Universidad de los Andes
Big Data and Machine Learning for Applied Economics
Taller 3

Integrantes:
David Santiago Caraballo Candela [1]
Sergio David Pinilla Padilla [2]
Juan Diego Valencia Romero [3]

14 de noviembre de 2022

GitHub: Repositorio Taller 3

1 Introducción

La industria inmobiliaria ha sido un tema de investigación central en la economía moderna por sus implicaciones en campos como la construcción, la inversión y el bienestar público (Zhang, 2021). El inminente crecimiento de las sociedades junto con la globalización han impulsado el desarrollo de nuevos proyectos de vivienda para satisfacer la demanda. Año tras año, los espacios disponibles para construir se reducen y las áreas circundantes sufren importantes cambios, por lo que los precios "siempre están al alza" y seguirles el ritmo es complejo. Hamizah et al. (2020) sostienen que el problema de *pricing* surge cuando existen numerosas variables, como la ubicación, el tamaño o la antigüedad, que pueden influir en el precio de la vivienda, por lo que la mayoría de las partes interesadas, incluidos los compradores y promotores, los contratistas y la industria inmobiliaria como tal, desearían conocer los atributos exactos o los factores precisos que influyen en el precio de la vivienda para ayudar a los inversores a tomar óptimas decisiones y a los constructores a fijar la valoración de mercado.

La correcta estimación de los precios de las viviendas supone un reto para la economía urbana ya que el desempeño del sector es un insumo esencial para la creación de planes urbanísticos y la fijación del presupuesto de infraestructura de las ciudades. Las aproximaciones más recientes se han orientado al uso de inteligencia artificial y aprendizaje de máquinas con métodos sofisticados que tienen en cuenta un sinnúmero de variables y parámetros que tienden a omitirse o dejarse de lado, lo que mejora la precisión de las predicciones. Tal es el caso de Truong et al. (2020) quienes calcularon los precios de las viviendas en Beijing, China (2018) con una base de datos que únicamente contiene los atributos (i.e. área, habitaciones, baños, tipo de vivienda, entre otras). De igual forma, Ihre & Engström (2019) implementaron dos algoritmos para predecir los precios del *Ames Housing Dataset* disponible en Kaggle y contrastarlos con aquellos fijados por las inmobiliarias del mercado.

El contexto anterior colapsa a una sola pregunta: ¿se puede monetizar con *machine learning*? En 2019 nació HABI, una plataforma digital fundada en Colombia que compra viviendas, las remodela y las vuelve a sacar a la venta -*flipping* en la jerga norteamericana-, generando retornos financieros en el proceso. Sin embargo, en 2021, Zillow de Estados Unidos se aventuró en este nicho de mercado, pero sus modelos sobrestimaron considerablemente el precio de las viviendas. Esta sobre-estimación supuso unas pérdidas de unos US$ 500 millones para la empresa y una reducción del 25% de su fuerza laboral, eliminando la subdivisión encargada. Luego, identificar los verdaderos precios de las viviendas no es desafío un fácil sino todo lo contrario, es un proceso que debe abordarse con exhaustividad. Los datos utilizados en este ejercicio provienen de PROPERATI, una firma de bienes raíces similar a HABI cuyo propósito radica en conectar a compradores y vendedores de manera rápida y efectiva.

[1] Código: 201813007.
[2] Código: 201814755.
[3] Código: 201815561.

Este trabajo constituye una aplicación práctica de los tópicos aprendidos en el curso de *Big Data and Machine Learning for Applied Economics* sobre predicción y estimación. El enfoque inicia con la limpieza de texto en las descripciones y título de los anuncios en la base de datos de entrenamiento. Luego, a partir de expresiones regulares y datos geo-espaciales se hallan nuevos predictores relacionados con características tanto internas como externas de cada vivienda. El modelo de predicción atiende a dos SuperLearner's (con y sin *prior*) con tres sub-modelos; i) XGBoost; ii) Redes Neuronales, y; iii) Lasso con una funcion de perdida *ad hoc* ABS-EXP que pondera en valor absoluto a las predicciones por debajo del precio de lista hasta por COP $40 millones y exponencialmente aquellos por encima y debajo de COP $40 millones.

El documento se organiza de la siguiente manera: la sección 2 contiene una descripción detallada de los datos a utilizar así como su procesamiento. En la sección 3 se describen los modelos de aprendizaje de maquinas cuyo objetivo es determinar el precio de las viviendas en la muestra de *test*, a la vez que se comenta sobre los resultados. Consideraciones finales en la sección 4.

2 Datos

Los datos aquí empleados, provenientes de PROPERATI, se organizan en dos bases distintas, ambas a nivel observacional de vivienda. Por un lado, el archivo nombrado `train.Rds` contiene una muestra de 51.437 observaciones y 12 variables recolectadas en las ciudades de Bogotá D.C. y Medellín. Por otro lado, el archivo `test.Rds` es una muestra de testeo (prueba) que contiene 5.000 observaciones con características idénticas a `train.Rds` pero con viviendas de la ciudad de Cali. En el presente ejercicio, cada una de las bases de datos cumple una función distinta. La base identificada como *test* no posee información sobre el precio de las viviendas y, por lo tanto, se busca predecir el valor comercial que aparece en la página web de PROPERATI. Para lograr esto, se utilizó la base de *train*, que sí poseía información acerca de los precios de las viviendas, como insumo para entrenar y evaluar la efectividad de los modelos de predicción esbozados en la sección 3. Lo anterior mediante la inclusión de nuevas variables extraídas a partir del título, la descripción y la ubicación (latitud-longitud) relativa de los bienes inmuebles dentro de la ciudad.

La Figura 1 (Anexos) muestra que la distribución de los precios en la muestra *train* está sesgada hacia la izquierda, con un valor medio de $\approx$ COP $745 millones por vivienda y con valores atípicos en la cola derecha ($>$ COP $2.200 millones). Asimismo, la Tabla 2 realiza un ejercicio de estática comparativa entre las variables disponibles para ambas muestras (i.e. tipo de propiedad[4], superficie -en m^2-, número de habitaciones y baños[5]). El 78% de las viviendas en la base *train* corresponden a apartamentos que promedian los 221,84 m^2, poseen 3,1 habitaciones y 2,65 baños. El 22% restante refiere a casas, lotes y/o bodegas con estructura (tamaño) similar. Por su parte, en la base *test* el porcentaje de apartamentos cae al 56%, pero que a su vez tienen mayor: i) metraje medio con 268,52 m^2 (+22%); ii) 3,75 habitaciones (+21%), y; iii) 3,11 baños (+17%). De esta base también se conoce de antemano que las viviendas están valuadas en $\approx$ COP $555 millones cada una con una desviación estándar de $\approx$ COP $650 millones.

Si bien la revisión preliminar de las variables en la base *train* es un primer paso para la estimación de los precios de las viviendas en *test*, la ausencia sistemática de atributos tanto dentro como fuera de la base (e.g. metraje, antigüedad, estrato, amenidades, entre otras) dificulta el ejercicio de predicción (ver Tabla 3, Anexos). Así las cosas, para seguir explotando la disponibilidad implícita de información, se aislaron las variables de "título" y "descripción" para extraer, a partir de expresiones regulares, un nuevo *set* de predictores que ayuden a identificar las características propias de las viviendas que influyen directa e indirectamente en su precio.

[4]Variable *dummy* donde 1 hace referencia a apartamento y 0 de lo contrario.

[5]Cuyos valores faltantes fueron aproximados, inicialmente, a partir del Anexo Técnico de Especificaciones de Vivienda y Obras de Urbanismo del Ministerio de Vivienda que erige la normativa vigente sobre los requerimientos mínimos de áreas de lavabo y aseo con la regla de, al menos, 0,75 baños completos (con ducha) por habitación.

La Tabla 4 da cuenta de los diez (10) bigramas[6] con mayor frecuencia descriptiva y concluye que los atributos referidos a las áreas comunes (i.e. sala, comedor); individuales (i.e. habitaciones, baños); de aseo (i.e. zona de ropas), la propiedad horizontal (i.e. vigilancia, parqueadero) y cercanías (i.e. comercios, vías de acceso) son aquellos que más tienen en cuenta los propietarios al listar sus viviendas. En esta misma línea, la Tabla 5 fue creada con el propósito de identificar la frecuencia descriptiva de las expresiones regulares, nuevamente bigramas, que comparte el decil más bajo de precio con el más alto (10) así como una evaluación individual (10 c/u) -para un total de 30-. A partir de este insumo junto con variaciones de las expresiones regulares se crearon nuevos predictores[7] (ver Tabla 7, Anexos) que dan cuenta de las diferencias narrativas en las descripciones de las viviendas. En especial, se encuentra que aquellas viviendas por debajo del umbral de los COP $270 millones son más propensas a contener 2 habitaciones y 2 baños así como de mencionar explícitamente que cuentan con red a gas, parqueadero cubierto, cercanía a paraderos de bus o piso en cerámica, pues son atributos que no se dan por sentado en las propiedades ya que no son comunes en los rangos de bajo precio. Mientras tanto, los inmuebles por encima de los COP $1.600 millones consignaron en la descripción que disponen de comedor independiente/auxiliar, ascensor privado, chimenea, 4 habitaciones, pisos en madera e incluso jacuzzi. Igualmente, la inclusión de estas características tienen como fin destacar la presencia de "lujos extra" tanto dentro como fuera de la vivienda.

Ahora bien, para complementar el *dataset* que se obtuvo del procesamiento de lenguaje natural, se buscaron fuentes externas especializadas que aportaran nueva información geo-espacial de Bogotá D.C., Medellín y Cali, tanto a nivel vivienda como de cada manzana. Las tres fuentes externas consultadas fueron: i) la pagina de Open Street Maps (OSM), en donde se obtuvo información de la ubicación de diferentes tipos de *amenities* que pueden estar influenciando los precios de los hogares en cada ciudad; ii) el Marco Geoestadístico Nacional (MGN), el cual permitió ubicar a la mayoría de los hogares en la base de datos dentro de una manzana específica en cada ciudad, y; iii) el Censo Nacional de Población de 2018 para Colombia, gracias al cual se extrajeron datos demográficos para cada una de las manzanas identificadas en el MGN.

De OSM se extrajo el dato de 13 *amenities* en cada ciudad, las cuales se presume pueden ser relevantes al realizar el ejercicio predictivo del valor de las propiedades. Diez (10) de estas variables están relacionadas con medidas de acceso a servicios dentro de la ciudad, en tanto corresponden a la distancia de cada hogar al: i) centroide de la ciudad en cuestión (*proxy* de acceso al trabajo); ii) aeropuerto; iii) parada de bus; iv) hospital; v) estación de policía; vi) tienda o centro comercial; vii) bar; viii) restaurante; ix) colegio, y; x) universidad más cercana(o). Las otras tres (3) variables estas relacionadas con medidas de densidad, al ser estas: i) la distancia promedio a todos los parques; ii) ríos y canales, y; iii) a las principales vías de la ciudad.

Las Figuras 2, 3 y 4 capturan los respectivos mapas de calor con el centroide de la ciudad (punto negro) los cuales van en línea con las expresiones regulares más usadas por los propietarios en las descripciones de las viviendas. Se encuentra que existe una fuerte correlación entre la densidad de paradas de bus (*proxy* de vías principales) y las tiendas o comercios para cada ciudad; Bogotá D.C. se concentra al norte mientras que Medellín y Cali al suroccidente/oriente. Por su parte, las zonas verdes de la muestra *train* exhiben una mayor aglomeración exceptuando los valores atípicos en Bogotá D.C. (e.g. Parque Simón Bolivar, Parque la Florida) y Medellín (e.g. Parque Natural Cerro El Volador) frente a los parques o jardines en Cali. Adicionalmente, para capturar posibles no linealidades (en especial concavidades) entre la distancia a los atributos y el precio de una vivienda, se incluyeron expresiones cuadráticas de todas las variables extraídas de OSM dentro de la base de datos de *train*. Estos datos geo-espaciales también permiten entender mejor las características de cada ciudad pues tal como se puede observar en la Tabla 6 (ver Anexos), Bogotá D.C. es la ciudad mas "dispersa", en donde los hogares tienden a estar mas alejados del centroide (6,7 km), y Medellín es la ciudad mas "compacta" (3,8 km).

[6] Grupo de dos letras, dos sílabas, o dos palabras.

[7] Variables *dummy* donde 1 hace referencia si la vivienda posee alguna de las amenidades o 0 de lo contrario.

Intuitivamente, esto indica que la vivienda en Bogotá D.C. no debería ser tan costosa respecto a las otras ciudades, porque estas largas distancias a sitios de trabajo corresponde a un atributo muy poco deseable. No obstante, estas mismas variables de distancia a *amenities* también ayudan a explicar porqué, a pesar de esta condición de ser una ciudad con grandes distancias entre puntos, la capital sigue teniendo las propiedades mejores valuadas entre las tres ciudades (*train* y *test*). Dentro de esta muestra, Bogotá D.C. es la ciudad con una menor distancia mínima promedio al transporte publico (i.e. estaciones de bus), y es la que más fácilmente puede proveer de seguridad (i.e. estaciones de policía), educación (i.e. escuelas y universidades) y entretenimiento (i.e. bares y restaurantes) a sus hogares. Mientras que Cali, la ciudad con propiedades de menor valor, es en la que hay una menor provisión de educación escolar y entretenimiento. Lo cual indica que, potencialmente, este tipo de atributos pueden llegar a ser buenos predictores del valor de las propiedades.

Adicionalmente, a partir de la indexación por manzanas extraídas del MGN, también se obtuvieron 4 variables demográficas del Censo 2018: La mediana del numero de cuartos por vivienda, la suma del total de personas que viven en cada manzana, la mediana del numero de hogares en una vivienda, y la mediana del estrato de las viviendas en cada manzana. Sin embargo, luego de agregar los datos de las manzanas con los datos del precio de los hogares estudiados, se encontró que el 32,7% de los hogares de la base de datos principal no fueron ubicados dentro de una manzana y, por lo tanto, surgió la necesidad de realizar un proceso de imputación de los datos demográficos sobre este conjunto de observaciones. Teniendo en cuenta que para esta base de datos la manzana media contiene dos (2) hogares, se decidió utilizar un algoritmo de *K-Nearest Neighbors (KNN)*, con un hiperparámetro $k^* = 3$, que únicamente considerara los datos de latitud y longitud de las propiedades, para realizar la imputación de los datos demográficos, debido a que es razonable esperar que: i) las observaciones dentro de la manzana mas cercana hagan parte de los vecinos geográficos mas cercanos y ii) si únicamente se toman 3 vecinos geográficos, en la mayoría de los casos las viviendas de la manzana mas cercana serán una mayoría dentro de los vecinos seleccionados.

En la Tabla 8 se evidencia que la manzana mediana de los hogares en esta base (aquella con los datos demográficos imputados) es muy similar entre las tres ciudades, y que, la "vivienda común" corresponde a una propiedad de estrato medio-alto, con 4 habitaciones y perteneciente a un solo hogar, que esta ubicada en una manzana con baja densidad poblacional. Entonces, teniendo en cuenta que la vivienda promedio no posee diferencias significativas en sus características demográficas, pero sí sobre los precios y el acceso a *amenities* entre cada ciudad, es factible pronosticar que los *features* extraídos de OSM son los que aportaran la información relevante para no sobre-estimar los precios de las viviendas que se quieren valuar.

3 Modelos & Resultados

La presente sección desarrolla la metodología para determinar el precio de las viviendas en Cali a partir de las observaciones de Bogotá D.C. y Medellín en la base de entrenamiento. Si bien en algunos casos, (≈ 300), fue posible recuperar el precio con expresiones regulares como "millones" precedidas de números (en número o texto) o tríos de ceros (000), el *core* del trabajo se centró en dos modelos SuperLearner (con y sin *prior*) que capturarán la información más relevante de tres sub-modelos con variación en los parámetros. En orden: i) *Boosting* por XGBoost (9 especificaciones); ii) Redes Neuronales (6 especificaciones), y; iii) Regularización por Lasso (8 especificaciones). Con el fin de mejorar la predicción, las variables continuas (j) fueron estandarizadas con media 0 ($\mu_j = 0$) y desviación estándar igual a 1 ($\sigma_j^2 = 1$). La técnica en *machine learning* de SuperLearner radica en un algoritmo que utiliza validación cruzada para estimar el rendimiento de múltiples modelos con diferentes ajustes o hiperparámetros. Después, crea una medida ponderada óptima de cada uno de estos modelos con base en el rendimiento de los datos de entrenamiento.

A diferencia de las *Deep Neural Networks* que emplean capas de unidades de procesamiento para aprender de la entrada de la capa anterior y "alimentarla", lo que le permite identificar características de nivel superior y facilitar la detección de entidades, el SuperLearner es asintóticamente tan preciso como el mejor algoritmo de predicción posible que se haya probado pues este no sufre de arquitecturas infinitas, opacidad en los resultados ni convergencia relativamente más lenta en conjuntos de datos pequeños. La intención es resolver para un objeto de la siguiente forma: $f^* = \text{argmin}_{f\in\mathcal{F}}\{\sum_{i=1}^{n} L[y_i, f(x_i)]\}$ donde L es la función de pérdida. En este sentido, el SuperLearner es una variación de *stacking* (apilamiento) o de validación cruzada (*K-Fold*), en la que los modelos individuales se entrenarían con una división de datos en K partes iguales (en tamaño), y luego se entrenaría un meta-modelo final con su salida. Por las razones anteriores, el SuperLearner es el enfoque de nuestra preferencia al ser el más robusto para abordar el problema de predicción de precios de las viviendas en la ciudad de Cali. La identificación viene dada por:

1. *Boosting* con XGBoost con una grilla de árboles en el grupo $\mathcal{A} = \{10, 100, 500\}$, una tasa de aprendizaje del 10% y una grilla de profundidad $\mathcal{D} = \{1, 5, 10\}$.
2. Redes Neuronales con una grilla de capas escondidas $\mathcal{H} = \{(100, 1), (250, 5)\}$ y una grilla de la potencia del término de regularización $\mathcal{N} = \{0.001, 0.1, 1.5\}$.
3. Regularización de Lasso con una grilla de potencia del término de regularización $\mathcal{L} = \{0, 0.05, 0.1, 0.2, 0.5, 1, 2, 5\}$.

Los argumentos propios del SuperLearner son; i) $K^* = 10$ *(10-Fold Cross-Validation)* y ii) una función de pérdida personalizada *ABS-EXP* construida a trozos la cual penaliza en valor absoluto cuando las predicciones se situán por debajo del precio de lista (aquel que aparece en PROPERATI) hasta por un desfase de COP $40 millones, exponencialmente si la estimación del precio está por encima del avalúo o debajo por más de COP $40 millones, o no penaliza si el precio de predicción coincide con el real.

En la segunda columna de la Tabla 1 se reportan los resultados del SuperLearner sobre la muestra de entrenamiento. Si se decidiera implementar este modelo para actividades de *pricing* de viviendas en Bogotá D.C. o Medellín, se podría esperar que, en promedio, el precio ofrecido a los vendedores estaría a una distancia de entre COP $95 millones y COP $475 millones de su precio real. Por tanto, vale la pena considerar si existen alternativas para mejorar el ajuste del modelo mediante la aplicación técnicas estadísticas como la econometría Bayesiana. Así pues, para la *train* se desarrolló un *prior* de la predicción de los valores de los hogares, el cual consistió en el mínimo valor real de las viviendas para cada decil de precios y, posteriormente, con esta nueva información se ajustó un nuevo SuperLearner sobre las predicciones del primero declarado y sobre el *prior* de la muestra *train*. La tercera columna contiene este componente Bayesiano.

Tabla 1. Estática comparativa de funciones de pérdida.

MÉTRICA	SuperLearner	SuperLearner+*prior*
log(ABS-EXP)	$56,14	$55,82
RMSE	$475,13	$420,46
MAE	$98,05	$151,50

Nota: Unidades en COP millones.

Con base en esta configuración del *prior*, se encontró que mientras el SuperLearner original tiende a ser mas preciso -menor sesgo-, también tiende a sobrestimar y/o subestimar mas fácilmente el precio de las viviendas en el *train* -mayor varianza-. Mientras que, puede estarse dando que el SuperLearner+*prior* tenga un menor desempeño en el *Mean Average Error (MAE)*, pero si es efectivo en el *shrinkage* de las predicciones iniciales, lo cual permite reducir la varianza -*Root Mean Square Error (RMSE)*- y lleva a que la divergencia entre las predicciones sea menor.

Para resumir, teniendo en cuenta el desempeño de los modelos SuperLearner implementados y la información recolectada durante el procesamiento del lenguaje natural, la estimación de los precios de las viviendas en Cali constó de tres etapas. Primero, se entreno el modelo SuperLearner original y a partir de este se presentaron predicciones preliminares del valor de las propiedades. En segundo lugar, se construyo un *prior* por deciles para las predicciones en el *test* a partir del uso de simulaciones de Montecarlo no paramétricas que seguían la forma de la distribución de los precios en la base *train*, pero que eran consistentes con las estadísticas descriptivas de los precios en Cali. Y, con las predicciones iniciales y el *prior* se construyó un segundo SuperLearner del cual se obtuvieron nuevas predicciones reducidas. Tercero, aprovechando que durante el proceso de procesamiento del lenguaje se lograron extraer datos puntuales sobre los precios de algunas propiedades de Cali, estos datos terminaron sustituyendo a los pronósticos de SuperLearner+*prior* para estas viviendas especificas.

4 Conclusiones

En este documento, se buscó desarrollar una metodología de predicción que permitiera identificar los precios de las viviendas a partir de sus características propias y ajenas. Lo anterior prestando especial cuidado de no cometer el error de sobrevalorar los inmuebles ni tasarlos por debajo de su precio real de mercado ($<$ COP \$40 millones). Luego de implementar las especificaciones de los SuperLearner's, se encontró que el mejor modelo es aquel con *prior* pues obtuvo una métrica (en términos logarítmicos para simplificar la escala) de la función de pérdida personalizada ligeramente mejor a la de su contraparte sin *prior*, lo cual también es consistente con un mejor comportamiento del *RMSE*, y demuestra que este segundo modelo con *shinkage* no incurre en pérdidas tan costosas como el primero (el no Bayesiano).

Nuestra aproximación práctica es un vistazo inicial a los procesos y actividades que llevan a cabo las empresas que, actualmente, hacen dinero con inteligencia artificial y/o aprendizaje de máquinas (e.g. Alphabet, Oracle, Amazon, entre otros). Un ejemplo aplicado es el *data-freemium* de Google que utiliza sus datos propios (recogidos por más de mil millones de búsquedas de usuarios cada día) para vender publicidad dirigida. Esto es, que recopila los patrones de navegación y sugiere promociones/ventas de bienes y/o servicios. En efecto, las firmas dedicadas *full-time* a este rubro económico deben invertir cientos de millones de dolares (US\$) al año para perfeccionar y calibrar sus modelos. Ciertamente, sin garantía de que estos vayan a funcionar o no pues la tasa de mortalidad de *start-ups* en este nicho de mercado es bastante alta ya que resulta una tarea ardua y rigurosa presentar un modelo de negocio que efectivamente genere rentabilidad financiera y tenga verdadero valor para satisfacer la demanda de los consumidores. De esta forma, se espera que las predicciones realizadas sean las mejores posibles para lograr maximizar los beneficios de las firmas de bienes raíces que quieran invertir en Cali, y que este algoritmo de valoración pueda ser replicable para futuros usos empresariales en el sector de bienes raíces.

Bibliografía

Hamizah, N., Rahman, S., Hasbiah, N., & Ibrahim, I. (2020). House price prediction using a machine learning model: A survey of literature. *Modern Education and Computer Science*, 6, pp. 46-54.

Ihre, A., & Engström, I. (2019). Predicting house prices with machine learning methods. *KTH VETENSKAP OCH KONST*.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. *Procedia Computer Science*, Volume 174, pp. 433-442.

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Hindawi*.

Anexos

Figuras

Figura 1. Precios de la vivienda (muestra *train*).

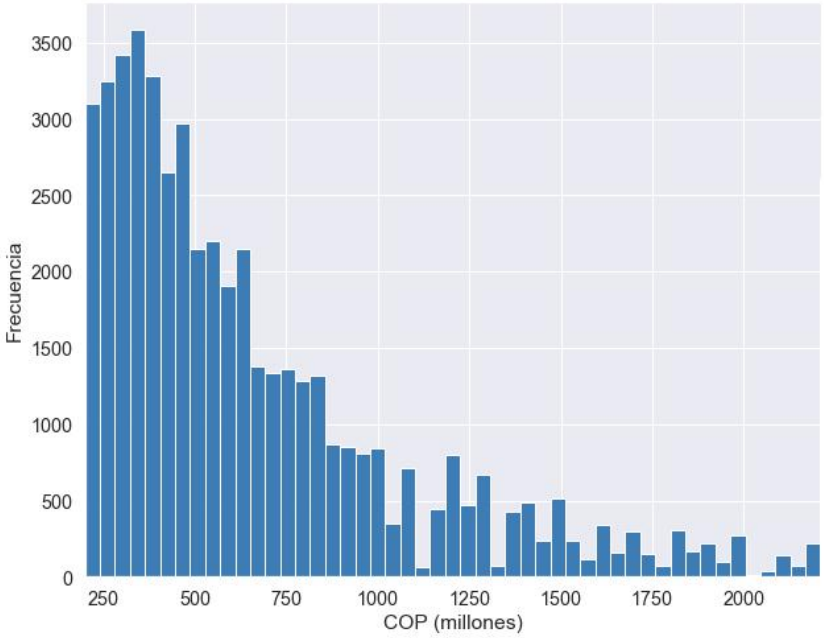


Figura 2. Distancia al paradero de bus más cercano por ciudad.

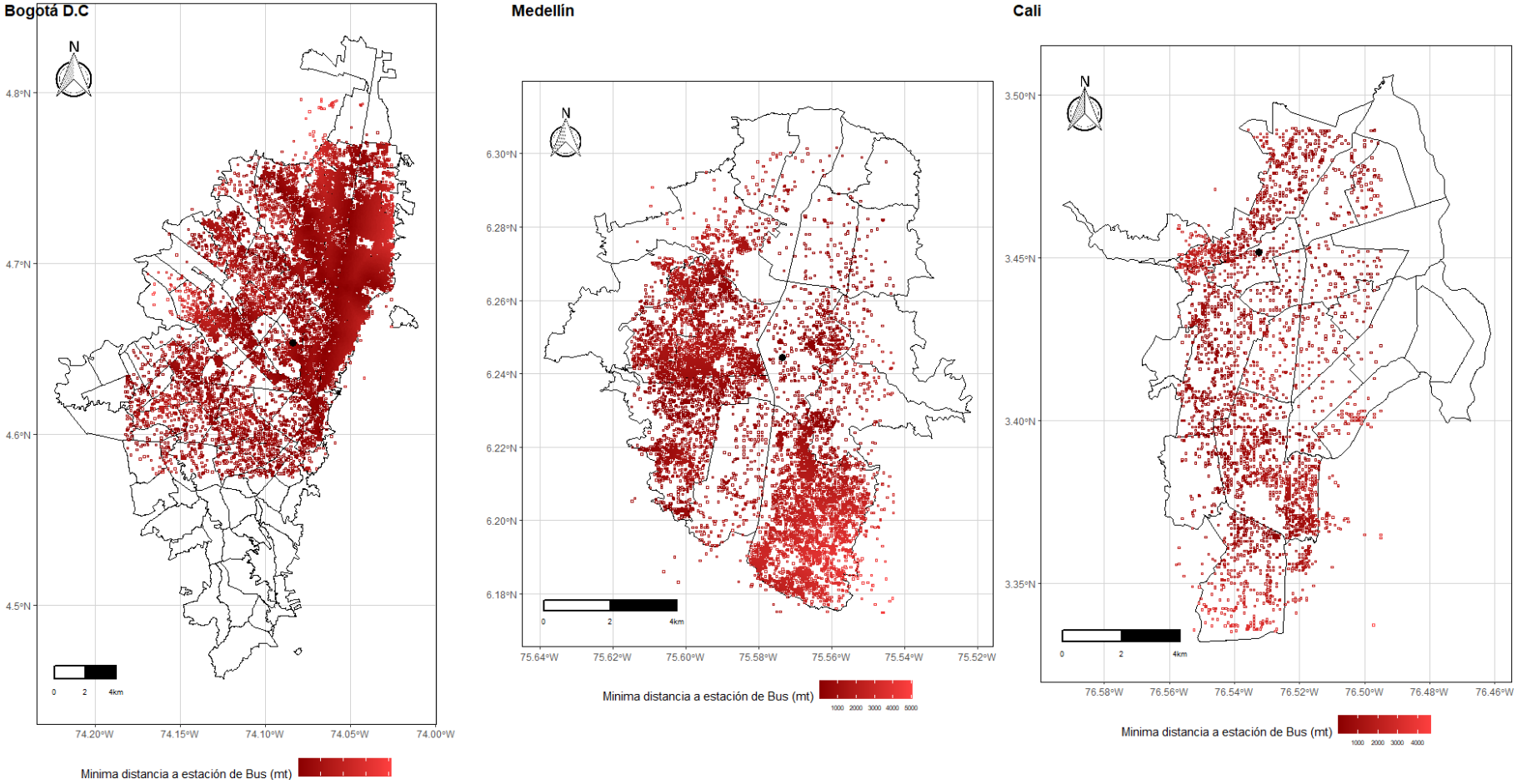


Figura 3. Distancia a la tienda o centro comercial más cercano por ciudad.

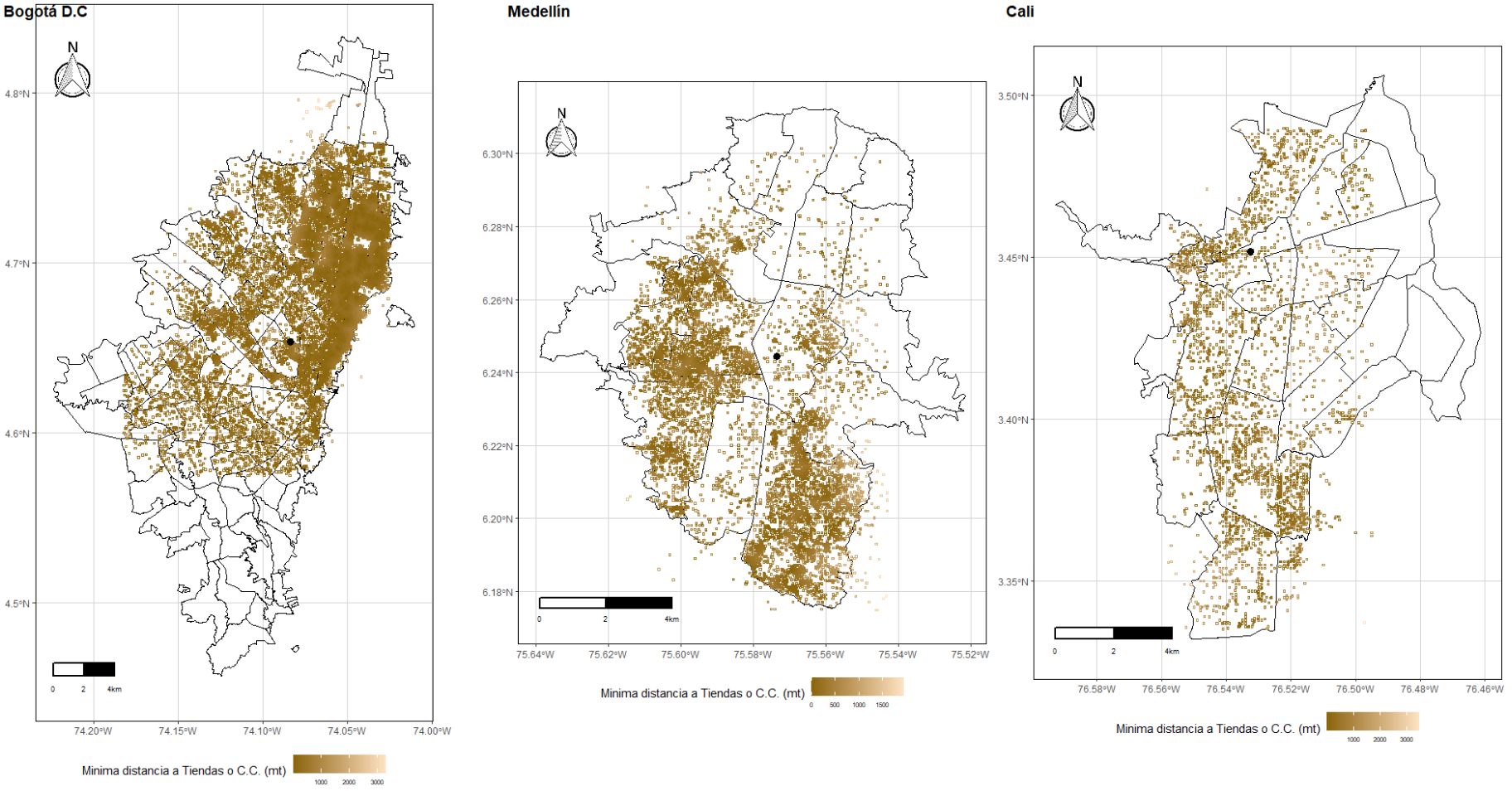
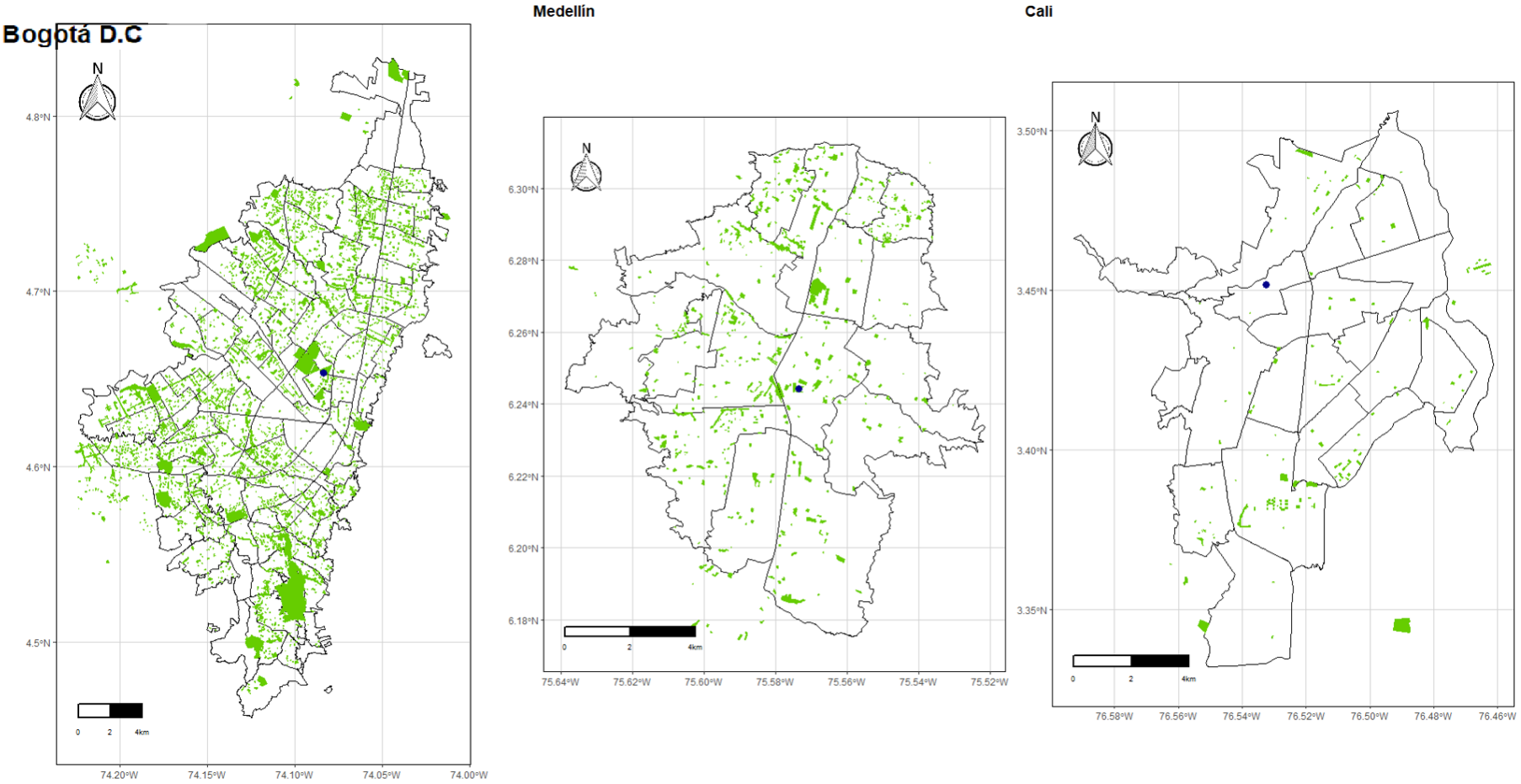


Figura 4. Distancia al parque o zona verde más cercano por ciudad.



Tablas

Tabla 2. Estadísticas descriptivas (sin expresiones regulares).

	Propiedad		Superficie		Habitaciones		Baños	
	Train	*Test*	*Train*	*Test*	*Train*	*Test*	*Train*	*Test*
Obs.	51437	5000	12393	1854	51437	5000	51437	5000
Media	0,78	0,56	221,84	268,92	3,10	3,75	2,65	3,11
Desviación	0,42	0,50	2637,04	913,93	1,36	1,60	1,19	1,43
Mínimo	0	0	15	10	0	0	0	0
Mediana	1	1	124	140	3	3	2	3
Máximo	1	1	198000	26000	11	11	20	13

Tabla 3. Valores faltantes en la base de entrenamiento.

VARIABLE	Cuenta	Porcentaje (%)
Superficie	39044	75,90
Superficie cubierta	41.745	81,15
Baños	15032	48,43
Título (anuncio)	15	0,03
Descripción (anuncio)	36	0,07

Tabla 4. Frecuencia descriptiva (muestra total).

BIGRAMA	Frecuencia
“sala comedor”	26433
“zona ropas”	17185
“cocina integral”	17076
“3 alcobas”	13523
“24 horas”	7317
“baño social”	7082
“zonas verdes”	6501
“2 baños”	6369
“vias acceso”	5980
“centro comercial”	5030

Nota: El valor de la frecuencia reportado hace alusión a las variaciones del conjunto de significado (e.g. sinónimos, expresiones derivadas). Sólo se incluyen los 10 bigramas de mayor frecuencia.

Tabla 5. Frecuencia descriptiva entre el decil inferior *vs.* superior (muestra total).

BIGRAMA	Decil 1	Decil 10
“venta apartamento”	4641	3670
“sala comedor”	3490	1649
“zona lavanderia”	2117	1353
“cocina integral”	2874	778
“3 habitaciones”	1719	1087
“zonas verdes”	1038	615
“baño social”	427	819
“vías acceso”	743	423
“parqueadero visitantes”	587	488
“zona infantil”	459	394
“baño servicio”	N/A	1714
“cada una”	N/A	1056
“comedor independiente”	N/A	701
“comedor auxiliar”	N/A	538
“cocina abierta”	N/A	527
“pisos madera”	N/A	500
“ascensor privado”	N/A	495
“espectacular apartamento”	N/A	444
“sala chimenea”	N/A	435
“4 habitaciones”	N/A	381
“2 baños”	2459	N/A
“24 horas”	1783	N/A
“2 habitaciones”	1744	N/A
“centro comercial”	1442	N/A
“salon social”	1234	N/A
“red gas”	734	N/A
“transporte publico”	651	N/A
“parqueadero cubierto”	530	N/A
“unidad cerrada”	398	N/A
“pisos ceramica”	390	N/A

Nota: El valor de la frecuencia reportado hace alusión a las variaciones del conjunto de significado (e.g. sinónimos, expresiones derivadas). N/A no excluye la expresión ni alude a que su frecuencia sea 0, sólo refiere a que no se encuentra dentro de los 30 bigramas con mayor frecuencia.

Tabla 6. Frecuencia de amenidades extraidas a partir de expresiones regulares (muestra total).

VARIABLE	Cuenta	Porcentaje (%)
Piscina	6642	12,91
Zonas verdes	6954	13,52
Chimenea	9026	17,54
“Espectacular”	5380	10,45
Ascensor privado	758	1,47
Ascensor estándar	10769	20,93
Piso en madera	4959	9,64
Comedor independiente	3944	7,66
Piso en cerámica	2172	4,22
Salón social	6574	12,78
Vigilancia	6842	13,28
Seguridad	5342	10,38
Red a gas	2835	5,51
Centros comerciales	4710	9,15
Transporte público	4154	8,07
Condominio	185	0,35
Campestre	1057	2,05
Penthouse	1044	2,02
Apartaestudio	1791	3,48
Remodelado	5509	10,71
Reformado	153	0,29
“Para estrenar”	1912	3,71
Duplex	3181	6,18
Jacuzzi	2153	4,18
Gimnasio	12106	23,53
“Esquinero”	1976	3,84
Aire acondicionado	102	0,19
Local comercial	2372	4,61
“Vista atractiva”	15734	30,58
Parqueadero privado	24209	47,06
Garaje	231	0,44
Terraza	13173	25,61
Balcón	15915	30,94

Tabla 7. Cercanía promedio (distancia mínima en km) de cada hogar a amenidades en las tres ciudades.

VARIABLE	Bogotá D.C.	Medellín	Cali
Centroide	6,66	5,87	3,83
Aeropuerto	7,71	6,77	2,63
Estación de bus	0,88	1,05	1,53
Hospital	0,81	0,49	0,69
Estación de Policía	0,68	0,86	1,02
Tienda y/o centro comercial	0,27	0,42	0,26
Bar	0,58	1,01	0,94
Universidad	0,87	0,95	1,11
Restaurante	0,25	0,45	0,34
Colegio	0,38	0,53	0,41
Parques (distancia media)	12,55	6,71	6,79
Ríos y canales (distancia media)	27,79	92,74	8,43
Vías principales (distancia media)	13,16	6,79	6,50

Tabla 8. Mediana de atributos demográficos por manzana.

VARIABLE	(estadística)	Bogotá D.C.	Medellín	Cali
Cuartos por vivienda	Mediana	4	4	4
Personas por manzana	Sumatoria	318	241	481
Hogares por vivienda	Mediana	1	1	1
Estrato	Mediana	5	5	5