

Metodología para obtención y análisis de datos inmobiliarios usando fuentes alternativas: estudio de caso en tres ciudades intermedias de Colombia

Methodology for the Collection and Analysis of Real Estate Data Using Alternative Sources: Case Study in Three Medium-Sized Cities of Colombia

Andrés E. Rosso-Mateus¹, Yeimy. M. Montilla-Montilla¹, Sonia. C. Garzón-Martínez¹

¹ Instituto Geográfico Agustín Codazzi, Centro de Investigación y Desarrollo en Información Geográfica - CIAF

Received: 01st-February-2021. Modified: 16th-November-2022. Accepted: 18th-March-2022

Resumen

Contexto: La política pública de Catastro Multipropósito necesita consolidar información inmobiliaria de diferentes fuentes para su análisis, tales como ofertas, transacciones y costos de construcción, entre otros. Las páginas web inmobiliarias forman parte de estas fuentes de información, aunque no han sido incluidas en el análisis comercial. Considerando lo anterior, es necesario revisar una metodología que permita acceder de forma óptima a estas plataformas web y facilite el análisis de las variables que allí se proveen, que son determinantes para el valor comercial de un inmueble. Se realiza un caso de estudio en tres ciudades colombianas: Fusagasugá, Manizales y Villavicencio.

Método: El método se desarrolla en dos etapas (i) web scraping, que permite obtener los enlaces de la información de páginas web inmobiliarias y descargar sus datos, y (ii) el análisis de datos inmobiliarios mediante el desarrollo de un flujo de trabajo que inicia con la exploración y la limpieza de los datos, continúa con el pre-modelado y finaliza con el modelado de las variables de interés en la determinación del valor de los bienes inmuebles usando técnicas de machine learning.

Resultados: A partir de la aplicación de técnicas de machine learning, fue posible automatizar la recolección, la limpieza, el almacenamiento y el análisis de datos inmobiliarios provenientes de plataformas web, así como delinear dos modelos (Ridge Regression y Random Forest) que, de acuerdo, con su error porcentual medio absoluto (0,34 y 0,35 respectivamente), permiten predecir el valor comercial de un inmueble considerando variables explicativas internas y externas.

Conclusiones: Obtener y analizar los datos inmobiliarios de fuentes alternativas como las plataformas web a través de desarrollos tecnológicos contribuye significativamente a atender la alta demanda de información del catastro del país. No obstante, es necesario ampliar el suministro de esta información a los ámbitos rurales, que cuentan con menos acceso y disponibilidad de la misma.

Idioma: Español

Palabras clave: Catastro Multipropósito, dinámica inmobiliaria, mercado inmobiliario, valor comercial, web scraping

Open access



© The authors; licensee: Revista INGENIERÍA. ISSN 0121-750X, E-ISSN 2344-8393. Cite this paper as: Author, F., Author, J., Author, S.: The Title of the Paper. INGENIERÍA, Vol. XX, Num. XX, 2015 pp:pp.
doi:10.14483/udistrital.jour.reving.23448393.17952

Abstract

Context: The Multipurpose Cadastre public policy needs to consolidate real estate information from different sources for analysis, such as offers, transactions, and construction costs, among others. Real estate websites are part of these sources of information, although they have not yet been included in commercial analysis. In light of the above, it is necessary to review a methodology that allows optimal access to these web platforms and facilitates the analysis of the variables provided therein, which are crucial to a property's commercial value. A study case was carried out in three Colombian cities: Fusagasugá, Manizales, and Villavicencio.

Method: The method is implemented in two stages: (i) web scraping, which allows obtaining the information links from real estate web pages and downloading their data, and (ii) analyzing real estate data by developing a workflow that starts with data exploration and cleaning, continues with pre-modeling, and ends by modeling the crucial variables in the determination of real estate value using machine learning techniques.

Results: By applying machine learning techniques, it was possible to automate the collection, cleaning, storage, and analysis of real estate data from web platforms, as well as to outline two models (Ridge Regression and Random Forest), which, according to their mean absolute percentage error (0,34 and 0,35, respectively), allow predicting the commercial value of a property while considering internal and external explanatory variables.

Conclusions: Obtaining and analyzing real estate data from alternative sources such as web platforms through machine learning techniques contributes significantly to addressing the high information demand of the country's cadastre. However, it is necessary to expand the supply of this information to rural areas, which have less access and availability to it.

Keywords: Multipurpose Cadastre, real estate dynamics, Real Estate Market, Commercial Value, web scraping

Language: Spanish.

1 Introducción

Las páginas web tanto en ámbitos científicos, políticos y de gobierno son hoy por hoy una fuente de datos rica e interesante para el desarrollo del análisis de grandes volúmenes de datos o “análisis de Big Data” [1]. El Web Scraping se dispone como una técnica de extracción automática de información textual de páginas web, donde se crea en primera instancia un árbol DOM (por sus siglas en inglés Document Object Model), para luego acceder a los datos necesarios por medio de este árbol. Otro método de Web Scraping, conocido como UzunExt, optimiza la labor de extracción de los datos mediante métodos de cadena que consiste en la búsqueda de un patrón dado, que permite luego el cálculo del número de elementos HTML de cierre para este patrón y finalmente la extracción del contenido del mismo [2].

Son varios los trabajos que proponen aplicaciones exitosas con el uso de Web Scraping, donde en el campo inmobiliario se busca extraer de sitios web información sobre el mercado de la vivienda para la generación de indicadores más precisos como el número de nuevas ofertas publicadas o la fluctuación de los precios a lo largo del tiempo para las ofertas existentes, así como avanzar con técnicas de aprendizaje de máquina para comparar los precios de venta de la oferta inmobiliaria versus los precios de negociación correspondientes y registrados por bases de datos notariales [3]; otros autores investigan sobre los precios de los alimentos, reconociendo en la técnica de Web Scraping un método efectivo para recopilar datos personalizados y de alta frecuencia en tiempo

real, siempre y cuando se mantengan los registros de precios de alimentos en línea [4]; por otro lado, a la hora de planificar grandes infraestructuras es necesario comprender el impacto potencial de los desarrollos de terceros que pueden afectar las obras previstas, donde con esta información, disponible en el dominio público, se usa la técnica de Web Scraping basada en un algoritmo de árbol reforzado, que luego analiza los datos con algoritmos de aprendizaje de máquina [5]; finalmente, un trabajo que aprovecha los datos de usuarios de redes sociales como Facebook, Instagram y Twitter usa Web Scraping para buscar, combinar y presentar de una mejor manera la información de acuerdo con las preferencias de usuarios y eliminando aquella información redundante [6]. En suma, son múltiples las aplicaciones de los métodos de Web Scraping, no encontrando limitaciones en el ámbito tecnológico sino más bien en aquellos asociados con el acceso legal a los datos que se disponen de forma libre por los distintos proveedores de contenidos en la web [7].

Considerando, los casos de éxito en la aplicación de Web Scraping en diferentes campos e identificando para Colombia, en el marco de la reciente política pública de Catastro Multipropósito [8], la necesidad de recopilar la información del mercado inmobiliario como ofertas, transacciones, costos de construcción, entre otras, que alimenten los Observatorios Inmobiliarios Catastrales del país [9], el presente proyecto propone desarrollar una metodología para la obtención, la depuración, el almacenamiento y el análisis de datos provenientes de las plataformas web inmobiliarias, considerando como variable objetivo los valores de ofertas inmobiliarias y las correspondientes variables explicativas derivadas de un ejercicio exploratorio, basado en algoritmos de aprendizaje de máquina, que dé idea del potencial de la información disponible en línea junto con el proceso de valuación masiva del suelo.

Se reconoce, por tanto, en las páginas web inmobiliarias una valiosa disposición de datos que caracteriza las ofertas de venta y alquiler de bienes inmuebles, documentando en parte la dinámica comercial de este sector y suministrando variables propias y del entorno asociadas al valor comercial de los bienes [10]. El Banco Interamericano de Desarrollo (BID) propone aplicar métodos avanzados de levantamiento de datos, como Web Scraping, para permitir el uso de la información inmobiliaria disponible en línea y servir de insumo para los observatorios del mercado inmobiliario (OMI), los cuales podrán a su vez elaborar modelos de valuación masiva [10]. De acuerdo con el anterior planteamiento se hace necesario explorar el potencial de la información proveniente de las plataformas web, las cuales son indicio de la dinámica inmobiliaria, para generar información actualizada del suelo en forma masiva. En este proceso un agente programado o software que realiza tareas repetitivas, predefinidas y automatizadas se encarga de explorar varios sitios web con el fin de obtener, clasificar, organizar y formatear la información.

El uso de técnicas de Machine Learning como soporte de tareas manuales es cada vez más frecuente en diferentes dominios. En el caso particular de establecer los precios de los inmuebles, estas herramientas pueden ser de gran ayuda al generalizar patrones comunes que sirven como método de estimación cercano. Para lograr identificar estos patrones es necesario contar con muchos ejemplos que permitan al algoritmo de Machine Learning generalizar basado en la evidencia; estos datos pueden ser muy costosos de obtener, razón por la cual los métodos de extracción automática, también conocidos como métodos de Web Scraping tienen un rol importante [11].

2 Materiales y Métodos

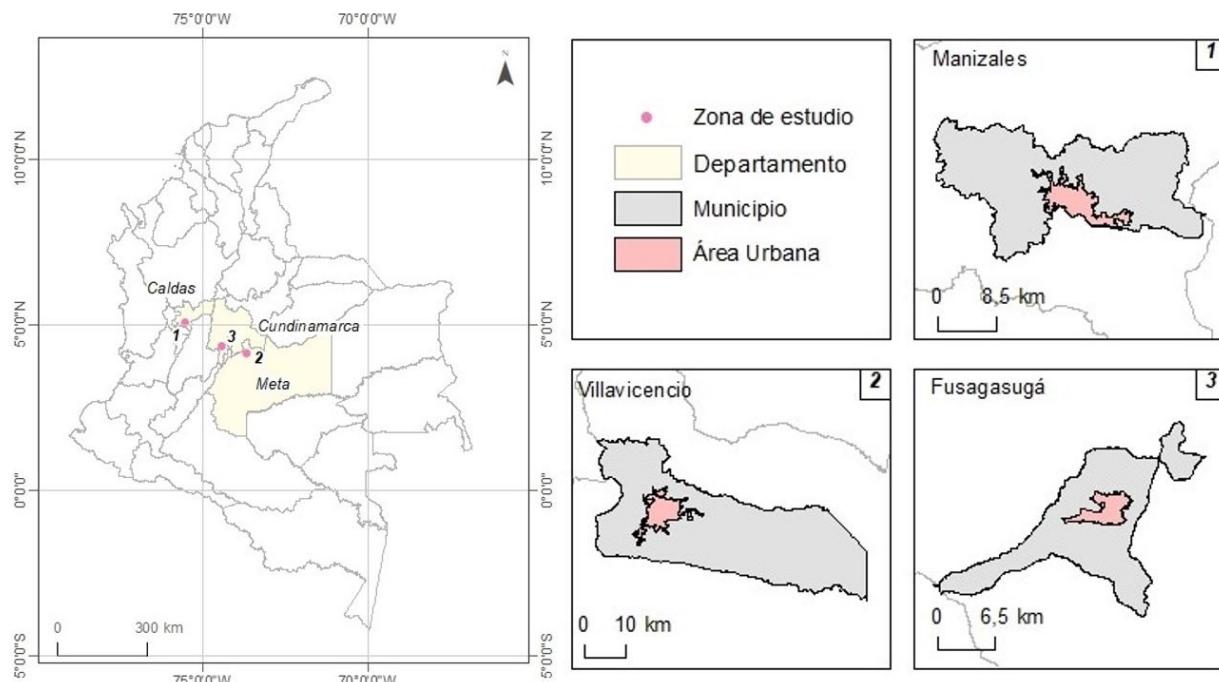
2.1 Área de estudio

El proyecto se desarrolló en tres áreas urbanas del país, como se observa en la figura 1:

- Manizales, Departamento de Caldas
- Villavicencio, Departamento del Meta
- Fusagasugá, Departamento de Cundinamarca

Manizales es la capital del departamento de Caldas, se ubica en el centro occidente de Colombia, en la Cordillera Central y en las inmediaciones del Nevado del Ruiz a una altura de 2.150 msnm aproximadamente [12]. De acuerdo con las cifras del censo nacional de población y vivienda del 2018, Manizales cuenta con una población de 400.436 habitantes [13]. En el ámbito catastral, conforme a la consulta realizada en el registro 1 de la base de datos catastral con cobertura nacional, el municipio cuenta con 141.239 predios, de los cuales 131.820 pertenecen al área urbana y 9.419 pertenecen al área rural [14].

Villavicencio es la capital del departamento del Meta, se ubica a una altura aproximada de 467 msnm en el piedemonte de la Cordillera Oriental y al noroccidente del departamento, a una distancia de 86 km de la ciudad de Bogotá [15]. De acuerdo con las cifras del censo nacional de población y vivienda del 2018, Villavicencio cuenta con una población de 451.212 habitantes [13]. En el ámbito catastral conforme a la consulta realizada en el registro 1 de la base de datos catastral con cobertura nacional, el municipio cuenta con 199.239 predios, de los cuales 142.977 pertenecen al área urbana y 54.515 pertenecen al área rural y 1.747 pertenecen a los centros poblados [14].



Fusagasugá es la capital de la provincia del Sumapaz en el departamento de Cundinamarca, se ubica a 59 km al suroccidente de la ciudad de Bogotá [16]. De acuerdo con las cifras del censo nacional de población y vivienda del 2018, Fusagasugá cuenta con una población de 134.658 habitantes [13]. En el ámbito catastral, conforme a la consulta realizada en el registro 1 de la base de datos catastral con cobertura nacional, el municipio cuenta con 71.318 predios, de los cuales 59.465 pertenecen al área urbana y 11.664 pertenecen al área rural y los 189 restantes pertenecen a los centros poblados [14].

2.2 Flujo metodológico

La propuesta metodológica, como se observa en la figura 2, se realiza en dos etapas organizadas de la siguiente forma: (i) Web Scraping; y (ii) análisis de datos inmobiliarios, usando métodos de Machine Learning. Para el desarrollo del componente de software de la primera etapa se sigue el modelo iterativo e incremental denominado Proceso Racional Unificado (RUP, Rational Unified Process por sus siglas en inglés) [17], caracterizado por la realización en paralelo de todas las etapas necesarias para la elaboración del software. Para la segunda etapa, donde se realiza el análisis de datos inmobiliarios se usan métodos de Machine Learning en el marco de la metodología CRISP-DM (por sus siglas en inglés, Cross-Industry Standard Process for Data Mining), la cual es empleada para llevar a cabo proyectos de minería de datos [18].

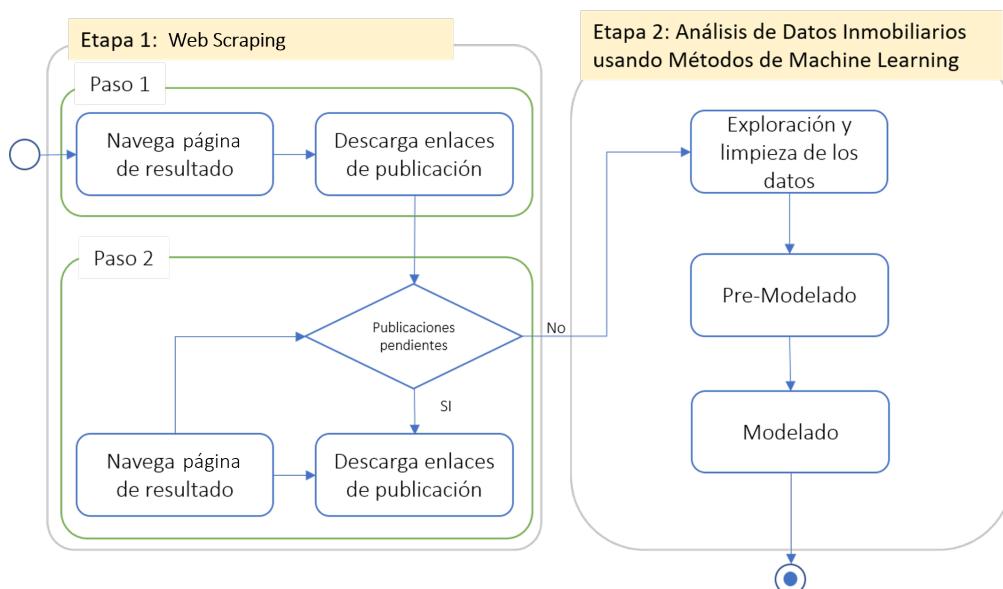


Figura 2: Flujo metodológico
Fuente: elaboración propia

2.3 Web Scraping

En la etapa 1 se realizan dos pasos [19] [20]: en el primero, el usuario hace manualmente la consulta en la plataforma web y el scraper se encarga de descargar uno por uno los enlaces a las páginas detalladas del anuncio conforme al resultado; en el segundo, con el listado de enlaces se realiza,

para cada una de las publicaciones, la descarga de la página html con la información detallada de la oferta; posteriormente, se extraen las variables de interés, enunciadas en la tabla II y se almacenan de forma estructurada en un archivo plano, el cual se convierte en uno de los insumos para la segunda etapa.

La implementación del scraper se desarrolla con Javascript y Python. El script generado se carga en la página web usando las extensiones Witchcraft: JS/CSS injector y Web Server for Chrome. Este script se encarga de navegar por las páginas resultado de la búsqueda, de la misma forma en que lo haría un usuario. Por otra parte, este también extrae los enlaces de cada una de las publicaciones, para luego navegar por estas usando NodeJS y la interfaz remota de Chrome. Para cada publicación descarga la información usando el script antes mencionado. Finalmente, los resultados se limpian y estructuran de forma tabular usando Python.

2.3.1 Datos de descarga

En los últimos años, se ha fortalecido la comercialización de inmuebles mediante el uso de portales web, convirtiéndose este en el principal medio para la publicación de la oferta inmobiliaria, con el consecuente relegamiento de la publicación mediante medios impresos. Lo anterior, incide favorablemente en la cantidad de información que se encuentra disponible en la red para alimentar un proceso de Web Scraping. Existe una gran cantidad de portales dedicados a la publicación del negocio inmobiliario entre los que se destacan los descritos en la tabla I.

Nombre	Página web
Metrocuadrado	www.metrocuadrado.com
Fincaraiz	www.fincaraiz.com.co

Tabla I: Plataformas Web

Fuente: elaboración propia

Las características más importantes que impulsan el valor de una vivienda tipo casa son el área y la localización; sin embargo, hay otras variables que pueden determinarlo tales como: número de habitaciones, de estacionamientos, o de baños, proximidad a algún sitio de interés por la prestación de algún servicio, entre otros, pero sin lugar a duda, la mayor instancia del valor es la demanda [21]. Para el desarrollo de esta investigación se tomó como fuente de datos la plataforma web Fincaraiz, de la cual se seleccionaron las variables que se disponen en la tabla II. Se consideraron de este sitio las variables de latitud y longitud para tener la posibilidad de definir patrones espaciales y de forma particular; se aclara que la variable precio, reportada en la plataforma, corresponde al valor de la oferta más no al valor pactado dentro de la negociación del inmueble, situación que se encontraría para cualquier plataforma web inmobiliaria. Por otro lado, para identificar los factores asociados con el entorno, que pueden influir en el valor de los predios, se tuvieron en cuenta las plataformas de Open Street Maps y Google Places, de donde se trajeron los sitios de interés.

2.4 Análisis de datos inmobiliarios usando métodos de Machine Learning

Para el desarrollo del proyecto se usó la metodología de desarrollo para proyecto de Machine Learning CRISP-DM [22]. Esta metodología proporciona un enfoque estructurado para la planificación del proyecto, siendo además una metodología robusta y bien probada.

Clasificación	Datos de la oferta	Ubicación del inmueble	Descripción del inmueble
Variable	Fecha Publicación ID publicación Precio Tipo de oferta Tipo de propiedad URL	Ciudad Localización Longitud Latitud	Área construida Número de baños Número de habitaciones Estacionamientos Estrato Administración Antigüedad

Tabla II: Variables de interés

Fuente: elaboración propia

Este modelo es una secuencia idealizada de etapas de un proyecto, pero a menudo las tareas pueden realizarse en un orden diferente o podría ser necesario volver a tareas anteriores y repetir ciertas acciones.

Las fases más relevantes o aquellas que requieren mayor trabajo son: (i) exploración y limpieza de los datos, (ii) pre-modelado, y (iii) modelado, que se aprecian en la figura 2.

2.4.1 Exploración y limpieza de los datos

En esta fase se realiza la exploración general de los datos, para luego avanzar con la limpieza de estos; posteriormente con los datos depurados se procede a la exploración geográfica y la exploración de la información secundaria.

Exploración de variables: corresponde a la revisión general de la calidad asociada a los datos que son recolectados mediante el Web Scraping, por el uso de tablas y gráficas que permiten:

- Exploración de datos faltantes en las publicaciones con el fin de identificar datos nulos en las variables de interés.
- Exploración de variables de texto, donde se revisa particularmente el campo de descripción del inmueble, el cual al ser un texto libre alimentado por el usuario puede arrojar información sobre aquellas palabras claves que con mayor frecuencia se usan para definir la oferta.
- Exploración de variables discretas, donde para cada variable discreta del conjunto de datos se realiza un diagrama de barras con el fin de identificar la frecuencia de los valores que estas toman.

Limpieza de los datos: este proceso consiste en la preparación de los datos para el pre-modelado y el modelado. Las principales tareas que se realizan son:

- Construcción de validadores, donde se plantean dos tipos: a. validador “builder” que contiene las coordenadas extremas de la ciudad objetivo con el propósito de corroborar la localización (latitud, longitud) de las publicaciones descargadas, b. validador “condition” que verifica el cumplimiento de ciertos criterios por parte de los registros, como por ej. la ciudad, la fecha, el tipo de oferta y el tipo de inmueble de interés.

- Imputación de valores vacíos por un valor lógico. Pueden existir varias publicaciones que no contienen algunos registros en ciertos campos, pero la ausencia no significa que se desconozca su valor. Para las variables "valor de la administración" y "número de garajes" se asume que el valor es cero (0) cuando existe ausencia del dato. Para la variable antigüedad en caso de ausencia del dato se asigna el valor de desconocido.
- Cálculo de la variable precio por m^2 "price m^2 " la cual se deriva de la división del precio de la oferta sobre el área del predio.

Finalmente se obtiene el número de publicaciones que cumplen con los validadores definidos y se consideran adecuadas para ser trabajadas en el siguiente paso.

Exploración geográfica: considerando el componente espacial del conjunto de datos, se realiza la construcción de mapas que facilitan la identificación de zonas en la ciudad seleccionada, que son de especial importancia de acuerdo con la variable que se visualiza.

- Mapa de ubicación de las publicaciones: que posibilita la visualización geográfica del conjunto de ofertas depuradas para la ciudad seleccionada.
- Mapa de cantidad de publicaciones por zona: que presenta las publicaciones agrupadas, demarcando zonas con etiquetas que denotan el número de publicaciones en ese conjunto.
- Mapa de concentración de publicaciones: que facilita la identificación de los sitios de la ciudad donde se presenta la mayor densidad de publicaciones.
- Mapa de concentración de precios: que ayuda con la identificación de los sitios que en la ciudad tienen la oferta de precios más altos. Es precisamente en estos sitios donde se realiza la exploración de las variables de entorno suministradas por Google Places u Open Street Maps, derivando en consecuencia los siguientes mapas.
- Mapa de ubicación de sitios de interés o amenities: que muestra la ubicación de los sitios de interés en la ciudad seleccionada.
- Mapa de concentración de sitios de interés: que presenta las zonas de mayor concentración de sitios de interés en la ciudad seleccionada.

Exploración variables explicativas vs variable objetivo: este proceso tiene como propósito contrastar diferentes variables del conjunto de datos como variables explicativas con respecto al valor del inmueble. Se hace uso de histogramas y diagramas de caja para abordar la exploración estadística de las variables, que faciliten la identificación de relaciones entre el precio de los bienes inmuebles y algunas características como: número de garajes, número de habitaciones, número de baños, años de antigüedad y estrato.

2.4.2 Pre-modelado

En esta fase se identifican las características internas y externas con más relevancia en la explicación del valor asociado al precio de la oferta. A continuación las técnicas usadas dependiendo el tipo de característica:

Identificación de las características externas: se realiza la cuantificación de los sitios de interés o amenities en un radio de un kilómetro en torno a la posición geográfica registrada de cada una de las publicaciones en la ciudad de análisis. Posteriormente, se tabula y analiza la tabla de frecuencias generada para estos sitios de interés.

Relevancia de las características internas y externas: para identificar las variables explicativas que influyen mayoritariamente en el valor de la oferta se consideran tres métodos: (i) selección univariada de características; (ii) eliminación recursiva de características; (iii) selección de características basada en árboles de decisión.

- Selección univariada de características: método estadístico mediante el cual se genera una regresión lineal univariada que prueba de forma individual el efecto de cada una de las variables explicativas sobre la variable objetivo. Se aplican dos estadísticos que permiten valorar la bondad de ajuste de los datos al modelo de regresión lineal simple:
 - Coeficiente de correlación lineal simple, que mide el grado de asociación lineal entre dos variables.
 - Análisis de varianza, que permite valorar hasta qué punto es adecuado el modelo de regresión lineal para estimar los valores de la variable dependiente.
- Eliminación recursiva de características: permite identificar cuáles variables dentro del conjunto de datos son de mayor relevancia para explicar la variable objetivo, la cual dentro del presente estudio corresponde con el valor de la oferta. La importancia de cada una de las variables se determina mediante una regresión lineal.
- Selección de características basada en árboles de decisión: se implementa un metaestimador que se ajusta a varios árboles de decisión aleatorios (también conocidos como árboles extra) en varias submuestras del conjunto de datos y usa promedios para mejorar la precisión predictiva y el control de sobreajuste.

2.4.3 Modelado

El objetivo de esta fase es tener un algoritmo que pueda predecir el precio de un inmueble dadas sus características, entre estas se tienen: ubicación geográfica, número de habitaciones, área, antigüedad. En el modelado se manejan tres etapas, que son: (i) entrenamiento del modelo, (ii) predicción y (iii) evaluación del desempeño del modelo.

Entrenamiento del modelo: el proceso de entrenamiento de un modelo de aprendizaje de máquina (ML, Machine Learning) consiste en proporcionar a un algoritmo ejemplos de entrenamiento para identificar patrones en los datos y posteriormente encontrar los hiperparámetros de forma automática usando cross-validation; para el presente caso, las publicaciones inmobiliarias son la fuente de datos para el entrenamiento, donde se cuenta con el precio del inmueble (Y) junto a las características internas y externas del inmueble (X). Listo el entrenamiento, se genera un modelo de ML que captura dichos patrones y por medio de una función de mapeo ($X \rightarrow Y$) estos se trasforman en un valor estimado (\hat{Y} = precio estimado).

En la etapa de entrenamiento es necesario que las características estén en la misma escala, por tanto se requiere usar técnicas de escalamiento o normalización como un paso previo al entrenamiento del modelo. Una vez encontrado el mejor modelo, se validan sus métricas contra los objetivos de modelado y si este es lo suficientemente efectivo se reserva para la fase de pruebas, de lo contrario, se debe iniciar el proceso de nuevo.

Modelo predictivo: en la fase de pruebas o predicción se toman las características del inmueble del que no se conoce el valor (Y), para ser escaladas y luego pasar por el modelo entrenado obtenido en el paso anterior. Se realizan por tanto varios experimentos para encontrar la combinación de variables y el mejor modelo que, basado en las características del inmueble, así como las características externas, logre predecir de mejor manera el precio asociado. Los parámetros de los modelos se ajustan por medio de la técnica "grid search", capaz de identificar los mejores hiperparámetros del modelo. Se prueban varios algoritmos de regresión, entre los cuales se encuentran:

- **Regresión lineal:** la regresión lineal (ajuste lineal) es un modelo matemático que aproxima la relación de dependencia entre una variable dependiente Y y las variables independientes X , más un término aleatorio ϵ .

El producto de la regresión es la estimación de una función de regresión dependiente de X . Con base en esta función estimada se pueden hacer predicciones sobre eventos futuros, ya que las variables están correlacionadas por medio de esta función.

- **Modelo de regresión Ridge:** tres de las limitaciones que aparecen en la práctica al tratar de emplear el modelo de regresión lineal son:

- Se ven perjudicados por la incorporación de predictores correlacionados.
- No realizan selección de predictores. Todos los predictores se incorporan en el modelo, aunque no aporten información relevante. Esto suele complicar la interpretación del modelo y reducir su capacidad predictiva.
- No pueden ajustarse cuando el número de predictores es superior al número de observaciones.

Una forma de atenuar el impacto de estos problemas es utilizar la estrategia de regularización Ridge, que fuerza a que los coeficientes del modelo tiendan a 0, minimizando así el riesgo de overfitting. Esto reduce la varianza, atenúa el efecto de la correlación entre predictores y disminuye la influencia en el modelo de los predictores menos relevantes. La regularización Ridge penaliza la suma de los coeficientes elevados al cuadrado. A esta penalización se le conoce como L2 y tiene el efecto de reducir de forma proporcional el valor de todos los coeficientes del modelo sin que estos lleguen a 0. El grado de penalización está controlado por el hiperparámetro λ . Cuando $\lambda = 0$, la penalización es nula y el resultado es equivalente al de un modelo lineal por mínimos cuadrados ordinarios.

- **Modelo Gradient Boosting:** Gradient Boosting o potenciación del gradiente es una técnica de aprendizaje automático utilizada para el análisis de la regresión y para problemas de clasificación estadística, la cual produce un modelo predictivo en forma de un conjunto de modelos de predicción débiles. Construye el modelo de forma escalonada como lo hacen otros

métodos de boosting, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable.

- **Modelo Random Forest:** en el método de Random Forest se entrenan diferentes modelos del tipo árbol de decisión. Se usan sobre diferentes particiones de los datos y la predicción se hace usando todos los modelos ajustados, los cuales al ser combinados tienen un rendimiento superior que si solo se usará un árbol de decisión.

Evaluación del desempeño del modelo: para comparar el desempeño de los modelos se usa la métrica de Error Porcentual Medio Absoluto (MAPE), la cual entrega la desviación en términos porcentuales y no en unidades como las otras medidas. Esta se interpreta como el promedio del error absoluto o diferencia entre el valor real y el pronóstico, expresado como un porcentaje de los valores reales. En la ecuación 1 se muestra la expresión matemática correspondiente.

$$MAPE = \frac{\sum_{i=1}^n 100 |Real_i - Pronóstico_i|}{\sum_{i=1}^n Real_i} \quad (1)$$

3 Resultados

La presentación de resultados se muestra organizada de acuerdo con las dos etapas de desarrollo. La sección 3.1 indica los resultados obtenidos en la etapa de Web Scraping y la sección 3.2 describe los resultados de análisis de la data recolectada.

3.1 Web Scraping

El proceso de Web Scraping se realizó el 29 de septiembre del 2020 para las zonas urbanas de los municipios de Manizales, Fusagasugá y Villavicencio. Para cada una de estas zonas se consultaron las casas y los apartamentos en arriendo y venta. Los resultados obtenidos se resumen en la tabla III.

Ciudad	Tipo	No. Venta	No.arriendo	Total
Manizales	casa	2.352	119	2.471
	apto	3.267	1.180	4.447
Villavicencio	casa	958	67	1.025
	apto	390	172	562
Fusagasugá	casa	539	11	604
	apto	209	15	224
	Total	7.715	1.564	9.333

Tabla III: Resultados del proceso de Web Scraping

Fuente: elaboración propia

Por otra parte, el estudio tuvo en cuenta variables externas (sitios de interés o amenities), las cuales fueron extraídas de Open Street Maps y Google Places. La tabla IV resume el número de puntos de interés considerados para cada una de las áreas de estudio.

Ciudad	No.puntos
Manizales	2.588
Villavicencio	3.317
Fusagasugá	957

Tabla IV: Número de sitios de interés por área de estudio

Fuente: elaboración propia

3.2 Análisis de datos inmobiliarios

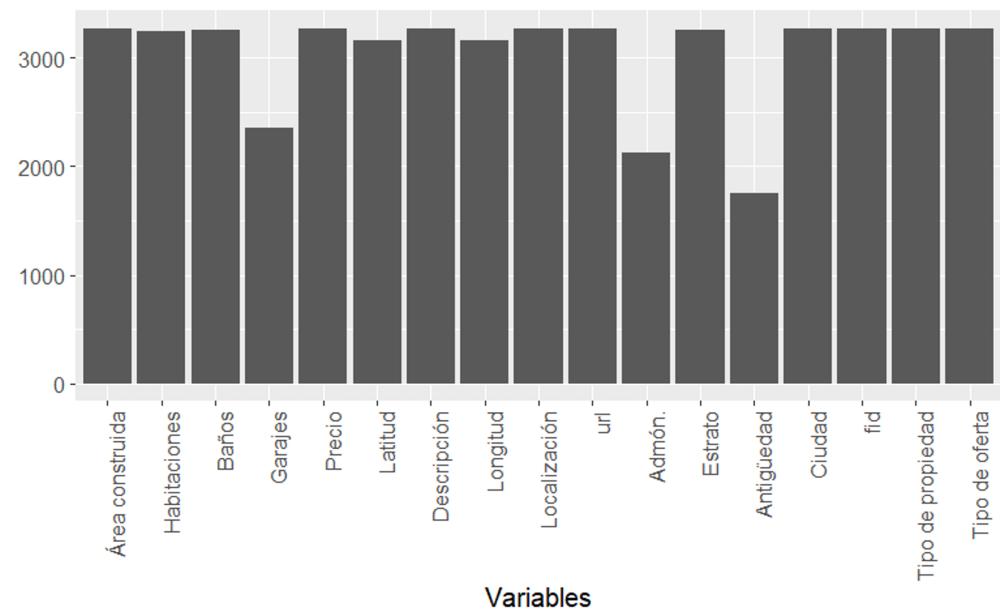
Los resultados de esta etapa se presentan en consideración de los siguientes pasos metodológicos: (i) exploración y limpieza de los datos, (ii) pre-modelado y (iii) modelado.

3.2.1 Exploración y limpieza de los datos

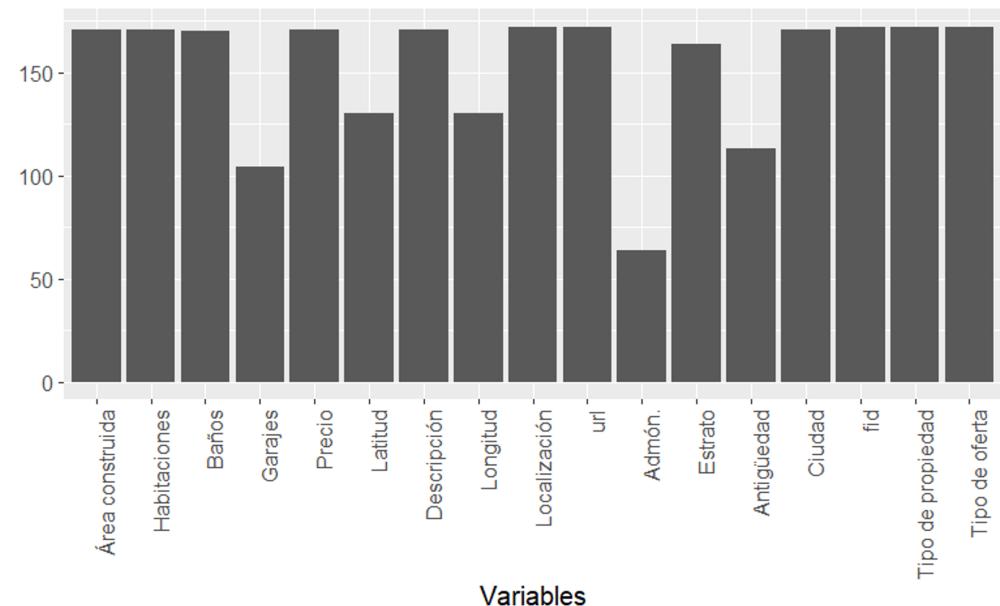
- **Exploración de variables:** se logró para cada área de estudio hacer la identificación de los datos faltantes, encontrando que las variables con mayor ausencia de datos en las publicaciones son: administración, garajes, antigüedad y las variables de carácter espacial longitud y latitud. Se observan en la figura 3 las gráficas de análisis de la completitud de variables para las ciudades de Manizales y Villavicencio.

De la exploración de las variables de texto, disponibles en las publicaciones inmobiliarias, se revisó específicamente el campo de descripción. Se identificó que las palabras que con mayor frecuencia se emplean en las ofertas de casas en venta, para las tres ciudades de estudio, fueron: cocina integral, sala comedor y baño privado. En la figura 4 se muestra la gráfica que representa el análisis de texto para las casas en venta de la ciudad de Villavicencio, donde el tamaño de las palabras es proporcional a la frecuencia de uso dentro de la descripción. El análisis de estos datos puede ser objeto de un estudio más profundo que involucre análisis semánticos, aspecto que se puede considerar para próximas investigaciones.

Con relación al análisis de variables discretas, se generaron gráficos de frecuencias asociados a los valores de cada variable que se disponen en la figura 5, con el fin de caracterizar el comportamiento de características de las viviendas en oferta para las ciudades de estudio. De este análisis se identificaron los siguientes aspectos sobre la dinámica inmobiliaria, que contribuyen con la definición de variables explicativas de los valores de oferta, así: en general, para las tres ciudades, la mayoría de las casas en venta poseen entre 3 y 4 habitaciones, en contraste con la venta de apartamentos que tienden a presentar 3 habitaciones, seguido en frecuencia de la oferta de 2 habitaciones para el mismo tipo de bien; en cuanto al número de baños, las casas en arriendo poseen un comportamiento muy similar para todas las ciudades, siendo las casas con 2 baños las de mayor frecuencia, seguidas de casas con 1 y 3 baños respectivamente; en cuanto al número de estacionamientos o garajes, en general tanto para casas como para apartamentos y sin importar la modalidad de oferta en la que se encuentren (venta o arriendo), presentan un comportamiento similar, siendo aquellos inmuebles de un garaje los más frecuentes de encontrar; con respecto a la distribución de los estratos, el acceso al bien inmueble propio (casa o apartamento) se encuentra en mayor medida concentrada en viviendas con características de estrato 3 o más, principalmente para las ciudades de Manizales y Fusagasugá; por otra parte, considerando la antigüedad del inmueble, en Fusagasugá la



a) Casas en venta en Manizales



a) Apartamentos en arriendo en Villavicencio

Figura 3: Completitud de variables

Fuente: elaboración propia



Figura 4: Exploración “descripción” casas en venta en Villavicencio
Fuente: elaboración propia

venta de casas tiene una mayor frecuencia en aquellas de entre 1 a 8 años, seguidas de entre 9 a 15 y 16 a 30 respectivamente.

• Limpieza de los datos:

Una vez aplicados los validadores e imputados los valores nulos, se consiguió el conjunto de datos por ciudades que se presenta en la tabla V. Los conjuntos de datos con menos de 50 publicaciones se descartaron de los análisis posteriores, ya que el número de muestras no era suficiente; por tanto, los conjuntos de datos de casas en arriendo para la ciudad de Villavicencio, casas y apartamentos en arriendo en la ciudad de Fusagasugá no se tuvieron en cuenta.

Ciudad	Tipo	No. Venta	No.arriendo	Total	Porc.Total datos
Manizales	casa	1.992	99	2.091	84,6
	apto	3.005	1.027	4.032	90,7
Villavicencio	casa	720	45	765	74,6
	apto	314	120	434	77,2
Fusagasuga	casa	310	3	313	51,8
	apto	137	8	145	64,7
	Total	6.478	1.302	7.780	83,4

Tabla V: Resultados del proceso de limpieza

Fuente: elaboración propia

- **Exploración geográfica:** en la figura 6.a se representa la distribución espacial del número de publicaciones de la ciudad de Manizales. En términos generales, para este caso, el número de ofertas de apartamentos en venta muestra una concentración que tiende a ubicarse en torno a las principales estructuras de movilidad vial, como es el caso de la Vía Panamericana, la Carrera 23 y la Carrera 14. Es importante mencionar que, si bien las vías atraviesan la ciudad

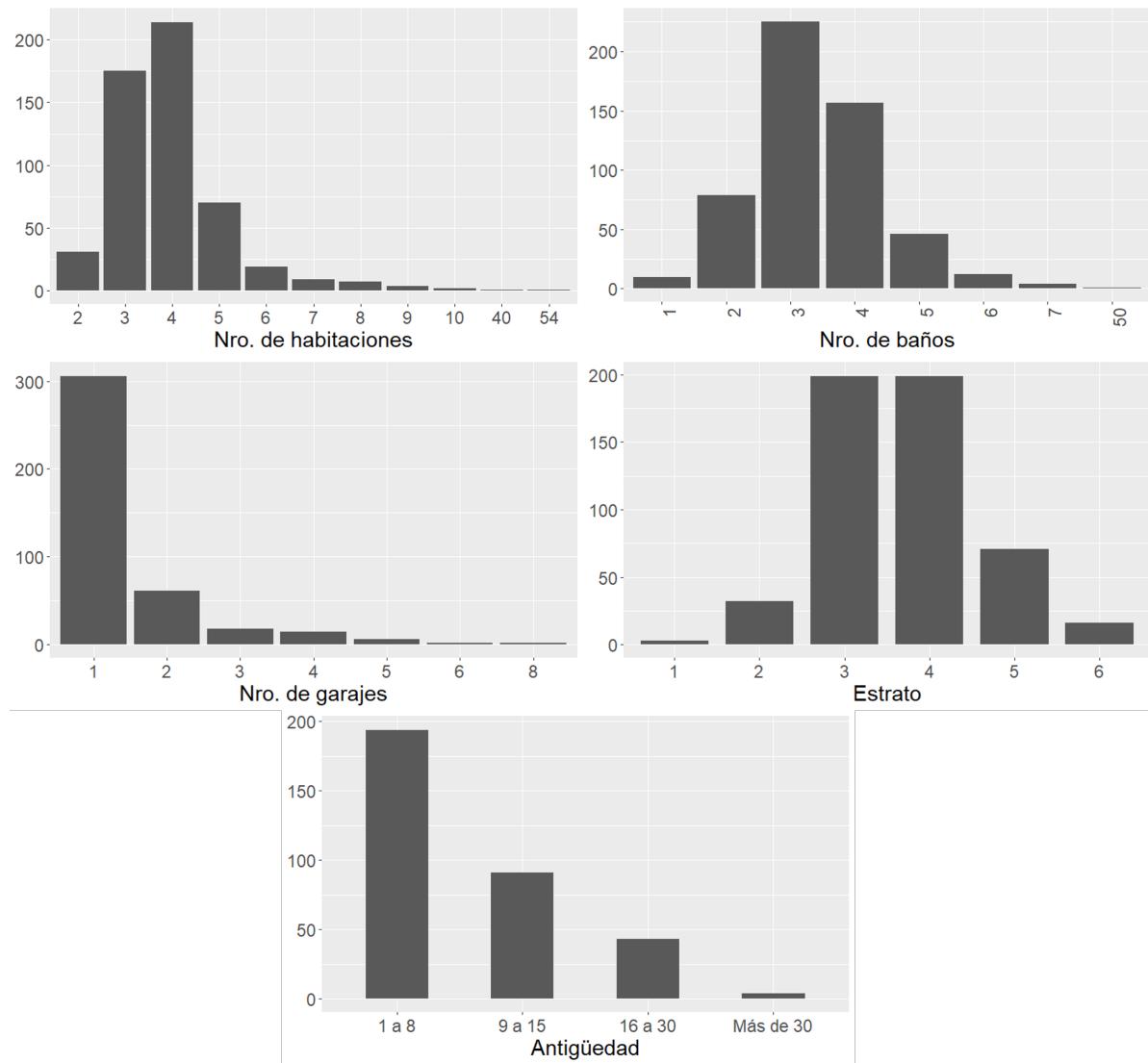


Figura 5: Exploración de variables discretas de casas en venta en Fusagasugá

Fuente: elaboración propia

de manera longitudinal y en sentido sur norte, el número de ofertas tiene un comportamiento radial, lo que implica que hay mayor número de ofertas hacia el centro de la ciudad, y disminuyen hacia la periferia de la misma, sobre todo en los extremos suroriental y noroccidental de Manizales.

El mapa de calor de la figura 6.b reafirma la relevancia de la estructura vial, los centros dotacionales y de comercio en la determinación de las zonas con mayor concentración de ofertas, evidenciando mayor número de ofertas hacia el centro de la ciudad y alrededor de las Carreras 23 y 14.

De acuerdo con la figura 6.c, la concentración de los sitios de interés en Manizales se distribuye a lo largo de las Carreras 25, 23 y 14, presentando concentraciones cerca de las infraestructuras de servicios tales como las universidades, el estadio Pologrande, centros turísticos como las Termales Tierra Viva y los centros comerciales Galerías Plaza, Fun-

dadores, Mallplaza Manizales y Parque Caldas.

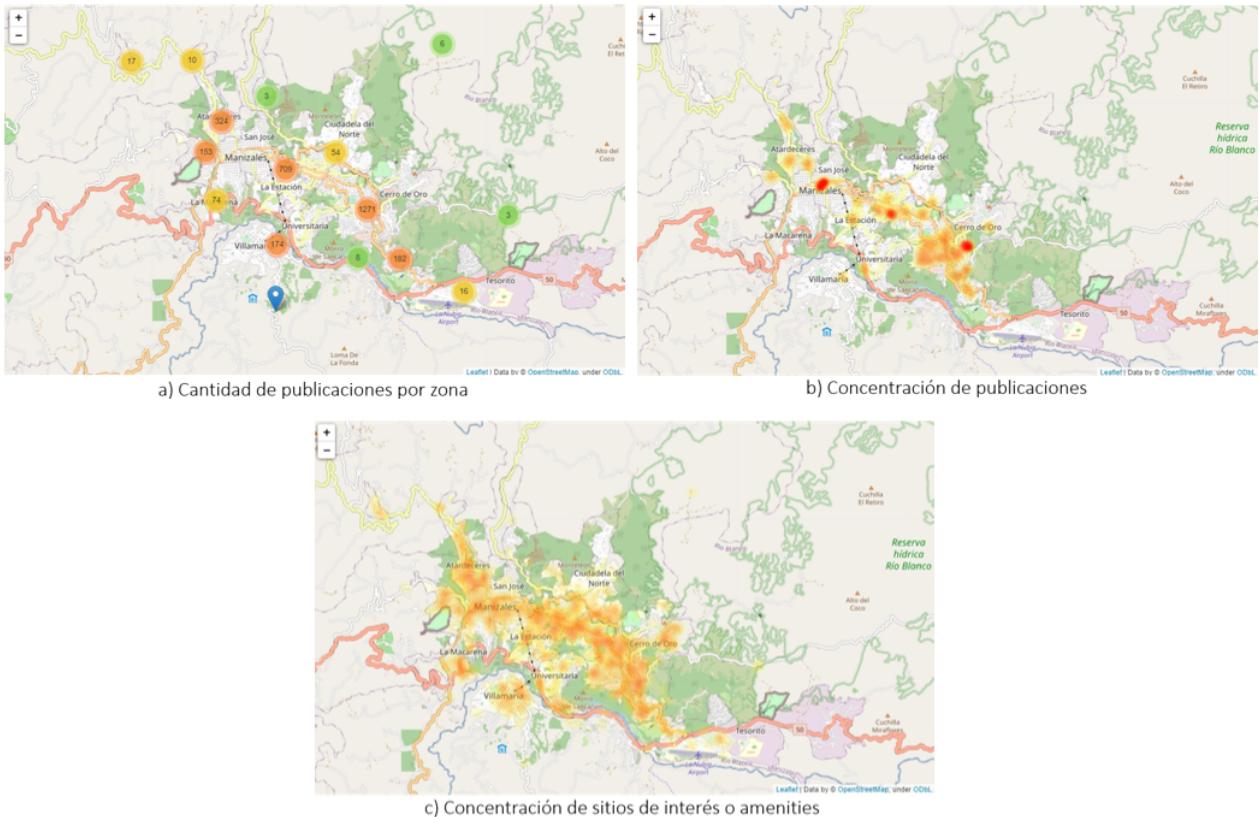


Figura 6: Exploración espacial de apartamentos en venta en la ciudad de Manizales
Fuente: elaboración propia

- **Exploración de variables explicativas vs variable objetivo:** producto de la exploración de variables para casas en venta de la ciudad de Fusagasugá se obtuvieron los resultados resumidos en la figuras 7 y 8. Primero se realizó la exploración de la distribución de las variables objetivo como se observa en la figura 7, a partir de las cuales se resalta que:

- El precio de las ofertas se concentra alrededor de 64.600 dólares, existiendo otras ofertas cercanas a los 820.000 dólares.
- La mayoría de las publicaciones presenta un valor por m^2 entre 450 dólares y 600 dólares; sin considerar los datos atípicos, existen ofertas donde se tiene un valor por m^2 de 1.400 dólares.

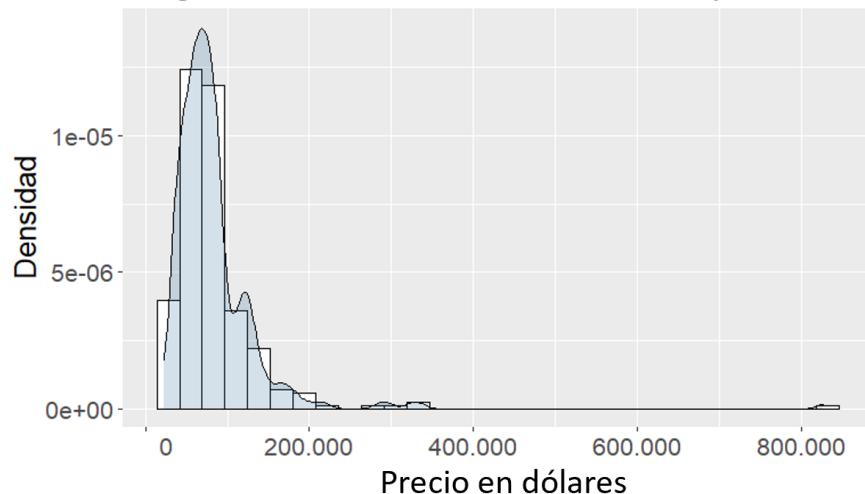
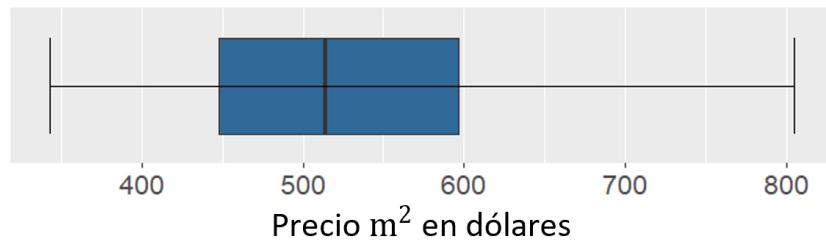
Diagrama de frecuencia de la variable precio**Diagrama de cajas de la variable precio m²**

Figura 7: Distribución de precios para casas en venta en Fusagasugá
Fuente:elaboración propia

Posteriormente se realizó la exploración de las variables objetivo versus las variables explicativas, mediante diagramas de cajas como se observa en la figura 8, a partir de la cuales se destaca que:

- Las variables explicativas número de habitaciones, número de baños y número de garajes versus el precio del inmueble al parecer comparten una relación directamente proporcional, por tanto, a medida que aumentan en número, también aumenta el precio de la oferta.
- En cuanto a la antigüedad se refiere, aparentemente esta no influye en el costo del inmueble, aunque es especialmente llamativo el comportamiento de los inmuebles con más de 30 años, el cual está influido por un inmueble con un precio de más de 820.000 dólares.
- De acuerdo con los diagramas de cajas para el precio de los inmuebles y el precio del metro cuadrado normalizados por estrato y sin considerar datos atípicos, se puede suponer que la variable estrato comparte una relación directamente proporcional con el precio de la oferta.

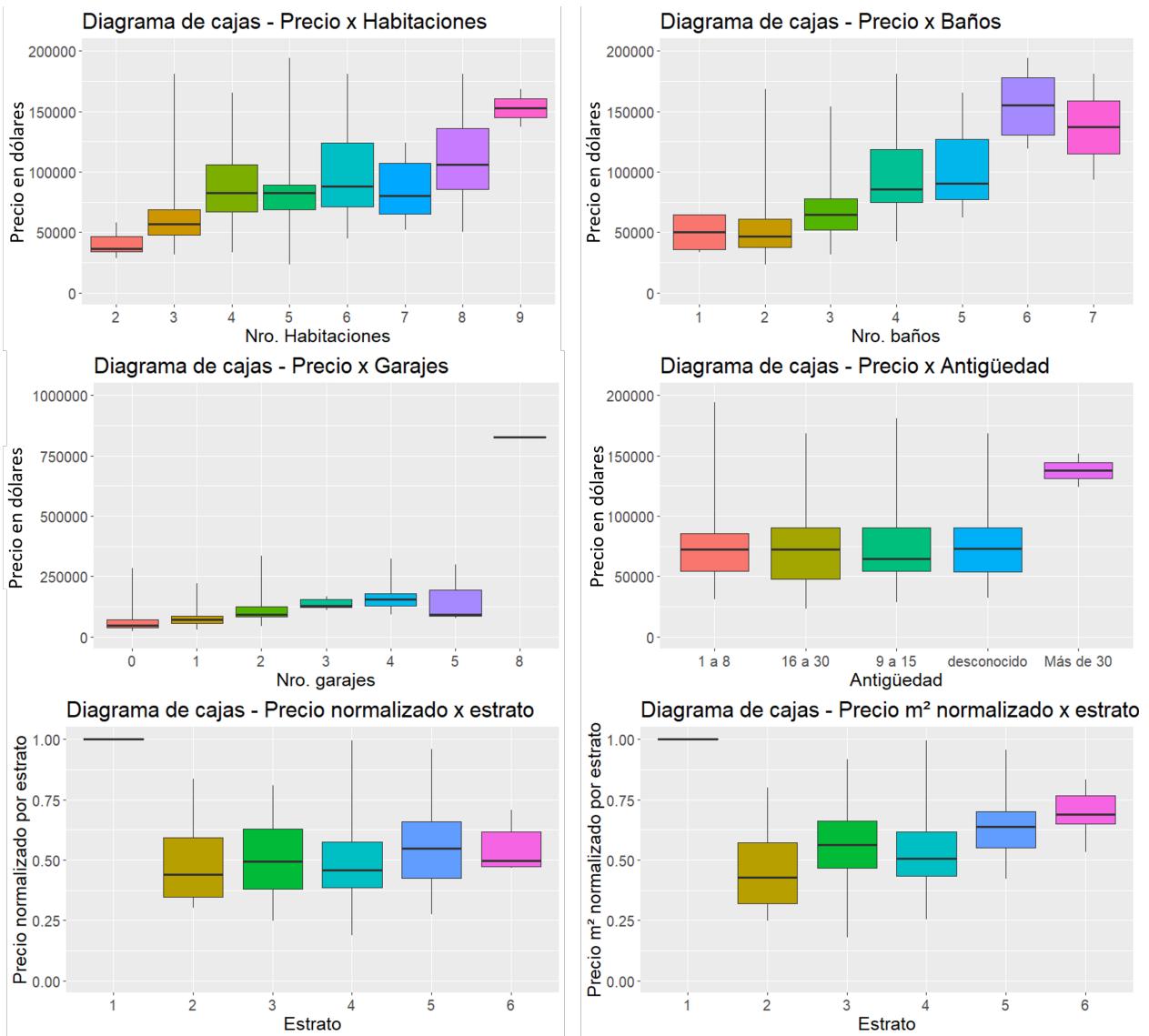


Figura 8: Exploración de variables explicativas vs Variable objetivo de casas en venta de Fusagasugá

Fuente: elaboración propia

3.2.2 Pre-modelado

Con el fin de observar la contribución de cada variable explicativa, tanto interna como externa, con respecto al valor del inmueble, las figuras 9.a y 9.c evidencian que, para la venta de casas en Manizales, los métodos de selección univariada y basada en árboles de decisión coinciden en que las variables área construida, estrato y número de baños, propias del inmueble, son las que mayor contribución generan al valor del mismo, en contraposición a la variable antigüedad que es la que menos contribución tiene. Los dos métodos anteriores se diferencian de la selección por eliminación recursiva de características (ver figura 9.b), en que este determinó que las variables administración, antigüedad y número de habitaciones son las que más influencia explicativa mues-

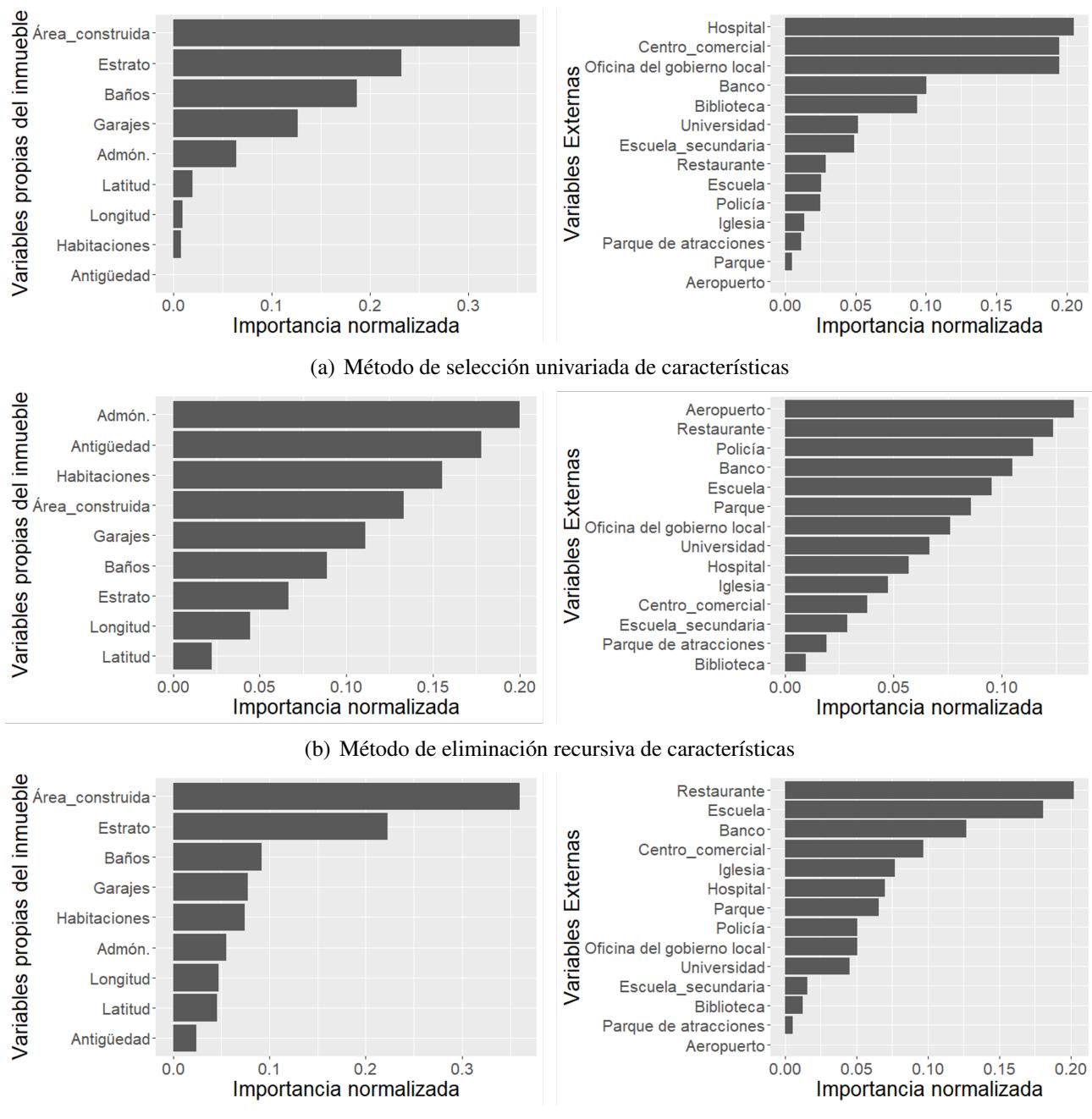


Figura 9: Importancia estadística de las características propias del inmueble y las variables externas en el valor de la oferta de casas en venta en Manizales

Fuente: Elaboración propia

tran.

En cuanto a las variables externas del inmueble no hay un consenso pleno; las técnicas de eliminación recursiva y selección basada en árboles de decisión son las únicas que coinciden en que la variable externa restaurante es una de las que mayor poder explicativo tiene, en contraste con la selección univariada que plantea que las tres variables externas que más contribuyen al modelo son hospital, centro comercial y oficina de gobierno local.

3.2.3 Modelado

Los resultados sobre la métrica definida obtenida por los modelos, luego del entrenamiento y ajuste de hiperparámetros, se muestran en la gráfica de barras de la figura 10.

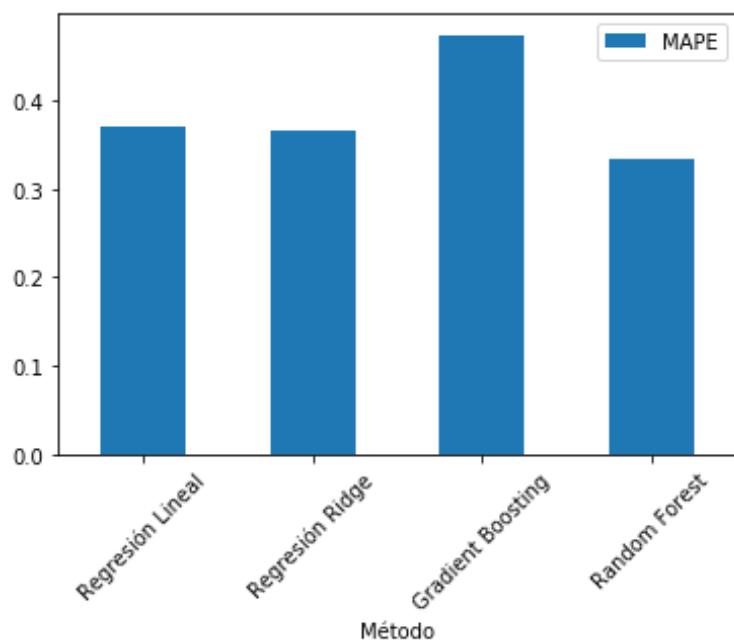


Figura 10: Comparación de modelos de predicción
Fuente: elaboración propia

Como se observa el mejor modelo fue el de Random Forest, seguido muy de cerca por el modelo de Regresión Ridge.

4 Principales contribuciones

Este proyecto tuvo como objetivo general establecer la metodología para la obtención, la depuración, el almacenamiento y el análisis de datos provenientes de plataformas web inmobiliarias, como fuente de información para Observatorios Inmobiliario Catastrales . Este propósito es consecuente con las necesidades que tienen los gestores catastrales de conocer la dinámica inmobiliaria de un territorio, en el marco de Catastro Multipropósito. Por tanto, los resultados obtenidos contribuyen específicamente con los siguientes aspectos:

- Exploración de la viabilidad técnica para hacer operativa la adquisición, la limpieza y el análisis de datos de ofertas inmobiliarias, disponibles en la web, que sean fuente de información para los observatorios inmobiliarios catastrales, haciendo uso de Web Scraping.
- Adquisición de conocimiento sobre la dinámica inmobiliaria, mediante la automatización del proceso para colecta e integración de datos inmobiliarios y fuentes abiertas de variables prediales externas, con incidencia en el valor comercial del inmueble para una región determinada.
- Estimación del precio del inmueble por medio de técnicas de Machine Learning con base en las características propias y la ubicación geográfica. Los modelos se entrenaron con los datos obtenidos del Web Scraping. Cada uno de los anteriores aspectos están asociado a un componente de software que permite automatizar la exploración de los datos obtenidos del Web Scraping así como la generación de modelos predictivos.

Adicionalmente, este tipo de soluciones permiten la programación de procesos de sincronización con las fuentes de información, con lo cual se generan herramientas para la actualización permanente de los Observatorios Inmobiliarios Catastrales. Desde el ámbito metodológico, el IGAC estará incursionado en el uso de nuevas técnicas para la mejora de los procesos catastrales requeridos en el marco de la implementación de la política pública de Catastro Multipropósito.

5 Conclusiones

El estudio realizado perfiló una ruta para la gestión y el análisis automático de la información inmobiliaria y de su entorno, teniendo como fuentes diferentes sitios web, que complementan las fuentes oficiales al brindar disponibilidad de datos donde se carezca de estos; lo anterior gracias a un proceso metodológico que permitió identificar las características propias y las variables externas que afectan el precio de los inmuebles en el mercado. Por tanto, los modelos de Machine Learning derivados del análisis de variables permiten estimar el precio de otros inmuebles de características similares. Lo anterior se convierte en un punto de partida para el proceso de valuación masiva basada en información real del mercado.

La implementación de la técnica de Web Scraping para la recolección automática de los datos de ofertas inmobiliarias se convierte en una herramienta que facilita el acceso a la información para diferentes interesados, entre ellos los gestores catastrales y la ciudadanía en general. Este tipo de soluciones tecnológicas brinda, por tanto, transparencia al permitir oportunidad, actualidad y mayor nivel de detalle en la caracterización de los mercados inmobiliarios en las diferentes zonas del país. Igualmente, ofrece una fuente de datos completa para los observatorios inmobiliarios, mediante la cual se puede examinar el mercado de viviendas en venta y en arriendo. Adicionalmente, el conjunto de datos recolectados es de provecho para la planificación territorial e investigaciones que busquen entender la dinámica inmobiliaria incorporando el componente espacial.

Es importante aclarar que para tener una buena aproximación de las predicciones es necesario contar con datos suficientes que alimenten los modelos de Machine Learning, logrando capturar correctamente los patrones estadísticos inherentes.

Agradecimientos

Se agradece al equipo técnico del Observatorio Inmobiliario y Catastral del Instituto Geográfico Agustín Codazzi por sus aportes en la identificación de variables inherentes a los predios, así como a los proveedores alternativos de información inmobiliaria.

References

- [1] Ulbricht L. 2020 Scraping the demos. digitalization, web scraping and the democratic project. <https://doi.org/10.1080/13510347.2020.1714595> **27**, 426–442. (doi:10.1080/13510347.2020.1714595).
- [2] Uzun E. 2020 A novel web scraping approach using the additional information obtained from web pages. *IEEE Access* **8**, 61726–61740. (doi:10.1109/ACCESS.2020.2984503).
- [3] Bricongne JC, Meunier B, Sylvain P. 2021 Web scraping housing prices in real-time: the covid-19 crisis in the uk. *SSRN Electronic Journal* (doi:10.2139/SSRN.3916196).
- [4] Hillen J. 2019 Web scraping for food price research. *British Food Journal* **121**, 3350–3361. (doi:10.1108/BFJ-02-2019-0081/FULL/XML).
- [5] Morshed R, Chu B, Huang E. 2019 Web scraping: Applications in infrastructure planning. *New South Wales*.
- [6] Dewi LC, Meiliana, Chandra A. 2019 Social media web scraping using social media developers api and regex. *Procedia Computer Science* **157**, 444–449. (doi:10.1016/J.PROCS.2019.08.237).
- [7] Krotov V, Johnson L, Silva L. 2020 Tutorial: Legality and ethics of web scraping. *Faculty & Staff Research and Creative Activity* **47**, 539–563. (doi:<https://doi.org/10.17705/1CAIS.04724>).
- [8] DNP. 2019. Conpes 3958. (doi:10.1109/ISSPA.1999.815782).
- [9] DANE. 2020. Decreto 148 de 2020.
- [10] Eguino, Huáscar; Erba, Diego; Da Silva, Everton; De Oliveira, Augusto; Piumetto, Mario; Iturre, Teresa; Rodríguez Ramírez A. 2020 *Catastro, valoración inmobiliaria y tributación municipal: Experiencias para mejorar su articulación y efectividad*. Inter-American Development Bank. (doi:10.18235/0002437).
- [11] Saurkar AV, Pathare KG, Gode SA. 2018 An overview on web scraping techniques and tools. *International Journal on Future Revolution in Computer Science & Communication Engineering* **4**, 4, 363–367.
- [12] Alcaldía de Manizales. 2020. Información General – Alcaldía de Manizales.
- [13] Departamento Administrativo Nacional de Estadística - DANE. 2020. ¿Cuántos somos?
- [14] Instituto Geográfico Agustín Codazzi - IGAC. 2020. Datos Abiertos Catastro — GEOPORTAL.
- [15] Alcaldía de Villavicencio. 2020. Presentación.
- [16] Alcaldía de Fusagasugá. 2020. Presentación.
- [17] Shafiee S, Wautelet Y, Hvam L, Sandrin E, Forza C. 2020 Scrum versus Rational Unified Process in facing the main challenges of product configuration systems development. *Journal of Systems and Software* **170**, 110732. (doi:10.1016/j.jss.2020.110732).
- [18] Huber S, Wiemer H, Schneider D, Ihlenfeldt S. 2019 DMME: Data mining methodology for engineering applications - A holistic extension to the CRISP-DM model. In: *Procedia CIRP*, vol. 79, pp. 403–408. Elsevier B.V. (doi:10.1016/j.procir.2019.02.106).
- [19] Nylen EL, Wallisch P. 2017 Web Scraping. In: *Neural Data Science*, pp. 277–288. Elsevier. (doi:10.1016/b978-0-12-804043-0.00010-6).
- [20] Glez-Peña D, Lourenço A, López-Fernández H, Reboiro-Jato M, Fdez-Riverola F. 2013 Web scraping technologies in an API world. *Briefings in Bioinformatics* **15**, 5, 788–797. (doi:10.1093/bib/bbt026).
- [21] Baldominos A, Blanco I, Moreno AJ, Iturrarte R, Bernárdez Ó, Afonso C. 2018 Identifying real estate opportunities using machine learning. *Applied Sciences (Switzerland)* **8**, 11, 2321. (doi:10.3390/app8112321).
- [22] Wirth R, Hipp J. 2000 Crisp-dm: Towards a standard process model for data mining. In: *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK.

Andrés Enrique Rosso Mateús

Ingeniero de Sistemas de la Universidad Distrital Francisco José de Caldas y Doctor en Ingeniería de Sistemas y Computación de la Pontificia Universidad Javeriana; Doctor en Ingeniería de la Universidad Nacional de Colombia. Correo electrónico: andres.rosso@igac.gov.co

Yeimy Maryuri Montilla Montilla

Ingeniera de Sistemas de la Universidad Nacional Abierta y a Distancia UNAD, Especialista en Análisis Espacial de la Universidad Nacional de Colombia, estudiante de Maestría en Geomática de la Universidad Nacional de Colombia. Correo electrónico: yeimy.montilla@igac.gov.co

Sonia Constanza Garzón Martínez

Ingeniera Catastral y Geodesta, Especialista en Sistemas de Información Geográfica, Magíster en Ciencias de la Información y las Comunicaciones de la Universidad Distrital Francisco José de Caldas. Correo electrónico: sonia.garzon@igac.gov.co