RANDOM FOREST COMO TÉCNICA DE VALUACIÓN MASIVA DEL VALOR DEL SUELO URBANO: UNA APLICACIÓN PARA LA CIUDAD DE RÍO CUARTO, CÓRDOBA, ARGENTINA.

Random Forest as a mass appraisal technique of the urban land value: an application for the city of Rio Cuarto, Córdoba, Argentina.

Juan Pablo Carranza
Universidad Siglo 21
Secretaría de Investigación

Micael Jeremías Salomón
Universidad Nacional de Córdoba
Facultad de Ciencias Económicas

Mario Andrés Piumetto
Universidad Nacional de Córdoba
Facultad de Ciencias Exactas, Físicas y Naturales, Centro de Estudios Territoriales

Federico Monzani
Universidad Nacional de Córdoba
Facultad de Ciencias Económicas, Instituto de Economía y Finanzas

Marcos Gaspar Montenegro
Universidad Nacional de Córdoba
Facultad de Ciencias Económicas, Instituto de Economía y Finanzas

Mariano Augusto Córdoba
CONICET, Universidad Nacional de Córdoba
Facultad de Ciencias Agropecuarias

Resumen:

El mercado inmobiliario desempeña un papel importante en la economía y la sociedad, por lo tanto, la desactualización de las valuaciones catastrales, en particular del suelo urbano, tiene efectos nocivos sobre las políticas públicas impositivas, territoriales y de vivienda, como en la estabilidad del sistema financiero. Los catastros afrontan el desafío de desarrollar valuaciones masivas de una jurisdicción con el fin de proveer datos actualizados y de calidad, de manera rápida y eficiente. Dado el avance tecnológico, la generación de grandes volúmenes de información y los progresos asociados a las ciencias de la computación, el objetivo del presente artículo consiste en evaluar la capacidad predictiva en la estimación del valor del suelo urbano mediante la aplicación de una técnica algorítmica de aprendizaje automático, conocida como Random Forest, en combinación con una técnica geo-estadística llamada Kriging Ordinario para el tratamiento de los residuos.

Palabras claves: Valor del Suelo, Valuación masiva, Machine Learning, Random Forest, Kriging Ordinario.

Abstract:

The real estate market plays an important role in the economy, therefore, the outdated fiscal value of lands has harmful effects on public policies, tax policies and management of the financial system. For this reason, it is imperative to have updated, good quality and accessible territorial information in order to fulfill the cadastral purposes and successfully face all challanges imposed by the dynamics of urban transformation.

The objective of this paper relies in the evaluation of the predictive capacity of a machine learning technique in estimating the value of urban land. It takes advantage of technological advance as well as large volumes of information which, at first, facilitates the challange of a mass appraisal. The technique used is known as Random Forest, which is combined, for error treatment, with a geo-statistical technique called Ordinary Kriging.

Key words: Land value, Mass appraisal, Machine Learning, Random Forest, Ordinary Kriging.

1. INTRODUCCIÓN.

El mercado inmobiliario desempeña un papel importante en la economía y la sociedad, influyendo en las políticas públicas impositivas, territoriales y de vivienda, la estabilidad del sistema financiero, el empleo y el gasto de los hogares.

La desactualización de las valuaciones del suelo urbano en el Catastro tiene efectos nocivos sobre la equidad del impuesto inmobiliario cobrado por los gobiernos locales, pero también en el desarrollo territorial de las ciudades. Siguiendo a Morales Schechinger (2007), el mercado de suelo está en movimiento constante, existiendo alteraciones estructurales que afectan en la misma magnitud a los precios de todos los terrenos, pero también alteraciones particulares que sólo afectan a terrenos específicos cuando cambian su uso o se densifican. Décadas de alteraciones urbanas no registradas en las valuaciones catastrales generan una estructura de bases imponibles regresivas, que gravan de manera laxa a las áreas urbanas más dinámicas que se han consolidado durante este período (principalmente las que se encuentran hacia la periferia), y de manera relativamente más exigente a áreas urbanas que con el paso del tiempo se han vuelto menos dinámicas (los centros geográficos urbanos típicos de las ciudades monocéntricas, que han perdido atractivo inmobiliario durante las últimas décadas). Esta situación se traduce en una elevada falta de equidad horizontal del sistema tributario local, entendido como una situación en la cual dos contribuyentes con igual capacidad de pago son gravados de manera diferente por el Estado.

Las ciudades latinoamericanas presentan una elevada segregación urbana que se ha potenciado en las últimas dos décadas (Sabatini, 2003), configurando un crecimiento hacia la periferia marcado por “la producción de territorios diferenciales que consolidan formas de vida antitéticas: por un lado, la segregación auto-inducida de los sectores de más altos ingresos y, por el otro, la segregación estructural (por expulsión) de los pobres urbanos” (Cervio, 2015). Una estructura de valores catastrales del suelo que no registre estos movimientos en la dinámica urbana se expresa en un impuesto inmobiliario que grava de igual manera a estos dos universos de contribuyentes que se encuentran segregados en la realidad, dotando al sistema tributario de

una notable falta de equidad vertical (situación en la cual dos contribuyentes de diferente capacidad contributiva son gravados de igual manera por el Estado).

Además, la desactualización de las valuaciones fiscales del suelo urbano tiene un impacto nocivo para la planificación urbana, dado que promueve la especulación inmobiliaria y el aumento general de los precios de la tierra. Siguiendo a Morales Schechinger (2007): "La retención de tierras es un ejemplo de conducta patrimonialista en la que participan todo tipo de propietarios cuando el entorno del mercado es desregulado y desgravado", situación que se traduce en grandes espacios vacantes que, al ser rodeados por la dinámica urbana, cuentan con acceso a múltiples servicios típicamente urbanos. El costo de oportunidad de estos espacios fragmentados es doble: no sólo se pone de manifiesto la contradicción entre zonas de viviendas precarias habitadas por hogares hacinados y grandes áreas urbanas vacantes que suele ser resuelta mediante procesos de ocupación informal de estos espacios, sino que se encarece la provisión de bienes y servicios públicos que deben sortear un espacio vacío para cumplir con su finalidad.

Por otro lado, la actualización del valor del suelo urbano que registran los Catastros es también relevante para el proceso de captura de plusvalías generadas por la inversión pública o la simple acción de los gobiernos locales que en ejercicio de sus potestades generan cambios en los usos del suelo. Según Morales Schechinger (2007), un porcentaje o el total de la renta de suelo puede convertirse en fuente de financiamiento de las ciudades. Por caso, este autor señala el incremento del precio del suelo dado por alguna acción pública a la cual se le puede atribuir ese incremento, como por ejemplo el cambio de patrón de uso, obras de infraestructura y equipamientos, entre otros. Todos estos factores generan aumentos en el valor del suelo por causas ajenas a los propietarios de los lotes, aunque éstos se vean beneficiados. Los esfuerzos por capturar parte de esta valorización se consideran una herramienta fundamental para fortalecer el financiamiento local y así el desarrollo socioeconómico de las ciudades (Reese, 2003).

En este sentido, el conocimiento del valor que el suelo urbano adquiere en el mercado como correlato de, entre otros factores, las transformaciones territoriales sucedidas o aún por suceder, se vuelve un elemento central para la aplicación de instrumentos de gestión y financiamiento urbano. Para el cumplimiento de los fines catastrales y afrontar exitosamente los desafíos que presentan las dinámicas de transformación urbanas, resulta esencial contar con información territorial no sólo de calidad, sino también accesible de manera libre y eficiente.

No obstante, los organismos catastrales presentan datos valuatorios altamente desactualizados. En Argentina, los valores vigentes en los Catastros Provinciales tienen un promedio de 20 años desde su último estudio del mercado inmobiliario (Piumetto, 2016). En parte, esa situación puede explicarse por la dificultad que implica modelar los mercados de suelo con resultados de calidad y presupuestos y plazos exigentes, usando enfoques metodológicos tradicionales y procedimientos con una alta carga manual y artesanal. Quizás la única excepción es la ciudad de Córdoba, con una cultura de actualización de valuaciones que ya reviste el carácter de política pública, con actualizaciones periódicas que registran su antecedente en el revalúo pionero realizado en el año 2008, que tuvo una impacto fiscal considerable y un bajo nivel de conflictividad en su aplicación, generando además herramientas que fueron aplicadas en el diseño de políticas de desarrollo urbano. La metodología aplicada ha sido consistente (se utilizó una técnica geoestadística conocida como Kriging Ordinario), lo que garantizó la coherencia del proceso de valuación masiva.

En este contexto, los modelos de valuación automatizada de inmuebles (*Automated Valuation Model* – AVM) que utilizan modelos estadísticos de última generación se presentan como una alternativa concreta para asegurar la sostenibilidad de una correcta y actualizada

valuación del suelo urbano que incorpore las rápidas transformaciones que se suceden en el territorio y la manera en que éstas se reflejan en los mercados inmobiliarios.

En las últimas décadas, de la mano del avance tecnológico, la generación de grandes volúmenes de información y los progresos asociados a las ciencias de la computación, se han desarrollado cada vez más métodos que pueden ser aplicados eficazmente para la valuación masiva de inmuebles. En los países y catastros que comenzaron a incorporar AVM se observa el uso de técnicas estadísticas clásicas tales como regresión lineal múltiple, regresión lineal con pesos espaciales, técnicas geo-estadísticas como kriging o co-kriging hasta más recientemente, algoritmos de aprendizaje automático como árboles de regresión, k-vecinos cercanos o redes neuronales.

En función a lo anterior, el objetivo del presente artículo consiste en evaluar la capacidad predictiva en la estimación del valor del suelo urbano mediante la aplicación de una técnica algorítmica de aprendizaje automático, conocida como Random Forest, en combinación con una técnica geo-estadística llamada Kriging Ordinario para el tratamiento de los residuos.

2. ÁREA DE ESTUDIO: CIUDAD DE RIO CUARTO.

El área de estudio corresponde a la ciudad de Río Cuarto[1], Provincia de Córdoba, Argentina, (Grafico 1), con una población de 170.000 habitantes y 72,2 km^2 de extensión su área urbana (mancha urbana). Se caracteriza por una estructura territorial e inmobiliaria monocéntrica, con un crecimiento desde su centro a la periferia (característico de las ciudades latinoamericanas), contenida por un anillo de circunvalación conformado por rutas nacionales (N°158, N°8) y una autopista (N°36).

En la base catastral, se registran en total de 67.267 parcelas con valuación urbana y 81.625 cuentas con valuación urbana. Como aproximación del volumen del mercado inmobiliario, considerando datos desde 2014 hasta la fecha, en el área de estudio se producen en promedio 2.207 compra-venta de inmuebles por año, es decir el 2,70% del total de cuentas.

Su centro comercial y administrativo es el espacio por excelencia que reúne un conjunto significativo de atributos urbanísticos (espacios públicos, equipamiento, servicios, infraestructura, mixtura de usos). La ciudad se encuentra dividida en sentido norte-sur por el río Cuarto, atravesado por una serie de puentes que conectan con vías principales hacia el noreste, intensificando estas el desarrollo en los sectores próximos.

[1] Se consideran dentro de la mancha urbana del conglomerado Río Cuarto, las localidades de Rio Cuarto propiamente dicha, Las Higueras y Santa Catalina (Holmberg).

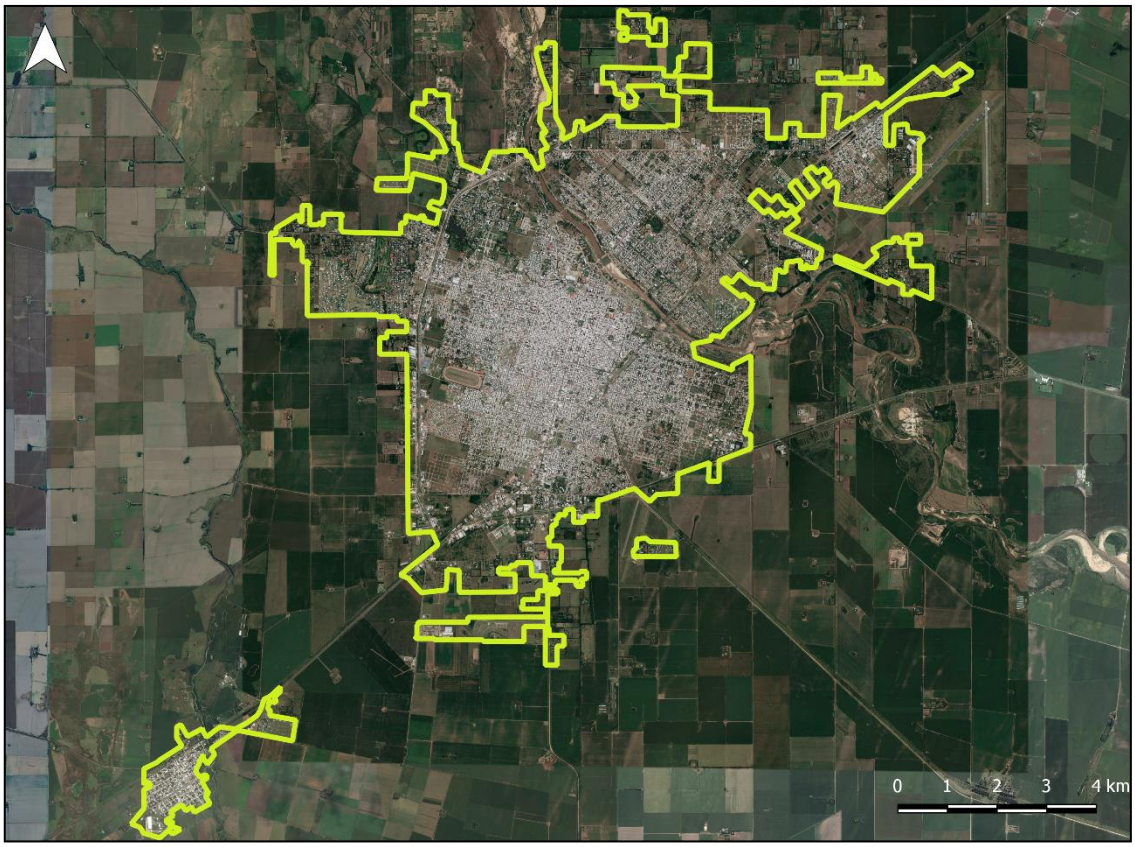
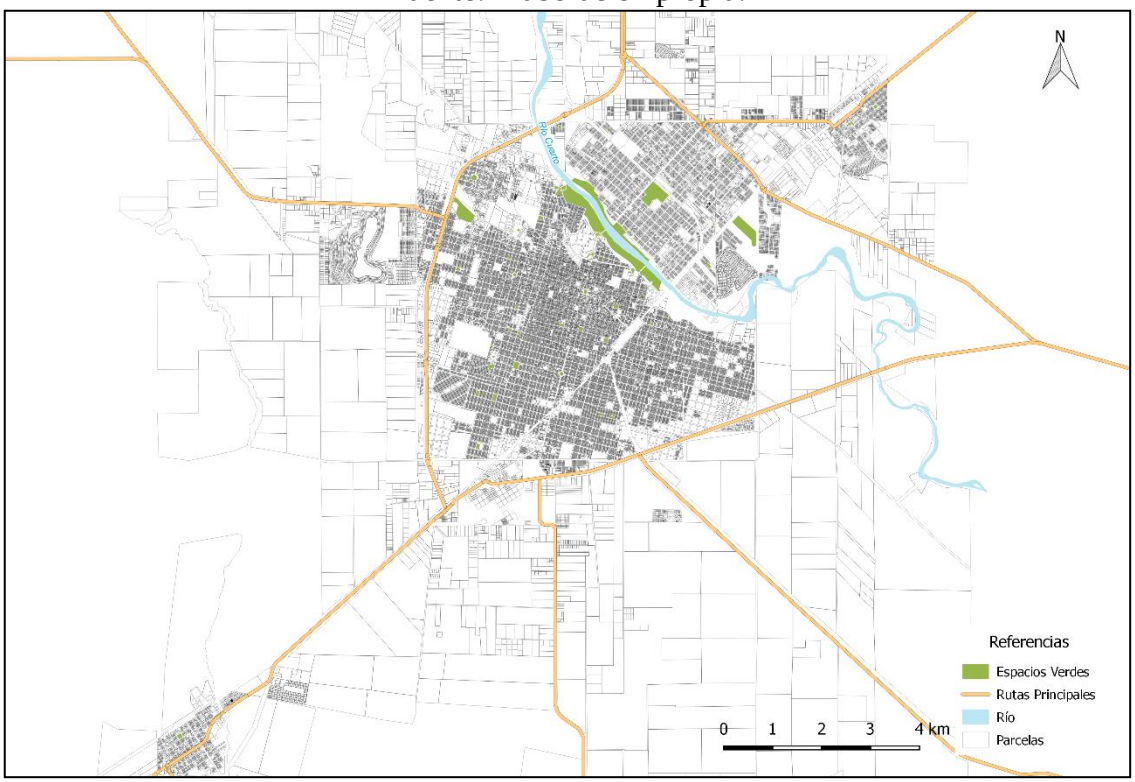


Figura 1 – Mapa de Rio Cuarto
Fuente: Elaboración propia.



Fuente: Elaboración propia.
Figura 2 – Parcelas de Rio Cuarto

3. ENFOQUE METOLODOGICO

Frecuentemente los análisis estadísticos se enfrentan a la exigencia de manipular grandes cantidades de datos complejos y mal estructurados que incluyen un gran número de variables, de las cuales se obtiene información con el propósito de describir una población, identificarla, clasificarla o jerarquizarla, e inferir tendencias, patrones, probabilidades y relaciones subyacentes en dicha estructura de datos.

El árbol de clasificación y regresión (CART) es un método multivariado que permite describir la variable de estudio (*output*), mediante medidas de posición y dispersión (ej. media, varianza), clasificarla y jerarquizarla en función de un conjunto de variables explicativas (*inputs*), e inferir qué valores puede asumir el *output* cuando se desconoce su valor, pero se tiene información relacionada a los inputs utilizados.

El CART constituye una alternativa a los modelos lineales aditivos para los problemas de regresión, en donde la variable de estudio es cuantitativa, y para los modelos logísticos aditivos, en donde la variable de estudio es cualitativa o factorial. CART, en cambio, y los métodos basados en árboles que de él derivan, están pensados para comportamientos no aditivos (anidados). Además, suelen ser de gran utilidad cuando el grupo de variables predictoras o inputs contiene una mezcla de variables cuantitativas y cualitativas.

El CART, y los métodos desarrollados a partir de este algoritmo, se denominan modelos basados en árboles ya que la manera de presentar los resultados es en forma de árbol binario. Cuando el output es cuantitativo se dice que el árbol es de regresión, mientras que si se trabaja con un output cualitativo se referirá a la aplicación de árboles de clasificación.

Dentro de cada árbol de decisión pueden encontrarse los siguientes elementos:

- Nodo de raíz: Representa toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.
- División (Split): Proceso de dividir la población en subconjuntos, llamados subnodos.
- Nodo de decisión: cuando un subnodo se divide en otros subnodos, se llama nodo de decisión.
- Hoja o Nodo final: Cuando finaliza el proceso de discriminación y los nodos no se dividen más, el resultado final es la Hoja o Nodo final.

Siguiendo a Breiman (1984) los CART están compuestos por un conjunto de reglas o procedimientos de particiones binarias recursivas, donde un conjunto de datos es sucesivamente particionado en función de la variable de estudio. En cada división los datos son separados en dos grupos mutuamente excluyentes. En cada instancia de separación el algoritmo analiza todas las variables *inputs* y selecciona, para realizar la partición, aquella que permite conformar dos grupos más homogéneos dentro de sí mismos y más heterogéneos entre sí.

En esta técnica todas las observaciones son consideradas como pertenecientes al mismo grupo. El grupo se separa en dos a partir de una de las variables input, de manera que la heterogeneidad medida sobre la variable output sea mínima dentro de los subnodos generados y máxima entre cada uno de ellos. Para medir la heterogeneidad dentro del grupo, se trabaja con suma de cuadrados corregida por la media $\sum(y_i - \bar{y})^2$. En función de este esquema, cada uno de los dos nodos originados en la primera partición se vuelve a separar nuevamente si:

a) Hay suficiente heterogeneidad para producir una partición de observaciones, y
b) El tamaño del nodo es superior al mínimo establecido para continuar el algoritmo.

El proceso recursivo se detiene cuando no se cumple al menos una de estas dos condiciones.

Siguiendo a Dobra (2002) y Hastie, Tibshirani y Friedman (2008), desarrollar un árbol de regresión clarificará el procedimiento. Dado una base de datos de tamaño N, con p *inputs* (x) y un *output* (y), por cada observación (n) del total N. Esto es, $(x_i, y_i)\ \forall\ i = 1,2, \dots, N$ con $X = (x_{i1}, x_{i2}, \dots, x_{ip})$. El proceso determina cuál es el input divisor (la variable independiente por la cual se genera la subdivisión) y el criterio de división (el valor por el cual el árbol se rige para generar nuevos nodos).

Comenzando desde una partición en M regiones $R_j\ \forall\ j = 1,2, \dots, M$ se busca estimar la variable output como una constante C_{mj}en cada región:

$$f(x) = \sum_{j=1}^{M} C_j\, I \quad \forall\ x\ \in R_j.$$

Si se adopta como criterio de minimización la suma de los cuadrados $\sum\ (y_i - f(x))$, es fácil observar que el mejor valor para C_{mj} es el promedio de y_i en la región R_m:

$$\widehat{C_j} = \frac{\sum_{i=1}^{n} (y_i | x_i)}{n} \forall\ i \in Rj.$$

Para encontrar la mejor partición del árbol en términos de la suma mínima de cuadrados, considerando toda la muestra, siendo k la variable divisora y s el criterio, se define el par de semi-planos:

$$R_1(k,s) = \{X | X_k\ \leq s\ \}\ \ y\ R_2(k,s) = \{X | X_k\ > s\ \}.$$

El objetivo consiste en buscar la variable divisora k y el criterio de partición s que resuelva la siguiente ecuación:

$$\min_{k,s}\left[\min_{c_1} \sum_{x_i\, \epsilon\, R_1\,(k,s)} (y_i - C_1)^2 + \min_{c_2} \sum_{x_i\, \epsilon\, R_2\,(k,s)} (y_i - C_2)^{\,2}\right]$$

Al elegir y minimizar k y s , se obtienen los valores mínimos para C_j correspondientes para cada nodo y así se genera una nueva partición en dos subnodos. Luego de encontrar la mejor partición, se dividen los datos en los dos subconjuntos resultantes y se repite el proceso en cada uno de ellos hasta que no exista suficiente heterogeneidad para continuar el algoritmo o hasta que el tamaño de los nodos sea inferior a un mínimo establecido a priori.

De esta manera, cada hoja de un árbol de regresión contiene aquellas observaciones lo suficientemente similares como para descartar la necesidad de generar nuevos nodos, representando un subconjunto homogéneo en función de los parámetros iniciales fijados en el algoritmo. El promedio del output para los datos de cada nodo se puede tomar como una predicción adecuada de aquellos valores ajenos a la muestra, para los cuales se desconoce el output, pero que tienen valores de inputs similares a los datos del nodo.

La varianza de los datos de cada nodo se puede tomar como una medida de impureza en la estimación. La razón para usar la varianza como medida de impureza se justifica en el hecho que el mejor estimador del *output* en un nodo es el promedio del valor de la variable predicha

en las observaciones que corresponden a dicho nodo. Por lo tanto, la varianza es el error cuadrático medio del promedio utilizado como estimador.

Si bien esta clase de modelos es de fácil interpretación, sobre todo al visualizar la estimación presentada en forma de árbol, se enfrenta a una gran desventaja: cuando el modelo "aprende" en detalle la base de datos de muestra, al generar gran cantidad de nodos sobre los datos de entrenamiento, se produce un impacto negativo en el desempeño del modelo en datos nuevos. Esta deficiencia, conocida como *overfitting* o sobre-ajuste, implica que el modelo presente un excelente ajuste para los datos utilizados en la muestra, como se observa en el Gráfico N°2 a través de la línea azul, pero al momento de predecir nuevos datos lo hace de manera imprecisa producto de una elevada varianza. Esto puede suceder ya que la población no se comporta exactamente igual a la muestra. En caso contrario, puede producirse un proceso de *underfitting* es decir un subajuste, donde el ajuste del modelo es muy suave, y el poder de predicción sigue siendo impreciso. Este resultado se visualiza en la line roja del mismo gráfico. La existencia de estos sesgos hace necesario el desarrollo de modelos que sean robustos, tanto al *overfitting* como al *underfitting*. Es decir, se debe desarrollar un modelo cuyo desempeño se pueda expresar gráficamente sea como la línea verde del Gráfico N°2 para poder generar una predicción efectiva.

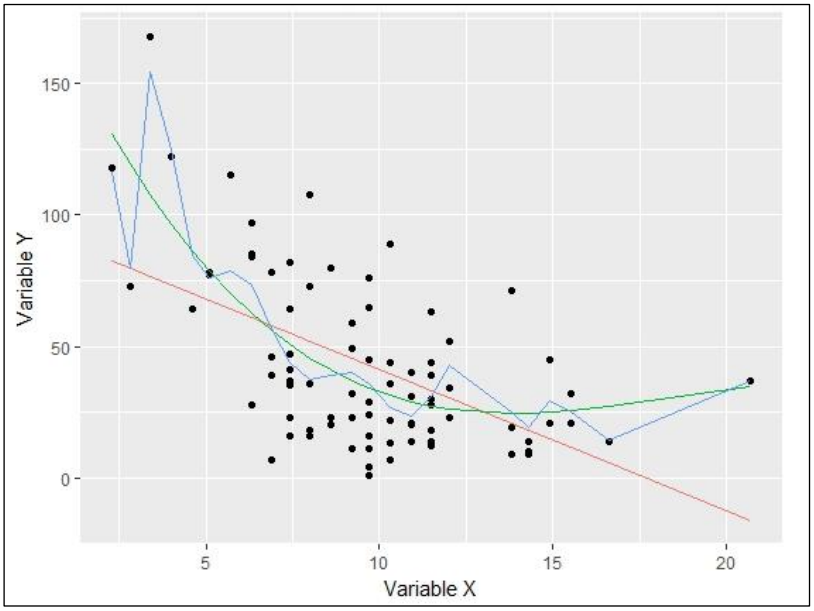


Figura 3 – Underfitting vs Overfitting
Fuente: Elaboración propia en base a Hastie, Tibshirani y Friedman (2008)

Por lo tanto, al elaborar un árbol de regresión se obtiene un modelo con mucha varianza y poco sesgo. Estos problemas se mantendrán independientemente de que se busque aumentar la complejidad de la estructura del CART utilizado. La solución a este problema consiste en generar artificialmente muchos árboles y combinar las predicciones de cada uno de ellos a través de diferentes métodos, para formar un "bosque" que reduzca la varianza y el sesgo implícitos en un solo árbol. Los métodos aplicados para la generación aleatoria de árboles de regresión y la combinación de las predicciones se conocen como *Bagging*, *Boosting*, *Stacking*, siendo el primero el método utilizado en este trabajo.

3.1. Bagging.

Siguiendo a Breiman op. cit., bajo un algoritmo de aprendizaje, la técnica de *Bootstrap Aggretation* consiste en tomar muestras aleatorias de igual tamaño y con reposición del conjunto original de datos de entrenamiento generando diferentes y nuevos datos de entrenamiento. Finalmente se "entrenan" estos datos para calcular la misma cantidad de predicciones y promediar para obtener resultados con mejor ajuste, mitigando el sesgo y la varianza.

3.2. Random Forest.

Random Forest (RF) es una combinación de árboles de decisión, tal que cada árbol depende de los valores de un vector aleatorio, independiente y con la misma distribución para cada uno de estos (Hastie, Tibshirani y Friedman, 2008). Todos los árboles tienen la misma distribución en el bosque (*forest*), pero son forzados a ser diferentes. Esto reduce la correlación. El método combina la idea de *Bagging* de Breiman op. cit. y la selección aleatoria de atributos, con el fin de reducir la correlación entre los árboles.

La idea en Random Forest es mejorar la reducción de la varianza de *Bagging* al reducir la correlación entre los árboles, sin aumentar demasiado la varianza. Esto se logra en el proceso de construcción de árboles mediante la selección aleatoria de *inputs*, específicamente mediante la aplicación de la técnica *bootstrapped* en la base, y dentro de cada nodo de división un subconjunto aleatorio de los inputs. Es decir, "*antes de cada partición, se selecciona m<M de los inputs como candidatos para ser variable de partición*" (Hastie, Tibshirani y Friedman, 2008).

Esquemáticamente, el funcionamiento de algoritmo en la generación de T_i (i=1, …, t) árboles puede representarse de la siguiente manera.

1. Se divide aleatoriamente la muestra en un conjunto de entrenamiento y un conjunto de testeo.
2. Se construye un bosque aleatorio en la base de entrenamiento, en donde cada árbol se construye de la siguiente manera:
   - I. Se toman aleatoriamente "*n*" datos con repetición (*bootstrap*) de la base de entrenamiento.
   - II. Esta nueva muestra será la utilizada para entrenar al árbol *i*.
   - III. Si existen *M* inputs, un número *m* de ellas será seleccionada aleatoriamente para utilizarse en la determinación de la decisión en cada nodo del árbol; *m* debería ser menor que *M*. El valor de m se mantiene constante mientras el bosque se construye.
   - IV. Se iteran todos los posibles valores de cada uno de los inputs seleccionados en *m* a los fines de realizar la mejor partición según los criterios establecidos anteriormente.
   - V. Se continúa con el proceso de partición de los nodos en dos nuevos subnodos hasta que se alcanza el tamaño del nodo deseado, obteniéndose finalmente el árbol *i*.
3. Se predice el conjunto de datos de testeo para evaluar la capacidad predictiva del modelo entrenado anteriormente, procediendo de la siguiente manera:

VI. Cada dato de la base de testeo se somete a los criterios de partición establecidos en cada árbol, desde el nodo raíz hasta las hojas, asignándose a cada uno de estos datos el valor estimado asociado a los nodos terminales.

VII. Este proceso se itera en todos los árboles, para, por último, promediar los valores predichos por cada árbol, siendo este último el valor predicho del bosque.

3.3. Incorporación de la dependencia espacial mediante Kriging Ordinario.

Si bien Random Forest es un método apropiado y ventajoso para hacer predicciones a gran escala considerando una gran cantidad de variables, no tiene en cuenta la dependencia espacial de los datos, algo que es clave al estudiar un sistema complejo con características territoriales. Resulta, entonces, de vital importancia incorporar al análisis algún método complementario que incorpore la dependencia espacial al análisis y predicción del valor del suelo urbano.

Un método adecuado es el conocido como Kriging Ordinario. Siguiendo a Oliver y Webster (2015), se trata de un método de predicción geoestadística esencialmente conformado por una técnica de interpolación lineal. En otras palabras, funciona como una combinación lineal de media móvil ponderada, en la que los pesos dependen del variograma o semivariograma y de la configuración de las observaciones muéstrales dentro del entorno (vecindario).

El método Kriging se basa en el supuesto de que la variable output (valor estimado en una determinada localización) es aleatoria y dependiente del espacio, en donde esta dependencia se expresa como un proceso estacionario con media constante y varianza dependiente de la distancia y dirección.

Suponiendo una función aleatoria Z_{x_i}, siendo cada punto y su localización $x_1, x_2, \dots, x_n$; para N datos, donde Z se distribuye normal con media (estacionaria) , covarianza dependiente de la distancia y dirección representando la variable output (en este caso Valor del Suelo). El Estimador es una combinación lineal ponderada de los datos:

$$z(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i)$$

Donde λ_i son las ponderaciones, que surgen al minimizar la función de semivariograma a través del lagrangeano. Para que el estimador resulte insesgado, la suma de los λ_i debe ser igual a 1. A su vez, x_0 es la localización donde se desea obtener la predicción. A gran escala, un vector X_0 de coordenadas se denominará grilla de predicción.

A través de un semivariograma empírico, que recoge la estructura de dependencia espacial de la variable dependiente, se identifica la función teórica que van estimar los ponderadores λ_i, siendo las más utilizadas - por su forma funcional – la función Esférica, la Gaussiana y la Exponencial. A partir de ello se estiman los parámetros de la función que describe la dependencia espacial, siendo los mismos el rango (distancia en la cual la varianza entre localizaciones deja de crecer), nugget (ordenada al origen) y sill (donde la semivarianza alcanza su valor máximo).

Cuando la tendencia y los residuos son ajustados por separados y luego sumados, se aplica una técnica combinada llamada Kriging con Regresión (Regression Kriging). Es decir, se realiza una estimación del output mediante una regresión lineal, a la cual se suma la interpolación de los residuos realizada mediante un Kriging Ordinario. Esta técnica combinada permite incorporar al análisis multivariado de la regresión lineal la dependencia espacial que ha sido capturada por los residuos del modelo lineal. Las predicciones que resultan de éste método visualmente pueden expresarse de la siguiente forma:

$$Predicción = Tendencia\ predicha\ usando\ regresión + Residuos\ predichos\ usando\ Kriging\ Ordinario$$

El modelo de Regression Kriging supone que los residuos se generan a partir de un proceso aleatorio de estacionalidad de segundo orden normalmente distribuido, esto es un proceso aleatorio que tiene una media y varianza constantes, y una correlación espacial que solo depende de la distancia de separación entre ubicaciones.

Siguiendo a Hengl (2015), para incorporar el análisis de relaciones más complejas entre inputs y entre éstos y el output, se puede reemplazar la regresión lineal por algún algoritmo de aprendizaje automático. En el presente artículo se utiliza el método Random Forest, desarrollado anteriormente, con el objetivo de incorporar al análisis relaciones más complejas entre variables. Por lo tanto, la predicción del suelo urbano se compondrá de la siguiente manera:

$$Predicción = Tendencia\ predicha\ con\ Random\ Forest + Residuos\ predichos\ usando\ Kriging\ Ordinario$$

4. DESCRIPCIÓN DE LA BASE DE DATOS.

Además de las muestras de valores del suelo, construidas en base a los valores de mercado de lotes baldíos urbanos (terrenos), se generó un conjunto de variables que pueden agruparse en tres conjuntos diferentes según sus características:

i) Variables de distancias respecto a: barrios cerrados, barrios populares, centro de la ciudad, espacios verdes, particularidades urbanas como la terminal de ómnibus o la Universidad, grandes superficies comerciales, parques industriales, vías de ferrocarril, vías principales, vías secundarias y zonas de depreciación (cementerio, cárcel, basural, planta cloacal, etc.).
ii) Variables de Servicios e Infraestructura: agua, cloaca, gas y pavimento, que luego se transforman en un índice de servicios.
iii) Variables respecto al entorno: cantidad de lotes urbanos baldíos en relación a la cantidad total de lotes en un radio de 500 metros a cada observación, cantidad de metros cuadrados edificados en relación a la cantidad de metros cuadrados de lotes urbanos en un radio de 500 metros a cada observación, cantidad de transacciones inmobiliarias realizadas durante el último año en un radio de 500 metros a cada observación.

Cada entrada en la base de datos contiene el valor unitario de la tierra (VUT) de mercado, el valor de cada una de las variables descriptas anteriormente y las respectivas coordenadas. Asimismo, se generó una base de datos de predicción, colocando un punto por cada eje de calle (un punto a la mitad de cada cuadra), con la misma estructura e información de cada uno de los inputs de la muestra. Los estadísticos descriptivos de la muestra utilizada pueden encontrarse en el Anexo.

5. PREPARACION MUESTRAL.

La base de datos cuenta con 283 observaciones, provenientes del mercado inmobiliario urbano, distribuidas en la zona urbana consolidad del área de estudio, relevados entre agosto y octubre de 2017. El tamaño muestral representa 0.4% del total de parcelas de la ciudad. Como puede observarse en el Gráfico N°3, cada punto representa una observación y el tamaño hace referencia al valor unitario (VUT).

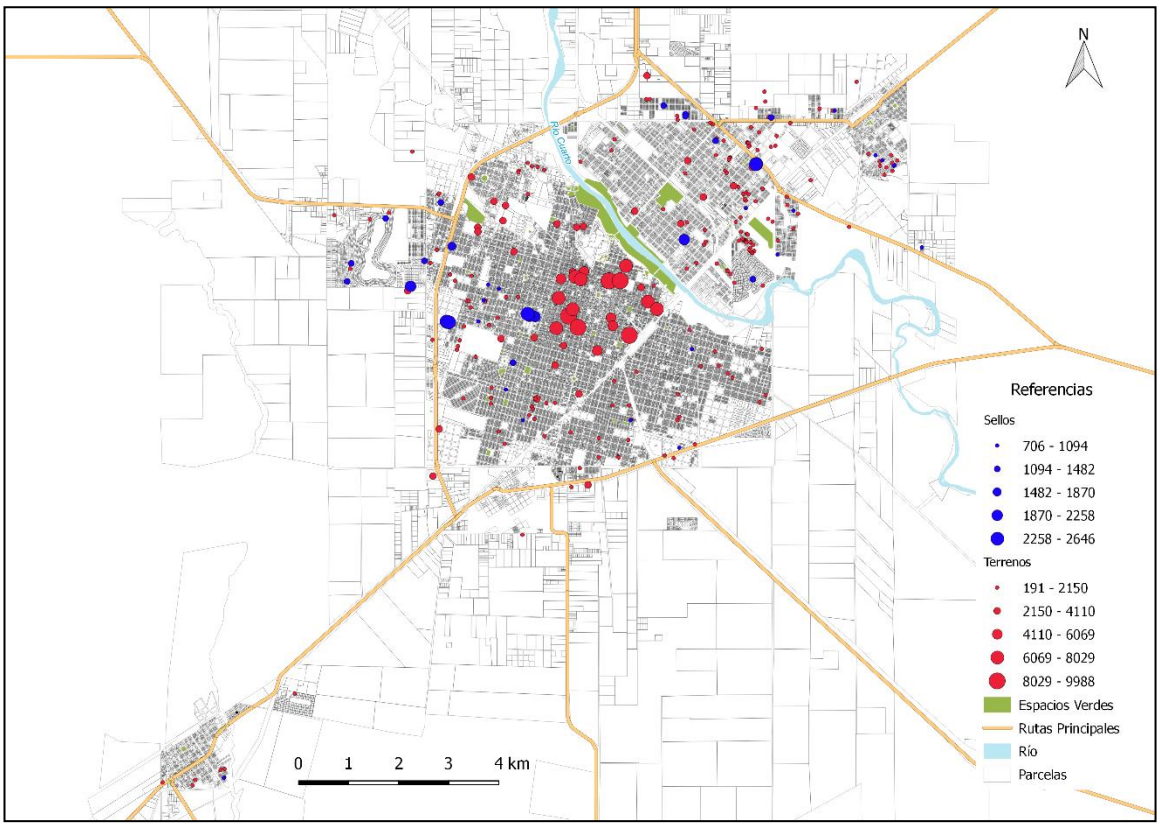


Figura 4 – Mapa de Rio Cuarto con observaciones muéstrales
Fuente: Elaboración propia.

A partir de fuentes como publicaciones web, inmobiliarias, agentes locales, periódicos, se registraron valores de oferta y ventas de inmuebles edificados y baldíos. En el caso de las ofertas se consideró el 87%[2] del valor pretendido para trasladarlo a un probable valor de venta, que se corresponde con un margen de negociación. También se incorporaron datos del Consejo de Tasaciones de la Provincia de Córdoba y determinados casos a partir de la base del Impuesto de Sellos 2017. Así mismo, se realizaron ad-hoc y por parte del equipo de trabajo, tasaciones auxiliares y estimación de valores de terreno a partir de viviendas, vía deducción de mejoras.

La base de datos está compuesta por 197 datos de ofertas y tasaciones, 30 valores estimados a partir de viviendas, 8 tasaciones ad-hoc, 3 remates y 45 ventas declaradas en el

[2] Porcentaje determinado en base a una regresión espacial en donde la variable dependiente es logaritmo natural del valor unitario de la tierra, y las variables independientes son: una variable categórica que informa sobre si la observación es oferta o venta, el rezago espacial de la variable dependiente y el rezago espacial del residuo.

impuesto a las transacciones inmobiliarias[3]. En todos los casos, los valores de la tierra se homogeneizaron a un terreno tipo de 10 m x 30 m, a partir de la aplicación de coeficientes de frente/fondo, superficie y de caso, según normativas y procedimientos del catastro provincial.

El Gráfico N° 4 presenta el histograma y box-plot de la muestra. Como puede apreciarse, la distribución de los VUT está sesgada hacia los valores más bajos.

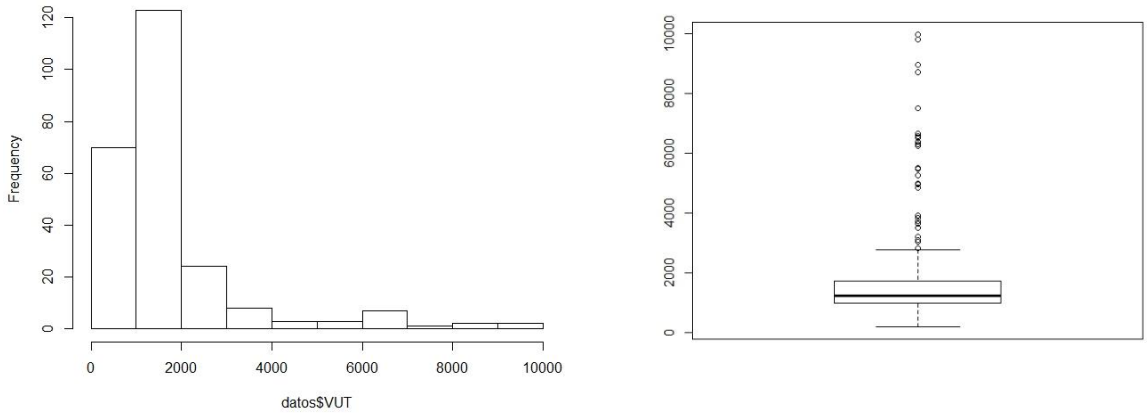


Figura 5 – Box Plot e Histograma del Valor del Suelo (VUT) en Rio Cuarto
Fuente: Elaboración propia.

En este escenario, eliminar valores atípicos o *outliers* aproximaría la distribución del VUT en la muestra a una distribución normal. Si bien esta depuración de datos es un procedimiento estadístico estándar, se eliminarían los datos con VUT más elevados, que se encuentran ubicados en la zona central de la ciudad. Estos casos atípicos, aun siendo extremos, son útiles para el análisis porque reflejan valores efectivamente observados en el mercado inmobiliario, y es importante para el análisis respetar la distribución espacial de las observaciones. En otras palabras, la variable dependiente no puede ser correctamente caracterizada por una distribución normal, lo cual quita capacidad predictiva a los modelos lineales clásicos y otorgan una ventaja relativa considerable a los métodos de aprendizaje automático que no imponen esta clase de supuestos sobre la distribución de la variable.

En función de lo expuesto en el párrafo anterior, se decidió no eliminar observaciones atípicas para respetar la dependencia espacial de la base de datos y para respetar la distribución original de los datos de mercado.

Sin embargo, si es necesario identificar y eliminar los puntos atípicos espaciales o *inliers* (Anselin, 2001). Estos son datos que difieren significativamente de lo observado en su vecindario, identificados de acuerdo a estadísticos elaborados utilizando una matriz de ponderaciones construida a través de la inversa de la distancia (euclidiana) o mediante la matriz de k vecinos más cercanos. Con este objetivo se calculó el índice de Moran local, que refleja el grado de similitud o diferencia entre el valor de cada observación en la muestra relación a sus vecinos. Los vecindarios se construyeron en un entorno de 1.000 metros y se asignó a cada

[3] Dado que existen marcados incentivos a la sub-declaración de las ventas para minimizar el pago del impuesto, estas 45 observaciones se seleccionaron en función de la correlación espacial entre el valor de la venta informado y los valores de mercado detectados en las zonas próximas.

valor una ponderación igual a la inversa de la distancia entre cada par de puntos. Esta técnica permitió identificar 53 observaciones como *inliers*, resultando así la base depurada con 230 datos, cuyas medidas de posición y dispersión pueden apreciarse en la Tabla N°1[4]:

Tabla N°1 – Estadísticas descriptivas del Valor de Suelo (VUT)

Min	Mediana	Media	Max	DS
$ 549	$ 1.834	$ 1.274	$9.988	$1.613,8

Fuente: Elaboración propia.

6. METODOLOGIA DE PROCESAMIENTO.

A los fines de llevar a cabo el análisis de la estimación del VUT para cada cuadra de la ciudad, se utilizará el siguiente procedimiento:

1. Se evalúan y procesan los datos a través de la depuración de datos atípicos (outliers) y/o datos atípicos en su entorno (outliers espaciales o inliers).
2. A cada punto de la muestra y las localizaciones a predecir, se le asignan las variables independientes (de distancia, de servicios e infraestructura y de entorno).
3. Se utiliza el algoritmo de aprendizaje automático Random forest y se estima el modelo con la base de datos para obtener predicciones del VUT en todas las localizaciones, tanto las muestras como en los puntos medios de cuadra.
4. Se determinan los residuos de la estimación a partir de los valores predichos y los valores observados en cada una de las muestras
5. Se aplica sobre los residuos el método Kriging Ordinario, ajustando el semivariograma correspondiente.
6. Finalmente, a la predicción original obtenida en cada una de los puntos medios de cuadra, mediante el algoritmo Random Forest, se suma la interpolación de los residuos obtenida mediante la técnica geoestadística Kriging Ordinario.

7. RESULTADOS.

7.1. Random Forest.

Se procede a generar un modelo predictivo mediante el algoritmo de aprendizaje automático de Random Forest. Este método genera un número finito de árboles y, a través del proceso de ensamblado *Bagging*, un promedio simple en los resultados que producirá las estimaciones pertinentes.

Para aplicar el método se crearon 500 árboles de regresión independientes, aunque según el Gráfico N°9 con sólo 200 árboles ya se hubiese logrado minimizar el error y estabilizar el modelo. Aun así, la incorporación de más árboles de los óptimos no genera problemas de *overfitting* ya que cada uno de ellos es independiente del resto, sino que aumenta los tiempos computacionales.

[4] Valores en pesos argentinos para el periodo agosto – octubre de 2017, contemplados a un tipo de cambio $17.5 por dólar estadunidense.

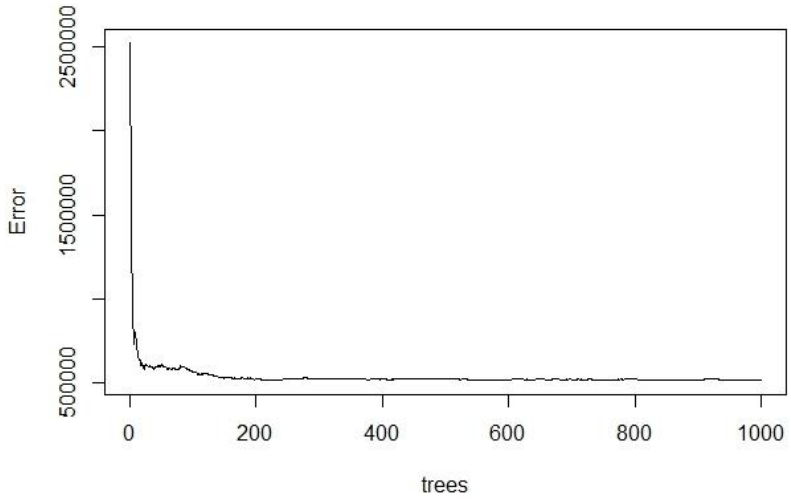


Figura 6 – Evolución del error en base a la cantidad de Arboles generados.
Fuente: Elaboración propia.

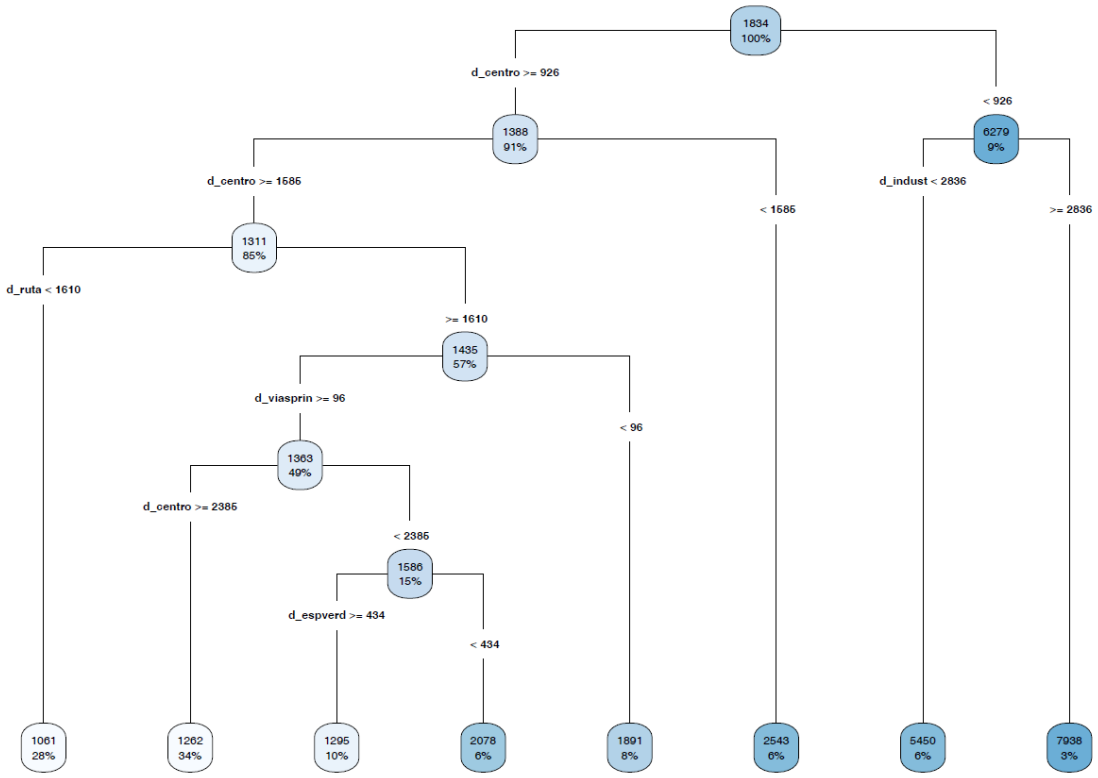


Figura 7 – Árbol de Regresión respecto a observaciones del Rio Cuarto
Fuente: Elaboración propia.

Del proceso de aprendizaje estadístico del método, mediante un procedimiento de validación cruzada con 5 grupos, surgió que el subconjunto óptimo de inputs a considerar en cada nodo es igual a 13, de un total de 27.

Una de las principales desventajas de los métodos basados en árboles es que no permiten cuantificar en una forma funcional la relación de cada variable input con el output bajo estudio, dada la gran cantidad de árboles aleatoriamente generados. Sin embargo, sólo a fines expositivos, se puede extraer del bosque de regresión un árbol aleatorio para analizar su composición, tal como se presenta en el Gráfico N°5:

La variable más influyente, en primera instancia, para subdividir las observaciones es "d_centro" que es la distancia que existe entre el punto observado y el área céntrica de la ciudad. Luego, en segunda instancia la variable divisora del nodo por un lado vuelve a ser "d_centro" y por el otro "d_indust" que corresponde a la distancia total hacia la zona industrial de la ciudad.

En tercera instancia, la distancia a la ruta toma relevancia como punto de referencia. Es interesante observar que en las hojas o nodo final del árbol aleatoriamente extraído se encuentra el VUT estimado para cada grupo de observaciones que cumplen con los criterios de partición previos.

Si bien, como se aclaró anteriormente, el método Random Forest no permite cuantificar una forma funcional que indique la relación de cada variable *input* con el *output* analizado, sí se puede identificar el aporte de cada *input* sobre la calidad predictiva del modelo. De esta manera, en el Gráfico N°6 puede apreciarse que, analizando los 500 árboles utilizados en la estimación, la importancia de las variables según su aporte a la reducción de la suma del error cuadrático medio[5] del modelo ratifica lo observado en el árbol aleatoriamente extraído en el Gráfico N°5. Además, variables que tienen encuentra el entorno y su nivel de desarrollo cobraron más relevancia a través del algoritmo (el detalle de las variables utilizadas, su significado y los estadísticos descriptivos más relevantes se pueden encontrar en el Anexo).

7.2. Kriging Ordinario.

Una vez obtenida la predicción inicial mediante la aplicación del algoritmo Random Forest, se determinan los residuos como la diferencia entre el valor observado del *output* y el valor predicho por el modelo.

Posteriormente, se procede a aplicar la técnica Kriging Ordinario para la interpolación espacial del valor de los residuos calculados. Para ello, es necesario determinar previamente el semivariograma teórico adecuado, que capture de manera más eficiente la dependencia espacial de los residuos de la estimación. En base a los datos muestrales, el modelo que mejor resuelve esta regresión no lineal es el exponencial, tal como puede apreciarse en el Grafico N°7.

[5] MSE por sus siglas en ingles, Mean Square Error. Siendo en el grafico - %IncMSE – incremento porcentual del error cuadrático medio.

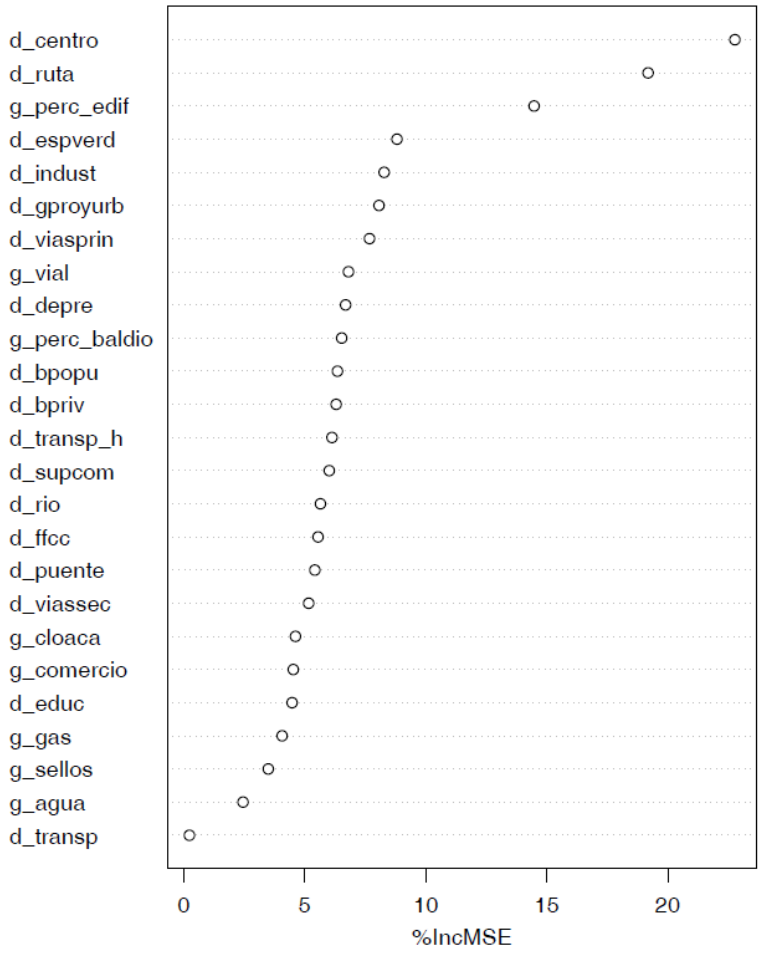


Figura 8 – Reducción del Error global por participación de las variables input
Fuente: Elaboración propia.

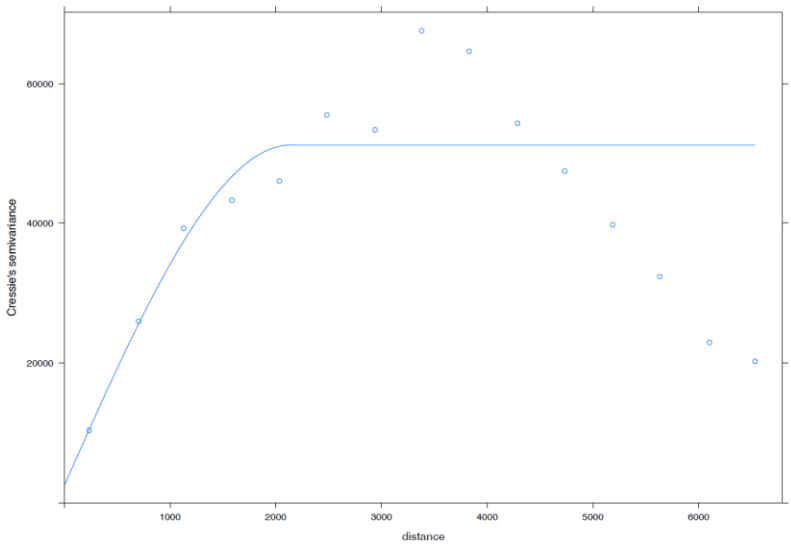


Figura 9 – Semivariograma teórico
Fuente: Elaboración propia.

7.3. Predicción Final.

Para la predicción final, se utiliza la base de predicción de Rio Cuarto. A partir de esto se genera la predicción del valor del suelo (VUT) para cada punto por cada eje de calle en base al modelo Random Forest. De la misma manera, se aplica la técnica Kriging Ordinario configurada para los residuos del modelo con el fin de predecir en cada punto correspondiente a los ejes. Estos resultados, al sumarse, se obtiene la predicción completa del valor del suelo en la Ciudad.

El Gráfico N°8 muestra el mapa de valores del suelo urbano calculado de la manera indicada en el párrafo anterior, siendo el color rojo los valores más elevados y el celeste los valores más bajos.

8. CALIDAD DE LOS RESULTADOS OBTENIDOS.

Para analizar la calidad predictiva del modelo aplicado se procede a realizar un proceso de validación cruzada. Esta técnica consiste en dividir de manera iterativa la base datos en dos subconjuntos: uno de entrenamiento, que se utiliza para entrenar el modelo, y otro de prueba, cuya finalidad es realizar la validación de los resultados obtenidos. Es decir, se genera una predicción con los datos de pruebas, el error predicho se calcula con el conjunto de datos reservados para realizar la validación.

Cuando el proceso se repite tomando distintos conjuntos aleatorios de datos de entrenamiento un número k de veces, sin reemplazo, utilizando cada subconjunto para validar el modelo entrenado con los otros k-1 subconjuntos, se denomina la técnica de validación cruzada k-fold. En el presente artículo se utilizará un $k = 10$.

Asimismo, como medidas calidad se tendrán en cuenta los siguientes estadísticos, que constituyen el estándar en la literatura estadística:

- Mean absolute percentaje error (MAPE):

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

- Median absolute percentaje error (MeAPE):

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

En donde:

$\hat{y}_i$ es el valor estimado o predicho,

y_i es el valor observado,

n es el tamaño de la muestra.

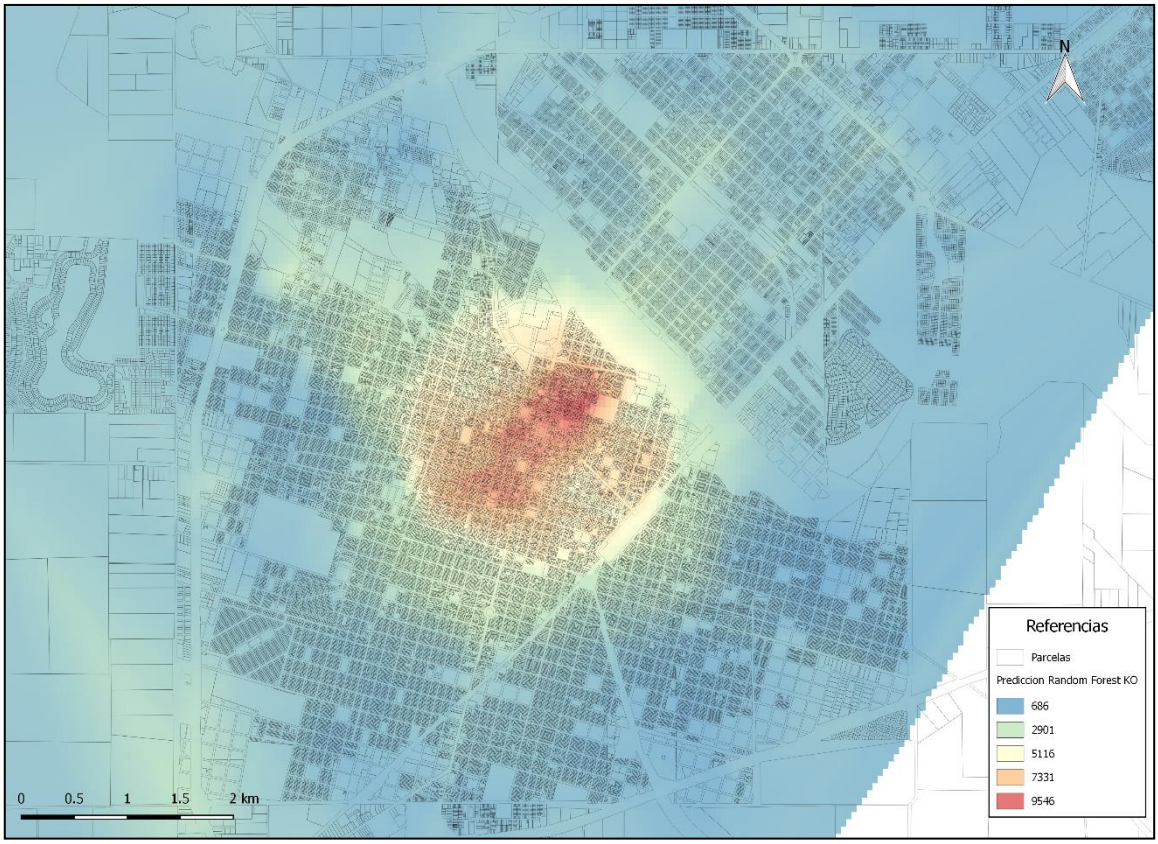


Figura 10 – Mapas de valor del suelo urbano en la ciudad de Rio Cuarto.
Fuente: Elaboración propia.

Siguiendo el organismo International Association of Assessing Officers (IAAO) abocado en la investigación en valuación de las propiedades, la administración y la política impuestos a la propiedad, aconseja para la comparación sobre el nivel de valuación, los siguientes parámetros, para medir la homogeneidad entre los valores predichos y las observaciones de mercado:

- Coeficiente de Variación (CV) en base al ratio entre el valor predicho y observado:

$$CV = \frac{\frac{\hat{y}_i}{y_i} - \text{media}(\frac{\hat{y}_i}{y_i})}{\text{media}(\frac{\hat{y}_i}{y_i})}$$

- Coeficiente de dispersión (CD) en base al ratio entre el valor predicho y observado:

$$CD = \frac{\frac{\hat{y}_i}{y_i} - \text{mediana}(\frac{\hat{y}_i}{y_i})}{\text{mediana}(\frac{\hat{y}_i}{y_i})}$$

Como puede apreciarse en la Tabla N°2, el error esperado de la predicción (en valor absoluto) es, en promedio, igual al 19% (aunque la mediana es igual al 13%). Los coeficientes calculados para evaluar la homogeneidad de la estimación en relación a los datos de mercado permanecen por debajo del 20%, en línea con lo estipulado por el IAAO (2003).

Tabla N°2 – Comparación de métodos predictivos

Modelo	**MAPE**	**MeAPE**	**CV**	**CD**
RF-KO	0.19	0.13	0.18	0.19

Fuente: Elaboración propia.

9. CONSIDERACIONES FINALES.

El objetivo del presente artículo consistió en evaluar la capacidad predictiva de una técnica algorítmica de aprendizaje automático para la estimación del valor del suelo urbano. Utilizando el avance tecnológico y la generación de grandes volúmenes de información, se partió de la hipótesis que la aplicación de este tipo de técnicas mostraría un desempeño más adecuado que las técnicas estadísticas y geo-estadísticas tradicionales.

Para ello, se procedió con la utilización de la técnica algorítmica conocida como Random Forest, la cual se combina, para el tratamiento de los residuos, con la técnica geo-estadística Kriging Ordinario. Partiendo de una base muestral y un campo objetivo como la Ciudad de Rio Cuarto, Provincia de Córdoba, se comprobó que la capacidad predictiva del algoritmo es altamente significativa.

Los resultados obtenidos se enmarcan dentro de los parámetros de calidad establecidos por el IAAO. El error relativo promedio en valor absoluto fue igual al 19%, mientras que la mediana de esta misma medida ascendió al 13%. En cuanto a la homogeneidad de los errores obtenidos, el coeficiente de variación fue igual a 0,18, siendo el coeficiente de dispersión igual a 0,19.

En función de los resultados obtenidos, se observa que la utilización de métodos de aprendizaje automáticos reduce en gran manera los tiempos que conllevan una valuación masiva, lo cual simplifica el proceso de actualización del valor del suelo frente a las constantes alteraciones estructurales que afectan los precios de todos los terrenos. Otra ventaja que provee este método es dotar al sistema tributario de una herramienta con alto respaldo estadístico que contribuya a la equidad del sistema fiscal, como también al área a los procesos de gestión urbana.

AGRADECIMIENTOS.

La información utilizada en el presente artículo fue generada en el marco del **Estudio Territorial Inmobiliario de la Provincia de Córdoba, Argentina,** financiado en conjunto por el Programa de Naciones Unidad para el desarrollo (PNUD) y el gobierno provincial. El proyecto, coordinado por la Secretaria de Ingresos Públicos y la Dirección General de Catastro, ambas dependientes del Ministerio de Finanzas, tiene por objetivo de actualizar las valuaciones catastrales de más de 2 millones de inmuebles urbanos y rurales, en una extensión de 165.000 km2; así mismo, modernizar los procesos de actualización, brindando un marco apropiado y sustentable de información y herramientas para la gestión de políticas territoriales. Entre las estrategias implementadas, se conformó un equipo de trabajo multidisciplinario de alto nivel y se desarrolló un Observatorio del Mercado Inmobiliario (OMI) que a julio de 2018 cuenta con más de 7.000 datos georreferenciados.

REFERENCIAS BIBLIOGRÁFICAS

ANSELIN, L. (1998). **GIS research infrastructure for spatial analysis of real estate markets.** Journal of Housing Research, 9, 113–133.

ANTIPOV E.; POKRYSHEVSKAYA, E. (2012). **Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics**. Expert System with Applications, 39, 1772-1778.

BREIMAN, L. (2001). **Random forests**. Machine Learning, 45, N°1. 5–32.

BREIMAN, L.; FRIEDMAN, J.; STONE, C.; OLSHEN, R. (1984). **Classification and regression trees.** California, Wadsworth, Inc.

HUANG, B.; WU, B.; BARRY, M. (2010) **Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices,** International Journal of Geographical Information Science, 24, N°3, 383-401.

CERVIO, A. L. (2015). **Expansión urbana y segregación socio-espacial en la ciudad de Córdoba (Argentina) durante los años '80.** Astrolabio,14.

HENGL T.; HEUVELINK G.; KEMPEN B.; LEENAARS J.; WALSH M.; SHEPHERD K. (2015). **Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions**. PLoS ONE, 10, N°6.

INTERNATIONAL ASSOCIATION OF ASSESSING OFFICERS (2003). **Standard on automated valuation.**

TOBLER, W. (1970). **A computer movie simulating urban growth in the Detroit region.** Economic Geography, 46, N° 2. 234-240.

JEREMY, M. (2006). **Mapping the Results of Geographically Weighted Regression.The Cartographic Journal.** 43, N°2. 171-179

JIAN, G. ; SHI, D. ; ZURADA, J. ; LEVITAN,A.; (2014) **Analyzing Massive Data Sets: An Adaptive Fuzzy Neural Approach for Prediction, with a Real Estate Illustration.** Journal of Organizational Computing and Electronic Commerce, 24, N°1. 94-112.

LOCKWOOD, T. ; ROSSINI, P.(2011) **Efficacy in Modelling Location Valuation models (AVMs). Within the Mass Appraisal Process.** Pacific Rim Property Research Journal, 17, N°3. 418-442.

MORALES SCHECHINGER, C. (2007). **Algunas reflexiones sobre el mercado de suelo urbano". Mercados de suelo urbano en las ciudades latinoamericanas**. Lincoln Institute of Land Policy (ed.).

PÉREZ-PLANELLS,L. ; DELEGIDO, J.; ET ALL. (2015). **Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos.** Revista de teledetección, 44. 55-65.

PIUMETTO, M. 2016. **Diagnósticos catastros provinciales e impuesto inmobiliario,** en Proyecto Modernización de los Sistemas de Gestión Financiera Pública Provincial, Argentina. BID, Ministerio del Interior, IERAL de Fundación Mediterránea (sin publicar).

QINGMIN, M. (2014). **Regression Kriging versus Geographically Weighted Regression for Spatial Interpolation**. International Journal of Advanced Remote Sensing and GIS, 3, N°1. 606-615.

REESE, E. (2003). **Instrumentos de gestión urbana, fortalecimiento del rol del municipio y desarrollo con equidad** - Lincoln Institute of land policy (Ed.).

SABATINI, F. (2003). **La segregación social del espacio en las ciudades de América Latina,** BID: Desarrollo Social. Documento de Estrategia. Washington DC.

ANEXO.

Tabla N°3 – Descripción de variables inputs

Variables Inputs	Definición	Tipo	Min	Max	Media	Mediana	DS
d_bpriv	Distancia Barrios privados	Continua	0	4546	1802	1637.6	1419.86
d_centro	Distancia Centro	Continua	0	6203	1487	1341.5	1094.37
d_educ	Distancia Centros educativos	Continua	39.17	2200.98	671.52	486.97	544.55
d_espverd	Distancia espacios verdes	Continua	21.72	938.35	348.33	338.8	198.58
d_ffcc	Distancia Ferrocarriles	Continua	54.01	2003.53	718.63	633.22	438.63
d_supcom	Distancia Centro Comerciarles	Continua	308.3	6239.9	2570.2	2647.5	1200.57
d_indust	Distancia a Centro Industrial	Continua	1133	7274	2987	3039	1101.6
d_ruta	Distancia Rutas	Continua	11.73	5325.46	2851.51	3100.02	1447.45
d_viasprin	Distancia Vías Principales	Continua	11.58	4962.03	856.4	663.15	815.16
d_viassec	Distancia Vías Secundarias	Continua	2.19	1262.85	279.68	200.19	270.12
d_depre	Distancia a zona de depreciación	Continua	840.7	7013.4	2667.2	2643.5	1072.98
d_front	Distancia a limite frontera	Continua	136.2	6318.5	2775.6	2516.6	1568.63
d_transp	Distancia a servicio transporte publico	Continua	0.93	9360	602.44	158.43	1529.62
d_gproyurb	Distancia a proyectos urbanos	Continua	143.2	13451.3	2682	2291	2147.86
d_bpopu	Distancia barrios populares	Continua	122.8	10904	2060	1660.3	1843.02
d_puente	Distancia a puente principal	Continua	360	14234.3	2902	2552.8	2247.09
d_rio	Distancia al rio	Continua	235.5	14208.6	2601	2382.9	2253.74
g_comercio	Densidad de comercio en entorno	Continua	0	60	3.26	0	6.72
g_perc_baldio	% Baldío en Entorno	Continua	0.01	0.846	0.33	0.37	0.24
g_perc_edif	% Edificado en entorno	Continua	0.001	1.06	0.34	0.25	0.27
g_sellos	Cantidad de transferencias	Continua	0	0.06	0.01	0.01	0.01
d_baja	Distancia a zonas de bajo valor	Continua	132.4	6318.5	1778.2	1659.4	1059.57
d_sube	Distancia a zonas de alto valor	Continua	0	6203.9	1487.5	1341.5	1094.37
g_agua	% de acceso al agua por manzana	Continua	0	1	0.83	1	0.3
g_cloaca	% de acceso al cloacas por manzana	Continua	0	1	0.68	1	0.42
g_vial	% de calle pavimentada por manzana	Continua	0	1	0.3	0	0.4
g_gas	% de acceso al gas por manzana	Continua	0	1	0.64	0.77	0.39

Fuente: Elaboración propia.