándar de la variable X_i. Esta transformación es particularmente relevante para modelos como *SVM* y regresiones penalizadas, ya que mejora la estabilidad numérica y evita sesgos por diferencias de escala.

Para abordar la dependencia espacial, se revisó el Índice de Moran, el cual permite evaluar la autocorrelación espacial entre el precio de renta de un predio específico y el precio de renta de sus predios vecinos. La especificación de dependencia espacial se realizó mediante la construcción de múltiples matrices de pesos espaciales $\mathbf{W}$, evaluando distintos criterios de vecindad. Se consideraron esquemas de contigüidad, como el método de la torre (*Rook contiguity*) y el método de la reina (*Queen contiguity*), además de grafos espaciales como la triangulación de Delaunay y vecinos relativos. También se analizaron matrices basadas en distancia, incluyendo el criterio de los k-vecinos más cercanos (k-*nearest neighbors*, con $k = 1, 2, 3, 4$).

Para seleccionar la mejor matriz de pesos espaciales, se compararon todas las especificaciones mediante pruebas de significancia del Índice de Moran, minimizando el valor p obtenido en cada caso. La matriz $\mathbf{W}$ final utilizada en el modelo fue aquella que mostró el menor p-valor, garantizando la mayor evidencia de autocorrelación espacial en los datos. Este proceso permitió incorporar de manera óptima la estructura espacial en el análisis y asegurar la robustez de los modelos en términos de dependencia geográfica.

En cuanto a la elección de modelos a trabajar, se adoptó un enfoque centrado en el uso de modelos de aprendizaje supervisado con dependencia espacial. Para ello, se implementaron modelos basados en regresión con dependencia espacial (*SAR - Spatial Lag Error*), complementados con la aplicación de un *Support Vector Machine* (SVM) con *kernel* polinomial. Además, se evaluaron modelos de *Random Forest* para explorar su desempeño en este contexto. La implementación de estos modelos se llevó a cabo en `R` utilizando los paquetes `spdep`, `e1071` y `randomForest`, gestionando su entrenamiento y validación a través de la librería `caret` para asegurar reproducibilidad en la estimación de los mismos.

Finalmente, en la etapa de validación, se utiliza la validación cruzada para comparar los modelos, determinar si la inclusión de la dependencia espacial mejora las estimaciones y evaluar el nivel predictivo de los modelos desarrollados. Se empleó validación cruzada k-fold con $k = 10$, complementada con remuestreo *bootstrap* para la estimación de intervalos de confianza.

La comparación final entre modelos se realizó utilizando métricas como el coeficiente de determinación ajustado (R^2_{adj}), el error cuadrático medio (*MSE*) y el error absoluto medio (*MAE*). Estos indicadores permitieron evaluar la precisión de las estimaciones

y determinar cuál de los enfoques proporcionaba una mejor predicción de la renta en predios rurales bajo dependencia espacial.

6. Análisis Exploratorio de las Variables para la Estimación de la Renta en Predios Rurales

El conjunto de datos consolidado contiene información inicial para 6459 predios rurales. Tras una revisión inicial de duplicados, se redujo a 5872 registros. Posteriormente, se llevó a cabo un proceso de depuración y transformación de las variables, pasando de 148 a 30. A partir de estas, se aplicaron métodos de selección de variables que redujeron el conjunto a 11 variables predictoras. Finalmente, se eliminaron los registros con datos faltantes, obteniendo un conjunto de datos definitivo compuesto por 11 variables predictoras, una variable dependiente y un total de 4495 registros. Este conjunto de datos optimizado será la base para la construcción del modelo espacial.

Los predios rurales están distribuidos en 31 de los 32 departamentos del país, específicamente en 503 de los 1123 municipios existentes. En total, estos predios abarcan un área de 334606.9 hectáreas. Según la clasificación de área del IGAC, el 43.04 % de estos predios son microfundios, mientras que el 26.27 % tiene áreas medianas, siendo estos dos los tamaños de predios más comúnmente encontrados en el conjunto de datos.
La variable de mayor interés en el estudio es 'ESTIMATIVO', la cual corresponde a los precios de renta de los diferentes predios. Esta variable posee un 1.52 % de datos faltantes y un 11.73 % de datos atípicos, aspectos que se deben considerar en análisis posteriores.

La tabla 12, ubicada en el apartado de anexos, contiene los porcentajes de datos atípicos para cada variable existente en el conjunto de datos.

Variables	Media	Desviación	Mínimo	Máximo
Área Predio (ha)	40.45	107.49	0.0029	2005.51
Estimativo	1,873,734	4,182,667	20035	49,906,349

Tabla 2: Estadísticas descriptivas básicas de las variables Área y Estimativo

Los resultados de la tabla 2 presentan las estadísticas descriptivas de las variables Área y Estimativo (Renta). La media del área de los predios rurales es de 40.45 hectáreas, con una desviación estándar de 107.49 hectáreas, lo que indica una variabilidad moderada en el tamaño de los predios. El área mínima registrada es de 0.0029 hectáreas, y el área máxima alcanza las 2005.51 hectáreas, reflejando la existencia tanto de pequeños predios como de extensiones de tierra considerablemente grandes en el conjunto de datos.

En cuanto a la renta estimada, la media es de 1,873,734 COP, con una desviación estándar de 4,182,667 COP, lo que evidencia una dispersión significativa en los valores de renta, aunque menor en comparación con los valores anteriores. El valor mínimo de renta es de 20035 COP, y el máximo es de 49,906,349 COP, lo que muestra la gran

disparidad en los arriendos rurales. Estos resultados destacan que, aunque existen predios con valores de renta notablemente bajos, estos no presentan valores extremos poco creíbles. Sin embargo, será importante realizar un análisis más detallado de los valores atípicos para garantizar que no afecten negativamente los resultados de la modelación.

Variables	**Q1**	**Mediana (Q2)**	**Q3**
Área Predio (ha)	0.6533	5.3764	32.0192
Estimativo	133690	513043	1,639,610

Tabla 3: Cuartiles de las variables Área y Estimativo

Los resultados de la tabla 3 proporcionan una visión detallada de la distribución de las variables 'Área Predio (ha)' y 'Estimativo' (Renta).

Para la variable 'Área Predio (ha)', sus resultados muestran una gran variabilidad en el tamaño de los predios, con una concentración significativa de predios más pequeños y una cola larga hacia predios de mayor tamaño. Esto sugiere que, aunque hay algunos predios muy grandes, la mayoría de los predios son relativamente pequeños.

En cuanto a la variable 'Estimativo' (Renta), estos resultados reflejan una considerable dispersión en los valores de renta, con una concentración significativa de predios con rentas más bajas y una cola larga hacia rentas más altas. Sugiriendo que, aunque hay algunos predios con rentas muy altas, la mayoría de los predios tienen rentas relativamente bajas.

La interpretación de estos cuartiles permite entender que es muy probable que existan diferencias significativas, en ambas variables y que estas repercutan directamente sobre la productividad y el valor de los predios.

En cuanto a la comparación del precio de renta con respecto al área de los predios, se observa lo siguiente:

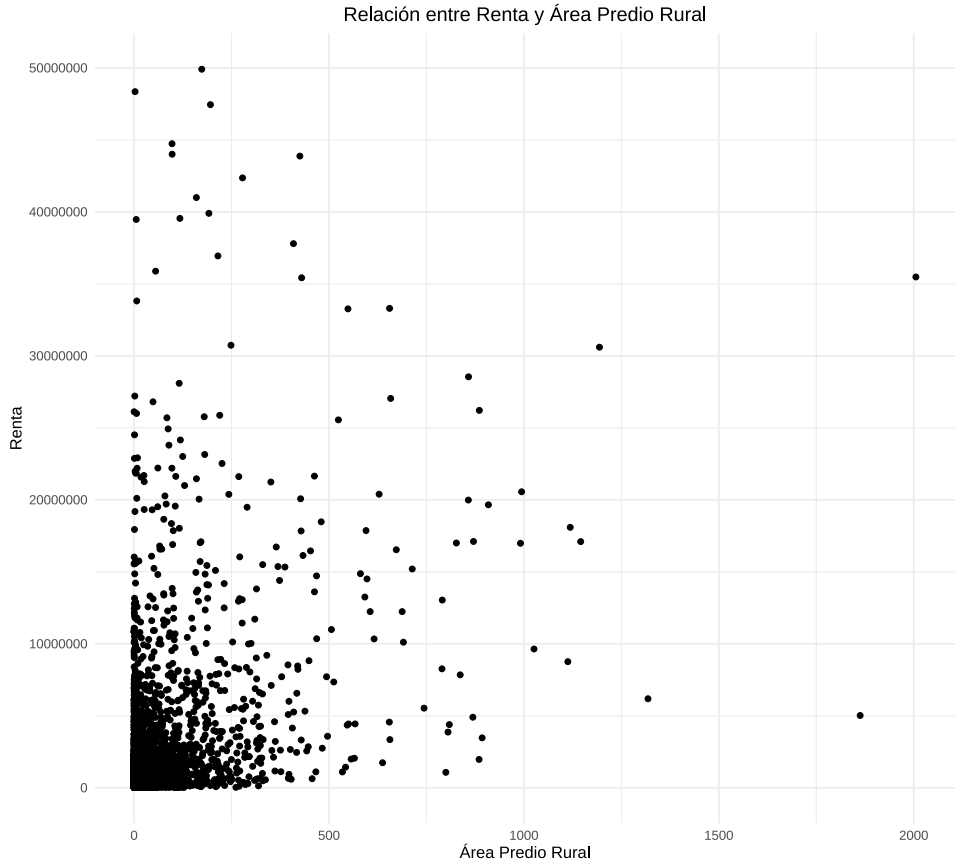

Figura 4: Dispersión entre la renta y el área en Ha del predio rural

La Figura 4 muestra que la mayoría de los predios tienen rentas por debajo de los 50 millones de pesos y áreas menores a 2500 hectáreas. Un aspecto particular del gráfico de dispersión, especialmente en relación con sus datos atípicos, es que no necesariamente los predios de mayor área son los que tienen un precio de renta más elevado. De hecho, existen predios con áreas muy reducidas que tienen altos valores de renta asociados. Este análisis subraya la importancia de considerar múltiples variables al evaluar el valor de los predios rurales, más allá de su tamaño.

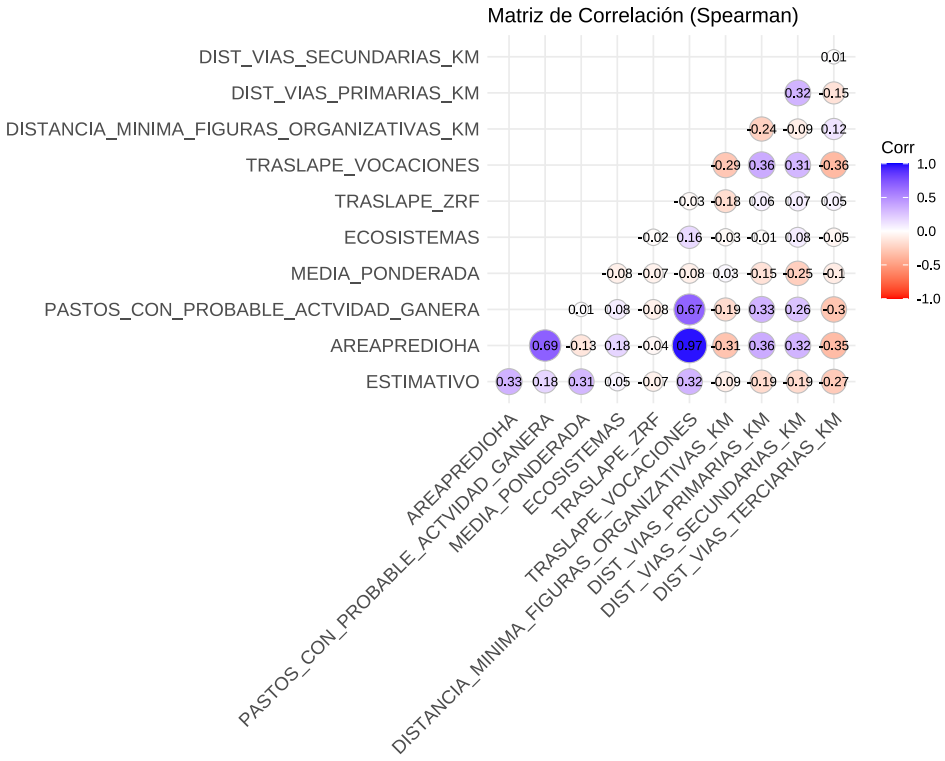

Figura 5: Matriz de correlación por método de Spearman

La matriz de correlación de Spearman (figura 5) revela que la variable **ESTIMATIVO**, que representa el cálculo del arriendo (renta) de cada predio, tiene una correlación positiva moderada con **AREAPREDIOHA** (0.329), lo que sugiere que los predios más grandes tienden a tener un mayor valor estimativo. Además, muestra una correlación débil con **PASTOS_CON_PROBABLE_ACTVIDAD_GANERA** (0.182), indicando que la presencia de actividad ganadera tiene una relación menos pronunciada con el valor estimativo. En cuanto a otras variables, **MEDIA_PONDERADA** presenta una correlación moderada con **ESTIMATIVO** (0.313), mientras que variables como **ECOSISTEMAS** y **DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM** muestran correlaciones muy débiles, indicando una mínima relación con el valor estimativo. Por otro lado, las correlaciones negativas con variables como **DIST_VIAS_PRIMARIAS_KM** (-0.186) y **DIST_VIAS_TERCIARIAS_KM** (-0.270) sugieren que los predios más cercanos a estas vías tienden a tener valores estimativos menores. En general, el valor estimativo se ve principalmente influenciado por el tamaño del predio y, en menor medida, por la actividad ganadera y las características relacionadas con la infraestructura vial.

El conjunto de datos de la SAE y presenta una amplia variabilidad en sus características. La variable **ESTIMATIVO**, que representa el valor estimado del arriendo, tiene un valor mínimo de 20,035 COP y un valor máximo de 49,906,349 COP, con una media de 1,873,734 COP, lo que refleja la diversidad de valores estimativos dentro de los predios rurales. En cuanto al tamaño de los predios, la variable **AREAPREDIOHA** muestra un rango que va desde 0.0029 hectáreas hasta 2,005.51 hectáreas, con una media de 40.45 hectáreas, indicando una gran dispersión en el tamaño de los predios. La presencia de actividad ganadera, representada por **PASTOS_CON_PROBABLE_ACTVIDAD_GANERA**, varía desde valores cercanos a 0 hasta un máximo de 1,534.31, con una media de 22.93, lo que sugiere que aunque algunos predios tienen una actividad ganadera significativa, la mayoría no la tiene de forma destacada.

Otras variables, como **MEDIA_PONDERADA**, presentan una media de 50.25, con un rango de 0 a 96, lo que refleja una amplia gama de valores en los indicadores utilizados. Este indicador, creado por la UPRA (Unidad de Planificación Rural Agropecuaria), se emplea para clasificar la calidad del suelo para fines productivos de siembra de productos. El valor de **MEDIA_PONDERADA** se encuentra en una escala que va de 0 a 100, donde los predios con valores entre 0 y 30 son considerados tierras improductivas, aquellos con valores entre 30 y 60 se clasifican como medianamente productivos, y los que superan 60 se consideran altamente productivos. Sin embargo, en este conjunto de datos, no se observa ningún predio que haya alcanzado el valor máximo de 100, lo que sugiere que todos los predios tienen una calidad de suelo que oscila entre mediana y alta, pero ninguno es calificado como de calidad máxima para la producción agrícola. Este indicador es crucial para evaluar la capacidad de los predios para cultivar productos y, por lo tanto, tiene un impacto directo en la rentabilidad y uso de la tierra para fines agropecuarios.

En cuanto a la infraestructura vial, las distancias mínimas a diferentes tipos de vías también varían considerablemente, destacando distancias menores a las vías primarias y secundarias, con medias de 3.82 km y 5.03 km, respectivamente. La **DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM** tiene una media de 5.85 km, lo que sugiere que los predios se encuentran a distancias relativamente grandes de figuras organizativas clave. En general, los predios presentados en este conjunto de datos son diversos tanto en su tamaño como en sus características relacionadas con la actividad económica, la infraestructura y las características geográficas.

7. Análisis y Preparación de Datos con Dependencia Espacial

Análisis de dependencia espacial Para obtener el Índice de Moran, se cargó un conjunto de datos que incluye información sobre predios rurales, como coordenadas y precios de renta (**ESTIMATIVO**). También se cargó un shapefile con los límites de los predios rurales en Colombia y se transformó a un sistema de referencia espacial adecuado (EPSG:21897). Las coordenadas de los predios se convirtieron de un sistema geográfico (EPSG:4326) a uno proyectado (EPSG:21897) para asegurar la consistencia con el shapefile. Luego, se visualizaron los datos utilizando gráficos estáticos y dinámicos para explorar la distribución espacial de los predios.

Posteriormente, se combinaron los datos del shapefile con la columna **ESTIMATIVO** del conjunto de datos, utilizando la columna **FMI** como llave. Se filtraron y limpiaron los datos para eliminar registros con valores nulos y duplicados, obteniendo un conjunto de predios únicos con valores válidos de **ESTIMATIVO**. Se definieron 20 matrices de pesos espaciales para modelar las relaciones de vecindad entre los predios. Finalmente, se calculó el Índice de Moran para cada matriz de pesos, identificando la matriz que proporcionaba el p-valor más bajo, lo que indica la mejor representación de las relaciones espaciales entre los predios. La matriz seleccionada fue la matriz de pesos **rook_nb_b**, que considera las relaciones de vecindad tipo rook"(vecinos compartiendo un borde) y utiliza el estilo de pesos "B".

Weights style	**n**	**nn**	**S0**	**S1**	**S2**
B	4264	18181696	9232	18464	120336

Tabla 4: Resumen de constantes de pesos

La matriz de pesos **rook_nb_b** generó los valores mostrados en la Tabla 4. Aquí, **n** representa el número de predios, **nn** es el número de pares de predios, **S0** es la suma de los pesos, **S1** y **S2** son sumas ponderadas que reflejan la estructura de la matriz de pesos.

Estadística	**Valor**
Moran I statistic	0.276
Expectation	-0.00023
Variance	0.00017
p-value	<2.2e-16
Moran I statistic standard deviate	20.951

Tabla 5: Resultados del Índice de Moran

El Índice de Moran se calculó utilizando la matriz de pesos seleccionada, obteniendo los resultados presentados en la Tabla 5. La estadística de Moran I es 0.2758, con una

expectativa de -0.0002 y una varianza de 0.0002. El valor p es menor que 2.2e-16, lo que indica una fuerte evidencia contra la hipótesis nula de no autocorrelación espacial.

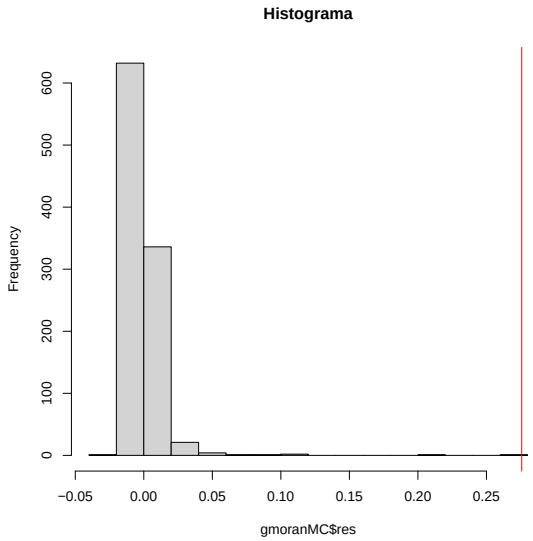

Figura 6: Gráfica del Índice de Moran utilizando el método de Monte Carlo

Para validar estos resultados, se utilizó el método de Monte Carlo, obteniendo la gráfica mostrada en la Figura 6. Esta gráfica refuerza la conclusión de que existe una autocorrelación espacial significativa en los precios de renta de los predios rurales.

Precio del entorno Una vez se ha demostrado la existencia de una autocorrelación espacial significativa en los precios de renta de los predios rurales, mediante el índice de Moran, surge la necesidad de construir una variable que permita incorporar la dependencia espacial en los modelos de aprendizaje automático. Este paso es fundamental, ya que algoritmos como SVM y Random Forest no están diseñados para modelar explícitamente la dependencia espacial, a diferencia de los modelos SAR - Spatial Lag Error, que lo hacen automáticamente. Por lo tanto, la variable denominada "precio del entorno"se construye con el propósito de capturar esta dependencia espacial y permitir que los modelos de machine learning consideren cómo los precios de renta de predios vecinos están relacionados con el precio de renta de un predio específico.

Cálculo del Precio del Entorno El cálculo del precio del entorno se lleva a cabo mediante un proceso iterativo y sistemático que garantiza una representación adecuada de la dependencia espacial. En primer lugar, para cada predio del conjunto de datos, se identifica su centroide como representación espacial. Con base en esta información, se generan buffers, que son áreas circulares alrededor de cada centroide. Estos buffers se construyen utilizando diferentes radios, que varían desde 1,000 hasta 30,000 metros,

con incrementos de 1,000 metros entre cada iteración.

Dentro de cada buffer, se calcula el promedio de los precios de arriendo de los predios contenidos en su interior. Sin embargo, si un buffer contiene únicamente un predio, el precio del entorno correspondiente se registra como faltante (NA), ya que no es posible calcular un promedio significativo con una única observación. Este procedimiento asegura que los valores registrados reflejen de manera adecuada la interacción espacial entre los predios.

Para cada radio de buffer considerado, se realiza un análisis detallado que incluye el cálculo de la correlación entre la variable objetivo (ESTIMATIVO), que contiene los precios de arriendo registrados por la SAE, y la variable precio del entorno. Además, se calcula el porcentaje de datos faltantes generado por esta metodología, considerando la cantidad de buffers que no cumplen con los requisitos mínimos para el cálculo del promedio (Al menos 2 predios al interior del Buffer).

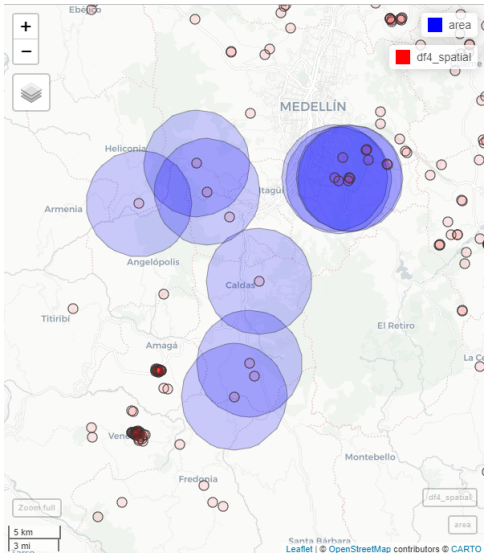

Figura 7: Buffers

A manera de ejemplo, para 10 predios (marcados como puntos rojos en la figura 7), los buffers, representados en color azul, muestran cómo se lleva a cabo el cálculo de los precios del entorno a una distancia específica. Cada buffer contiene predios vecinos cuyos precios de arriendo se promedian para estimar el precio del entorno. Sin embargo, si un buffer contiene únicamente un predio, el precio del entorno correspondiente se registra como faltante (NA), ya que no es posible calcular un promedio representativo.
Para determinar la distancia más adecuada para los buffers, se utiliza un método de selección basado en criterios cuantitativos. Esto asegura que la variable de precio del

entorno capture de manera efectiva la dependencia espacial, optimizando su representatividad y utilidad en el modelo.

Selección de la Distancia Óptima La selección de la distancia óptima para los buffers se realiza equilibrando dos criterios fundamentales. Por un lado, se busca maximizar la correlación entre el precio del entorno y el precio de arriendo, asegurando que la variable calculada sea representativa de la dependencia espacial. Por otro lado, es necesario minimizar el porcentaje de datos faltantes, ya que los algoritmos de aprendizaje automático suelen requerir conjuntos de datos completos, y los valores faltantes implican la eliminación de registros del conjunto de datos.
Para facilitar esta decisión, se genera un gráfico combinado que ilustra la correlación (representada por una línea azul) y el porcentaje de datos faltantes (representado por una línea roja) en función de las distintas distancias de buffer evaluadas. Adicionalmente, se destaca con una línea vertical la distancia sugerida, que representa el mejor compromiso entre una alta correlación y un bajo porcentaje de datos faltantes.
Los resultados obtenidos se consolidan en una tabla que detalla las correlaciones y los porcentajes de datos faltantes para cada distancia evaluada. Este análisis permite identificar patrones y determinar qué configuración de distancia es más adecuada para representar la dependencia espacial sin comprometer la cantidad de datos disponibles para el modelo. De esta forma, la metodología asegura que la variable de precio del entorno capture de manera efectiva la dependencia espacial, permitiendo a los modelos de aprendizaje automático incorporar esta característica crucial para mejorar la precisión de sus predicciones.

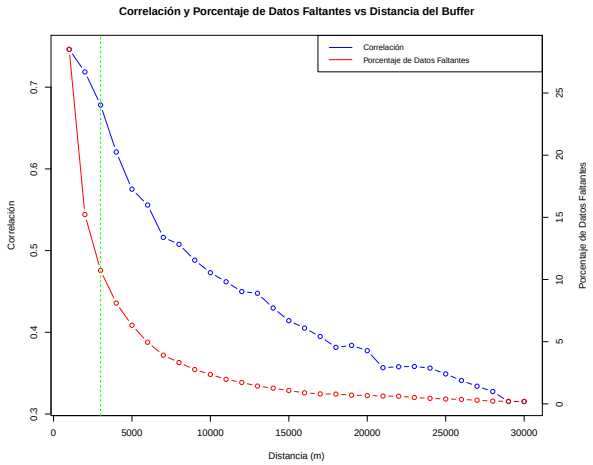

Figura 8: Gráfica de correlación y porcentaje de datos faltantes.

Distancia (m)	Correlación	% Faltantes
1000	0.7462	28.51
2000	0.7187	15.22
3000	0.6782	10.73
4000	0.6207	8.11
5000	0.5751	6.32
6000	0.5558	4.96
7000	0.5162	3.92
8000	0.5077	3.32
9000	0.4883	2.76
10000	0.4730	2.37
11000	0.4619	1.98
12000	0.4499	1.72
13000	0.4478	1.43
14000	0.4298	1.26
15000	0.4143	1.07
16000	0.4052	0.89
17000	0.3948	0.80
18000	0.3816	0.78
19000	0.3840	0.70
20000	0.3775	0.66
21000	0.3567	0.63
22000	0.3579	0.61
23000	0.3582	0.51
24000	0.3562	0.44
25000	0.3491	0.39
26000	0.3411	0.36
27000	0.3340	0.29
28000	0.3275	0.22
29000	0.3153	0.20
30000	0.3154	0.19

Tabla 7: Correlación y porcentaje de datos faltantes según distancia del buffer.

La figura 7 muestra cómo las líneas que representan la correlación y el porcentaje de

datos faltantes se cruzan después de los 30 kilómetros. Esto sugiere que, a partir de esa distancia, los buffers son lo suficientemente grandes como para garantizar la presencia de al menos dos predios en cada buffer. Sin embargo, distancias tan amplias no son apropiadas para capturar la dependencia espacial, ya que exceden el rango en el cual los predios pueden considerarse espacialmente relacionados. Por esta razón, se recurrió a la tabla 7 para analizar las distancias evaluadas y seleccionar una opción coherente que ofreciera un equilibrio entre un nivel aceptable de correlación y un porcentaje de datos faltantes que no redujera significativamente el tamaño del conjunto de datos. La distancia que mejor cumple con estos criterios es un radio de 3 kilómetros para los buffers.

Selección de Variables En este proceso, se implementaron tres estrategias principales de selección de variables: *backward selection*, *forward selection* y *stepwise selection*, utilizando el criterio de información de Akaike (AIC) como métrica de evaluación. Estas estrategias son métodos de selección automatizados que optimizan el modelo balanceando la complejidad y el ajuste a los datos, lo que resulta en modelos más interpretables y con mejor desempeño generalizado.

En el método de *backward selection*, el proceso comienza con un modelo completo que incluye todas las variables predictoras disponibles. A partir de este punto, se eliminan iterativamente aquellas variables que no contribuyen significativamente al modelo, basándose en el AIC. El proceso se detiene cuando la eliminación de más variables no mejora el modelo según este criterio.

Por otro lado, en el método de *forward selection*, el procedimiento comienza con un modelo nulo que no incluye predictores. A partir de este punto, se agregan iterativamente las variables que más contribuyen a mejorar el modelo, evaluando en cada paso el AIC. Este proceso continúa hasta que no se observe una mejora significativa al incluir nuevas variables.

El método *stepwise selection* combina los enfoques anteriores, comenzando generalmente con el modelo completo y realizando tanto inclusiones como exclusiones de variables en cada iteración. Este enfoque permite evaluar dinámicamente el impacto de agregar o eliminar predictores, buscando un equilibrio óptimo en el ajuste del modelo.

Para implementar estos métodos, se emplearon funciones del paquete `MASS` en `R`. Se definieron inicialmente el modelo completo, que contiene todos los predictores, y el modelo nulo, que no incluye ninguna variable. Posteriormente, se aplicaron las estrategias mencionadas y se calcularon métricas como el AIC, R^2, y el R^2 ajustado para cada modelo resultante.

Los resultados de la selección se resumieron en la tabla 6 que incluye el método implementado y las métricas usadas. Además, se identificó el mejor modelo basándose en el

valor mínimo del AIC. Por último, se compararon las variables seleccionadas por cada método para evaluar si las estrategias producían resultados consistentes.

Tabla 6: Resultados de la selección de variables según el método y métricas de evaluación.

Método	**AIC**	**R^2**	**R^2 Ajustado**
Backward	162490.5	0.5586750	0.5548420
Forward	162490.3	0.5583464	0.5546905
Stepwise	162490.5	0.5586750	0.5548420

Resultados del Proceso de Selección de Variables El modelo resultante del proceso de selección de variables optimizado mediante el criterio de información de Akaike (AIC) incluyó 43 predictores significativos. Estos predictores representan un conjunto diverso de variables relacionadas con las características de los predios, el entorno geográfico y socioeconómico, y otras condiciones que influyen en el precio de renta estimado (*ESTIMATIVO*).

Entre las variables seleccionadas destacan factores relacionados con el uso del suelo, como `CLASIFI_AC`, que identifica la clasificación de los predios, y `AREAPREDIOHA`, que mide el área del predio en hectáreas. También se incluyen variables ambientales y socioeconómicas como `PASTOS_CON_PROBABLE_ACTVIDAD_GANERA`, `DIVERSIDAD_DE_CULTIVOS`, `IPM_SECCION_RURAL` y `AMENAZA_AMBIENTAL`. Asimismo, el modelo incorpora variables relacionadas con infraestructura y acceso, como `DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM`, `DIST_VIAS_PRIMARIAS_KM`, `DIST_VIAS_SECUNDARIAS_KM` y `SERVICIOS_BASICOS`.

Es importante destacar la inclusión de la variable construida `precio_entorno`, que captura la dependencia espacial identificada previamente a través del índice de Moran y permite incorporar este efecto en el modelo de aprendizaje automático. Esta variable resultó altamente significativa ($p < 2e-16$) y tuvo un impacto considerable en la mejora del ajuste del modelo.

El modelo final presentó un R^2 de 0.5587 y un R^2 ajustado de 0.5548, lo que indica que las variables seleccionadas explican más del 55 % de la variación en el precio estimado. El valor del estadístico F de 145.8 con un p-valor $< 2{,}2e-16$ refuerza la significancia global del modelo.

Selección de Variables Final La selección inicial de variables generó un modelo con 23 variables que resultaron en 43 predictores, lo cual no era viable para el desarrollo de los modelos de machine learning seleccionados: SVM con kernel polinomial y Random Forest. Para abordar esta limitación, se diseñó un algoritmo que explo-

ró diferentes combinaciones de las variables previamente seleccionadas, evaluando su desempeño mediante criterios de parsimonia y significancia. Como resultado de este proceso, se seleccionó un modelo con 11 predictores clave, logrando un equilibrio entre simplicidad y capacidad explicativa.

El modelo final presenta un R^2 ajustado de 0.5316, ligeramente inferior al 0.5548 del modelo anterior, pero reduce significativamente la cantidad de predictores necesarios, lo cual facilita su implementación en modelos más complejos de machine learning. Este modelo incluye variables cruciales para el contexto de la SAE, como `MEDIA_PONDERADA`, que mide la calidad del suelo, un factor determinante para establecer actividades productivas. También se destacan variables relacionadas con distancias a vías primarias, secundarias y terciarias, que son esenciales para evaluar la complejidad logística en la extracción de productos agrícolas o ganaderos. Además, se incorpora la variable `DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM`, relevante porque las tierras administradas por la SAE están destinadas principalmente a poblaciones rurales organizadas, para quienes el acceso a estas figuras resulta fundamental.

Otra variable clave en el modelo es `precio_entorno`, que captura la dependencia espacial identificada en los análisis previos y refleja la influencia de los predios vecinos en los precios de renta. Su inclusión no solo mejora la capacidad explicativa del modelo, sino que también asegura que esta característica espacial sea considerada en los modelos de machine learning subsecuentes.

El modelo optimizado mantiene una buena capacidad explicativa con un F-statistic de 516.3 (p-valor $< 2{,}2e-16$), lo que confirma su robustez. Además, la eliminación de predictores redundantes o menos significativos mejora la interpretabilidad y simplifica el proceso de modelación en etapas posteriores. Esta reducción de variables también facilita la calibración y validación de los modelos de aprendizaje automático, mejorando su desempeño en la predicción de los precios de renta de los predios rurales administrados por la SAE. Finalmente, las variables seleccionadas para la modelación mediante los métodos de *machine learning* son: `AREAPREDIOHA`, `PASTOS_CON_PROBABLE_ACTVIDAD_GANERA`, `MEDIA_PONDERADA`, `ECOSISTEMAS`, `TRASLAPE_ZRF`, `TRASLAPE_VOCACIONES`, `DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM`, `DIST_VIAS_PRIMARIAS_KM`, `DIST_VIAS_SECUNDARIAS_KM`, `DIST_VIAS_TERCIARIAS_KM` y `precio_entorno`.

Transformación lineal En el análisis de regresión lineal, es fundamental que la variable dependiente cumpla con ciertos supuestos, como la normalidad y homocedasticidad de los residuos. Cuando estos supuestos no se satisfacen, las transformaciones de la variable dependiente, como la logarítmica o la raíz cuadrada, pueden ser útiles para estabilizar la varianza y aproximar la normalidad.

Al aplicar una transformación logarítmica a la variable dependiente log(ESTIMATIVO), el modelo resultante presentó un R^2 de 0.4506 y un R^2 ajustado de 0.4494. Por otro

lado, al utilizar la transformación de raíz cuadrada $\sqrt{\text{ESTIMATIVO}}$, se obtuvo un R^2 de 0.5741 y un R^2 ajustado de 0.5731. Estos valores indican que la transformación de raíz cuadrada proporciona un mejor ajuste en comparación con la transformación logarítmica.

Para evaluar la eficacia de las transformaciones, se compararon los modelos con y sin transformación de la variable dependiente. A continuación, se presenta una tabla resumen de los coeficientes de determinación:

Modelo	R^2	R^2 **ajustado**
Sin transformación	0.5327	0.5316
Raíz cuadrada	0.5741	0.5731
Logarítmica	0.4506	0.4494

Tabla 7: Comparación de R^2 y R^2 ajustado entre modelos con diferentes transformaciones de la variable dependiente.

Los resultados sugieren que la transformación de raíz cuadrada mejora el ajuste del modelo en comparación con el uso de la variable dependiente sin transformar y con la transformación logarítmica.

8. Validación de Estimaciones de Renta

En este capítulo se aborda la construcción de modelos de aprendizaje supervisado con dependencia espacial, utilizando como base los métodos de SVM con kernel polinomial y Random Forest. Los modelos se desarrollaron con un enfoque estándar de entrenamiento y prueba, utilizando un 70 % de los datos para entrenamiento y un 30 % para prueba.

Para garantizar que las variables predictoras fueran comparables en escala y estuvieran listas para el modelado, se realizó un preprocesamiento que incluyó la estandarización (centrado y escalado) de los predictores. Este paso se implementó con la función `preProcess` del paquete `caret` en `R`, generando un conjunto de datos ajustados. La fórmula base utilizada para los modelos es:

$$\begin{aligned}\text{ESTIMATIVO} \sim\ & \text{AREAPREDIOHA} + \text{PASTOS\_CON\_PROBABLE\_ACTVIDAD\_GANERA} \\ & + \text{MEDIA\_PONDERADA} + \text{ECOSISTEMAS} + \text{TRASLAPE\_ZRF} \\ & + \text{TRASLAPE\_VOCACIONES} \\ & + \text{DISTANCIA\_MINIMA\_FIGURAS\_ORGANIZATIVAS\_KM} \\ & + \text{DIST\_VIAS\_PRIMARIAS\_KM} + \text{DIST\_VIAS\_SECUNDARIAS\_KM} \\ & + \text{DIST\_VIAS\_TERCIARIAS\_KM} + \text{precio\_entorno}.\end{aligned}$$

Para asegurar la reproducibilidad de los resultados, se utilizó la semilla `set.seed(1523)` en todos los modelos.

8.1. Modelos SVM con Kernel Polinomial

El método de SVM con kernel polinomial (`svmPoly`) se seleccionó por su capacidad para modelar relaciones no lineales en los datos. Dos variantes del modelo fueron desarrolladas:

- **Modelo sin transformación:** Utilizando la variable dependiente `ESTIMATIVO`.
- **Modelo con transformación:** Utilizando la raíz cuadrada de `ESTIMATIVO`.

El kernel polinomial se caracteriza por tres hiperparámetros clave:

- **degree:** El grado del polinomio que define la complejidad del modelo.
- **scale:** Escala del kernel para ajustar la variación en los predictores.
- **C:** Parámetro de regularización que controla el balance entre el ajuste al entrenamiento y la generalización.

Resultados de los Modelos SVM A continuación, se comparan los resultados de los dos mejores modelos seleccionados basados en el menor RMSE:

Tabla 8: Comparación de los dos mejores modelos SVM.

Modelo	**RMSE**	R^2	**MAE**
Sin transformación	2992799	0.5193	1190317
Con transformación	595.397	0.5973	377.150

El mejor modelo fue el que utilizó la raíz cuadrada de la variable dependiente ($\sqrt{\text{ESTIMATIVO}}$). Este modelo presentó el menor RMSE (595.397), un R^2 más alto (0.5973) y un MAE significativamente menor (377.150) en comparación con el modelo sin transformación. Estos resultados demuestran que la transformación de la variable dependiente mejora tanto el ajuste como la capacidad predictiva del modelo.

8.2. Modelos Random Forest

Se evaluaron cuatro modelos de Random Forest basados en diferentes configuraciones:

- **modelSFit5:** Random Forest con la variable dependiente sin transformar.
- **modelSFit6:** Random Forest con la transformación de raíz cuadrada de la variable dependiente ($\sqrt{\text{ESTIMATIVO}}$).
- **modelSFit7:** Conditional Inference Random Forest (`cforest`) con la variable dependiente sin transformar.
- **modelSFit8:** Conditional Inference Random Forest (`cforest`) con la transformación de raíz cuadrada de la variable dependiente ($\sqrt{\text{ESTIMATIVO}}$).

Diferencias entre Algoritmos y Configuraciones Los modelos `Random Forest` y `Conditional Inference Random Forest` difieren en su implementación y propósito. Mientras que `Random Forest` utiliza el algoritmo tradicional de árboles de decisión basado en bootstrap y selección aleatoria de predictores (`mtry`), `Conditional Inference Random Forest` ajusta los umbrales de división en función de pruebas estadísticas condicionales, lo que reduce el sesgo hacia predictores con múltiples niveles.

En cuanto al tiempo de ejecución, los modelos `cforest` suelen ser más lentos que los modelos tradicionales de Random Forest debido al cálculo de valores p y las pruebas estadísticas para determinar los puntos de corte óptimos en cada nodo.

Hiperparámetros de los Random Forest

- `mtry`: Número de predictores seleccionados aleatoriamente para evaluar en cada división.

- `ntree`: Número de árboles en el modelo (fijo internamente por la función `train` en 500).
- **Transformación de la variable dependiente:** La transformación de raíz cuadrada ($\sqrt{\text{ESTIMATIVO}}$) estabiliza la varianza y mejora el ajuste del modelo en la mayoría de los casos.

Resultados de los Random Forest Los resultados obtenidos para los mejores modelos de cada configuración se resumen en la siguiente tabla:

Tabla 9: Comparación de los mejores modelos Random Forest.

Transformación	**Algoritmo**	**RMSE**	R^2	**MAE**
Sin transformación	Random Forest	2736623	0.5716	1103016
Con transformación	Random Forest	521.2813	0.6871	305.1283
Sin transformación	cforest	2728637	0.5734	1109401
Con transformación	cforest	525.1817	0.6824	313.3746

Entre los modelos evaluados, el mejor desempeño corresponde a **modelSFit6** (Random Forest con transformación de raíz cuadrada). Este modelo presentó el menor RMSE (521.2813), el mayor R^2 (0.6871) y el menor MAE (305.1283), superando tanto a los modelos con `cforest` como a los que no aplicaron transformación. La transformación de la variable dependiente resultó fundamental para mejorar la capacidad predictiva del modelo.

El modelo `cforest` con transformación (modelSFit8) también mostró resultados competitivos, pero su tiempo de ejecución más prolongado lo hace menos práctico para aplicaciones con grandes volúmenes de datos.

8.3. Validación Cruzada en SVM y Random Forest

Para garantizar la robustez y generalización de los modelos, tanto los modelos de SVM (`svmPoly`) como los de Random Forest (`rf` y `cforest`) se validaron mediante un esquema de validación cruzada por remuestreo bootstrap. Este método implica tomar múltiples muestras con reemplazo del conjunto de datos original, generando diferentes combinaciones de entrenamiento y prueba en cada iteración.

8.3.1. Detalles del Remuestreo Bootstrap

- **Número de repeticiones:** 25.
- **Muestras:** Cada muestra tiene el mismo tamaño que el conjunto de datos original ($n = 4995$).

- **Métricas:** En cada iteración se calculan las métricas de desempeño (RMSE, R^2, y MAE) y al final se promedian para obtener una evaluación global.

Aunque los resultados finales se presentan como el promedio de todas las iteraciones del bootstrap, cada una de las 25 repeticiones representa un fold independiente donde el modelo es entrenado en una muestra y evaluado en los datos restantes. Esto permite obtener una distribución de las métricas de desempeño y evaluar su estabilidad a lo largo de los diferentes remuestreos.

8.3.2. Comparación de Validación entre SVM y Random Forest

El esquema de validación cruzada es el mismo para todos los modelos, incluyendo los basados en `svmPoly`, `rf`, y `cforest`. Esto asegura una comparación justa entre ellos, ya que cada modelo se evalúa bajo las mismas condiciones y con el mismo número de muestras.

Adicionalmente, se evaluó la dispersión de las métricas obtenidas a lo largo de los 25 folds del bootstrap. Se calculó el intervalo de confianza del 95 % para cada métrica de desempeño, lo que permite determinar la estabilidad de los modelos y la consistencia de sus resultados a lo largo de las iteraciones.

La validación por bootstrap ofrece ventajas como:

- **Estabilidad de las métricas:** Reduce la variabilidad en los resultados al promediar múltiples iteraciones.
- **Uso eficiente de los datos:** Maximiza el uso del conjunto de datos disponible al emplear muestras con reemplazo.
- **Evaluación robusta:** Permite evaluar el desempeño del modelo en situaciones más generales.

8.3.3. Comparación entre el Mejor Modelo de SVM y Random Forest

La siguiente tabla resume los resultados obtenidos para los mejores modelos seleccionados de cada enfoque:

Tabla 10: Comparación entre el Mejor Modelo de SVM y Random Forest.

Modelo	**Método**	**RMSE**	R^2	**MAE**
SVM	svmPoly (con transformación)	595.397	0.5973	377.150
Random Forest	rf (con transformación)	521.281	0.6871	305.128

Si bien la tabla presenta los valores promedio de las métricas obtenidas en los diferentes folds del bootstrap, también se evaluó la variabilidad de los resultados. Se observó

que la desviación estándar de las métricas en cada fold fue baja, lo que indica que la selección del mejor modelo es estable y no depende de una sola partición del conjunto de datos.

De acuerdo con la tabla, el mejor modelo global es el **Random Forest con transformación de raíz cuadrada (`modelSFit6`)**. Este modelo supera al SVM con transformación en todas las métricas principales (RMSE, R^2, y MAE). La ventaja del Random Forest radica en su capacidad para manejar interacciones y no linealidades de forma eficiente, junto con la ventaja adicional de la transformación aplicada en la variable dependiente para mejorar el ajuste.

9. Conclusiones y trabajos futuros

9.1. Conclusiones

1. **Caracterización de los Predios Rurales:** El análisis del conjunto de datos revela una notable diversidad en las características de los predios rurales en Colombia. El tamaño del predio (`AREAPREDIOHA`) es el factor más determinante para la estimación del valor del arriendo (`ESTIMATIVO`), con una correlación positiva moderada, lo que sugiere que los predios más grandes tienden a tener mayores valores estimativos. Aunque la actividad ganadera, representada por `PASTOS_CON_PROBABLE_ACTVIDAD_GANERA`, tiene una relación débil con el valor de renta, su impacto podría ser más relevante en ciertas áreas rurales, donde la ganadería es una fuente principal de ingresos. La infraestructura vial también juega un papel crucial, con distancias a vías primarias y terciarias presentando correlaciones negativas con los valores de renta, lo que podría estar asociado con la accesibilidad o conectividad a mercados urbanos. En cuanto a la calidad del suelo, medida por `MEDIA_PONDERADA`, los predios en su mayoría tienen una calidad de suelo de media a alta, lo que influye en su rentabilidad, especialmente para actividades agrícolas.

2. **Incorporación de la Dependencia Espacial:** La inclusión de la variable de precio del entorno es esencial para capturar la dependencia espacial en los precios de renta de los predios rurales. La metodología aplicada, que utiliza buffers espaciales de diferentes radios, permite integrar esta dimensión en modelos de machine learning, como `SVM` y `Random Forest`, que por sí solos no consideran la dependencia espacial. Esta incorporación mejora la robustez de los modelos, ya que permite capturar patrones geográficos en los precios de los predios rurales.

3. **Impacto de la Selección de Variables y Transformaciones:** La selección de variables clave mediante métodos como `backward`, `forward` y `stepwise` resultó en un modelo más simple y fácil de interpretar, sin sacrificar la capacidad explicativa. El modelo optimizado, que redujo el número de predictores de 43 a 11, incluyendo la variable `precio_entorno`, facilitó la implementación en algoritmos de machi-

ne learning. Además, la transformación de la variable dependiente utilizando la raíz cuadrada de `ESTIMATIVO` mostró ser más eficaz que otras transformaciones, como la logarítmica, mejorando el ajuste del modelo y aumentando su capacidad predictiva.

4. **Desempeño de los Modelos de Predicción:** Se compararon diversos modelos de machine learning para la estimación del precio de renta en los predios rurales. Los modelos con la transformación de la variable dependiente ($\sqrt{\texttt{ESTIMATIVO}}$) demostraron un mejor rendimiento. En particular, el modelo `Random Forest` con transformación de la raíz cuadrada de `ESTIMATIVO` destacó, logrando un `RMSE` de 521.2813, un $\texttt{R}^2$ de 0.6871 y un `MAE` de 305.1283. Este modelo superó tanto a los modelos de `SVM` como a los modelos sin transformación, demostrando la capacidad de `Random Forest` para manejar interacciones y no linealidades de manera eficiente. A pesar de que el modelo `cforest` mostró resultados competitivos, el tiempo de ejecución más largo lo hace menos adecuado para aplicaciones en grandes volúmenes de datos.

5. **Recomendaciones y Utilidad del Modelo:** Con base en los resultados obtenidos, se recomienda utilizar el modelo `Random Forest` con la transformación de la variable dependiente como el modelo más adecuado para la estimación de los precios de renta de los predios rurales. Su capacidad para capturar complejas interacciones entre variables y su rendimiento superior lo convierten en la opción preferida para aplicaciones futuras en la estimación de precios en el contexto rural colombiano. Además, es crucial seguir trabajando en la automatización de la integración de datos geoespaciales, ya que la construcción del conjunto de datos en este estudio dependió del cruce manual de información geográfica, lo cual aún no está completamente automatizado.

9.2. Trabajos Futuros

En trabajos futuros, se podría explorar la inclusión más profunda de dependencia espacial en los modelos de aprendizaje automático mediante técnicas avanzadas como los árboles de decisión ponderados geográficamente y los Campos Aleatorios de Vectores de Soporte (CAVS). Los árboles de decisión con dependencia espacial permiten considerar la heterogeneidad espacial de los datos al integrar matrices de ponderación espacial y usar un análisis de regresión local. Esta metodología podría mejorarse al incorporar la autocorrelación espacial de la variable objetivo dentro de los nodos del árbol, lo que optimiza la división de los grupos de datos y puede proporcionar un ajuste más preciso para las áreas rurales donde la dependencia espacial es significativa. Además, los modelos de bosque aleatorio local permitirían calcular la importancia de variables específicas para cada unidad espacial, lo que puede generar predicciones más detalladas y localizadas para los predios rurales.

Por otro lado, la implementación de CAVS, que utiliza campos aleatorios de Markov y condicionales para modelar explícitamente las correlaciones espaciales, podría ser un avance importante para abordar la complejidad de los datos espaciales en el contexto rural. Comparar estos métodos con los modelos actualmente implementados en el trabajo de grado, como SVM y Random Forest, podría revelar diferencias en el rendimiento y en la capacidad para capturar patrones espaciales no lineales. Estos enfoques permitirían una validación más robusta de los resultados obtenidos, lo que ofrecería una visión más completa de la efectividad de cada modelo en la estimación de los precios de renta, considerando tanto las interacciones espaciales como las no lineales presentes en los datos.

Referencias

[1] J. C. Muñoz Mora y H. Cardona Jaramillo, «El precio de la tierra: estado del arte de las metodologías de valoración de predios rurales y su aplicación en Colombia,» *Suma de Negocios*, vol. 4, n.º 1, págs. 21-31, 2013.

[2] H. Lim y M. Park, «Modeling the Spatial Dimensions of Warehouse Rent Determinants: A Case Study of Seoul Metropolitan Area, South Korea,» *Sustainability*, vol. 12, n.º 1, pág. 259, 2019. DOI: 10.3390/su12010259. dirección: https://www.mdpi.com/2071-1050/12/1/259.

[3] A. Grybauskas, V. Pilinkienė y A. Stundžienė, «Predictive analytics using Big Data for the real estate market during the COVID-19 pandemic,» *Journal of big data*, vol. 8, n.º 1, págs. 1-20, 2021.

[4] S. W. Hegerty, «Are Rents Excessive in the Central City?: A Geospatial Analysis,» *arXiv.org*, 2021.

[5] *Sobre La Sociedad de Activos Especiales - SAE - SAE SAS*, https://www.saesas.gov.co/informacion_ciudadano/preguntas_respuestas/sobre_sociedad_activos_especiales, Accessed: 2023-11-19, 2023.

[6] S. Savary, S. Waddington, S. Akter et al., «Revisiting food security in 2021: an overview of the past year,» *Food Security*, 2022. dirección: https://link.springer.com/article/10.1007/s12571-022-01266-z.

[7] R. C. Patel, «Food Sovereignty: Power, Gender, and the Right to Food,» *PLOS Medicine*, 2012. dirección: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001223.

[8] P. A. Castro, J. P. Bustos y P. Rueda-Guevara, «Strengthening strategies of food security and food sovereignty in the context of the COVID-19 emergency in Colombia,» *Biomédica*, 2022. dirección: http://www.scielo.org.co/scielo.php?pid=S0120-41572022000500026&script=sci_abstract.

[9] *Gobierno firma 'Acuerdo de Támesis' para entrega de tierras a campesinos*, https://www.eltiempo.com/politica/gobierno/gobierno-firma-acuerdo-de-tamesis-para-entrega-de-tierras-a-campesinos-713612, Accessed: 2023-11-19, 2022.

[10] M. G. Bridge, «1The meaning of personal property,» en *Personal Property Law*, Oxford University Press, nov. de 1996, ISBN: 9781854315816. DOI: 10.1093/acprof:oso/9781854315816.003.0001. eprint: https://academic.oup.com/book/0/chapter/155292903/chapter-pdf/39432495/acprof-9781854315816-chapter-1.pdf. dirección: https://doi.org/10.1093/acprof:oso/9781854315816.003.0001.

[11] *Predio | Qué es, Tipos y Características y más... - Arquitasa*, https://arquitasa.com/predio/, Accessed: 2023-11-19, 2022.

[12] T. W. Merrill, «Economics of Leasing,» *Journal of Legal Analysis*, vol. 12, págs. 221-272, 2020. dirección: https://doi.org/10.1093/jla/laaa003.

[13] *Rent | Definition, Impact & Benefits | Britannica Money*, https://www.britannica.com/money/topic/rent-economics, Accessed: 2023-11-19, 2022.

[14] *What is Exploratory Data Analysis? | IBM*, https://www.ibm.com/topics/exploratory-data-analysis, Accessed: 2023-11-19, 2022.

[15] R. Behar y M. Yepes, *Estadística: Un enfoque descriptivo*. Ed. Feriva, 1995. dirección: https://www.editdiazdesantos.com/wwwdat/pdf/9788479789923.pdf.

[16] M. Komorowski, D. C. Marshall, J. D. Salciccioli e Y. Crutain, «Exploratory Data Analysis,» *Secondary Analysis of Electronic Health Records*, págs. 185-203, 2016. dirección: https://doi.org/10.1007/978-3-319-43742-2_15.

[17] *Types of Variables in Research & Statistics | Examples - Scribbr*, https://www.scribbr.com/methodology/types-of-variables/, Accessed: 2023-11-19, 2022.

[18] J. A. Nelder y R. W. Wedderburn, «Generalized Linear Models,» *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, n.º 3, págs. 370-384, 1972. DOI: 10.2307/2344614.

[19] P. Mccullagh y J. A. Nelder, *Generalized Linear Models* (Monographs on Statistics and Applied Probability), 2nd. London: Chapman y Hall/CRC, 1989.

[20] L. Anselin y A. K. Bera, «Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics,» en *Statistics Textbooks and Monographs*, vol. 155, CRC Press, 1998, págs. 237-290.

[21] P. A. P. Moran, «Notes on Continuous Stochastic Phenomena,» *Biometrika*, vol. 37, n.º 1/2, págs. 17-23, 1950. DOI: 10.2307/2332142. dirección: https://doi.org/10.2307/2332142.

[22] P. Moraga, *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny* (Chapman & Hall/CRC Biostatistics Series). Chapman y Hall/CRC, 2019, ISBN: 9780367357955. dirección: https://doi.org/10.1201/9780429341823.

[23] F. Ploeckl, «Spatial Modeling,» en *Handbook of Cliometrics*, C. Diebolt y M. Haupert, eds., Springer, 2019. dirección: https://link.springer.com/referenceworkentry/10.1007/978-3-030-00181-0_56.

[24] C. Gaetan y X. Guyon, *Spatial Statistics and Modeling*. Springer, 2009. dirección: https://link.springer.com/book/10.1007/978-0-387-92257-7.

[25] M. Dumelle, M. Higham y J. M. Ver Hoef, «spmodel: Spatial statistical modeling and prediction in R,» *Journal of Statistical Software*, vol. 98, n.º 13, págs. 1-31, 2021. dirección: https://www.jstatsoft.org/article/view/v098i13.

[26] L. Anselin, *Spatial Econometrics: Methods and Models*. Springer, 1988. dirección: https://link.springer.com/book/10.1007/978-94-015-7799-1.

[27] L. Anselin, «Under the hood: Issues in the specification and interpretation of spatial regression models,» *Agricultural Economics*, vol. 27, págs. 247-267, 2002. dirección: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1574-0862.2002.tb00120.x.

[28] A. Burkov. «The Hundred-Page Machine Learning Book.» (2019), dirección: https://millengustavo.github.io/hundred-page-ml/.

[29] T. Hastie, R. Tibshirani y J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics), 2.ª ed. New York, NY: Springer, 2009, vol. 2, ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

[30] L. Breiman, «Random Forests,» *Machine Learning*, vol. 45, n.º 1, págs. 5-32, 2001. DOI: 10.1023/A:1010933404324.

[31] L. Breiman, «Bagging Predictors,» *Machine Learning*, vol. 24, n.º 2, págs. 123-140, 1996. DOI: 10.1007/BF00058655.

[32] C. Cortes y V. Vapnik, «Support-vector networks,» *Machine Learning*, vol. 20, n.º 3, págs. 273-297, 1995. DOI: 10.1007/BF00994018.

[33] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Information Science and Statistics), 2.ª ed. New York: Springer Science & Business Media, 1999, ISBN: 978-0-387-98780-4. DOI: 10.1007/978-1-4757-3264-1.

[34] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998, ISBN: 978-0-471-03003-4.

[35] D. Stojanova, M. Ceci, A. Appice, D. Malerba y S. Dzeroski, «Global and local spatial autocorrelation in predictive clustering trees,» en *Discovery Science: 14th International Conference, DS 2011, Espoo, Finland, October 5-7, 2011. Proceedings 14*, Springer, 2011, págs. 307-322.

[36] S. Quiñones, A. Goyal y Z. U. Ahmed, «Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA,» *Scientific Reports*, vol. 11, n.º 1, pág. 6955, 2021.

[37] Y. Luo, J. Yan y S. McClure, «Distribution of the environmental and socioeconomic risk factors on COVID-19 death rate across continental USA: a spatial nonlinear analysis,» *Environmental Science and Pollution Research*, vol. 28, págs. 6587-6599, 2021.

[38] C.-H. Lee, R. Greiner y M. Schmidt, «Support vector random fields for spatial classification,» en *Knowledge Discovery in Databases: PKDD 2005: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, Porto, Portugal, October 3-7, 2005. Proceedings 9*, Springer, 2005, págs. 121-132.

[39] P. Schratz, M. Becker, M. Lang y A. Brenning, «mlr3spatiotempcv: Spatiotemporal Resampling Methods for Machine Learning in R,» *arXiv preprint arXiv:2110.12674*, 2021, Available at https://arxiv.org/abs/2110.12674. dirección: https://arxiv.org/abs/2110.12674.

[40] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji y T. Nauss, «Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation,» *Environmental Modelling & Software*, vol. 101, págs. 1-9, 2018. DOI: 10.1016/j.envsoft.2017.12.001. dirección: https://doi.org/10.1016/j.envsoft.2017.12.001.

[41] H. Meyer, C. Reudenbach, S. Wöllauer y T. Nauss, «Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction,» *Ecological Modelling*, vol. 411, pág. 108 815, 2019. DOI: 10.1016/j.ecolmodel.2019.108815. dirección: https://doi.org/10.1016/j.ecolmodel.2019.108815.

[42] L. Fahrmeir, T. Kneib, S. Lang y B. D. Marx, *Regression: Models, Methods and Applications*. Springer, 2021. dirección: https://link.springer.com/book/10.1007/978-3-662-63882-8.

[43] D. C. Montgomery, E. A. Peck y G. G. Vining, *Introduction to Linear Regression Analysis*. Wiley, 2021. dirección: https://books.google.com/books/about/Introduction_to_Linear_Regression_Analys.html?id=tCIgEAAAQBAJ.

[44] D. A. Griffith, *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization* (Advances in Spatial Science). Springer, 2003, ISBN: 978-3-540-00820-0. DOI: 10.1007/978-3-540-24806-4.

[45] M. P. Bohorquez, *Estadística Espacial y Espacio-Temporal para Campos Aleatorios Escalares y Funcionales*. Bogotá, Colombia: Universidad Nacional de Colombia, 2020.

[46] B. Nikparvar y J.-C. Thill, «Machine learning of spatial data,» *ISPRS International Journal of Geo-Information*, vol. 10, n.º 9, pág. 600, 2021.

[47] P. Melin, J. C. Monica, D. Sanchez y O. Castillo, «Analysis of spatial spread relationships of coronavirus (COVID-19) pandemic in the world using self organizing maps,» *Chaos, Solitons & Fractals*, vol. 138, pág. 109 917, 2020.

[48] P. Agarwal y A. Skupin, *Self-organising maps: applications in geographic information science*. John Wiley & Sons, 2008.

[49] H. Shafizadeh-Moghadam, J. Hagenauer, M. Farajzadeh y M. Helbich, «Performance analysis of radial basis function networks and multi-layer perceptron networks in modeling urban change: a case study,» *International Journal of Geographical Information Science*, vol. 29, n.º 4, págs. 606-623, 2015.

[50] A. S. Fotheringham, C. Brunsdon y M. Charlton, *Quantitative Geography: Perspectives on Spatial Data Analysis*. SAGE Publications, 2000. dirección: https://us.sagepub.com/en-us/nam/quantitative-geography/book205876.

[51] I. G. A. C. (IGAC), *Geoportal*, 2024. dirección: https://geoportal.igac.gov.co/.

[52] U. de Planificación Rural Agropecuaria (UPRA), *Cálculo de la Unidad Agrícola Familiar Paso a Paso*, 2024. dirección: https://upra.gov.co/es-co/Publicaciones/20221220_Cartilla_UAF.pdf.

[53] D. A. N. de Estadística (DANE), *Pobreza Multidimensional*, 2024. dirección: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-multidimensional.

A. Anexos

Tabla 11: Clasificación de Variables por Categoría

Categoría	**Variable**
PRODUCTIVIDAD	PASTOS_CON_PROBABLE_ACTVIDAD_GANERA
PRODUCTIVIDAD	MEDIA_PONDERADA
PRODUCTIVIDAD	DIVERSIDAD_DE_CULTIVOS
PRODUCTIVIDAD	AREA_DISPONIBLE
PRODUCTIVIDAD	CLASIFICACION_IGAC
PRODUCTIVIDAD	AREAPREDIOHA
PRODUCTIVIDAD	precio_entorno
COMUNIDADES	CLASIFI_AC
COMUNIDADES	DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM
COMUNIDADES	CERCANIA_MINIMA_COMUNIDADES
COMUNIDADES	DIST_ZRC_KM
COMUNIDADES	DIST_TITULOSCOLECTIVOS_KM
COMUNIDADES	DIST_RESGUARDOS_KM
APARTADO AMBIENTAL	ECOSISTEMAS
APARTADO AMBIENTAL	TRASLAPE_RUNAP
APARTADO AMBIENTAL	AMENAZA_AMBIENTAL
APARTADO AMBIENTAL	IPM_SECCION_RURAL
APARTADO AMBIENTAL	DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM
APARTADO AMBIENTAL	TRASLAPE_ZRF
INFRAESTRUCTURA Y APARTADO SOCIAL	AREA_MINERIA
INFRAESTRUCTURA Y APARTADO SOCIAL	TRASLAPE_EXTRACCION_COMBUSTIBLE
INFRAESTRUCTURA Y APARTADO SOCIAL	DISTANCIA_A_VIAS_PRINCIPALES
INFRAESTRUCTURA Y APARTADO SOCIAL	DIST_VIAS_PRIMARIAS_KM
INFRAESTRUCTURA Y APARTADO SOCIAL	DIST_VIAS_SECUNDARIAS_KM
INFRAESTRUCTURA Y APARTADO SOCIAL	DIST_VIAS_TERCIARIAS_KM

Tabla 12: Variables con Datos Atípicos

Variable	**Porcentaje de Datos Atípicos**
ESTIMATIVO	11.73
AREAPREDIOHA	12.65
PASTOS_CON_PROBABLE_ACTVIDAD_GANERA	16.76
MEDIA_PONDERADA	0.00
ECOSISTEMAS	5.81
TRASLAPE_ZRF	7.99
TRASLAPE_VOCACIONES	12.39
DISTANCIA_MINIMA_FIGURAS_ORGANIZATIVAS_KM	1.82
DIST_VIAS_PRIMARIAS_KM	9.75
DIST_VIAS_SECUNDARIAS_KM	5.63
DIST_VIAS_TERCIARIAS_KM	4.36
PRECIO_ENTORNO	9.11