VALUACIÓN MASIVA DE LA TIERRA URBANA MEDIANTE INTELIGENCIA ARTIFICIAL. EL CASO DE LA CIUDAD DE SAN FRANCISCO, CÓRDOBA, ARGENTINA[1]

MASS APPRAISAL OF URBAN LAND VALUE USING ARTIFICIAL INTELLIGENCE. THE CASE OF SAN FRANCISCO CITY, CÓRDOBA, ARGENTINA

Juan Pablo Carranza[2]; Mario Andrés Piumetto[3]; Micael Jeremías Salomón[4]; Federico Monzani[5]; Marcos Gaspar Montenegro[6]; Mariano Augusto Córdoba[7]

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. Por tal motivo, 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. Los resultados obtenidos permiten resaltar la ventaja de 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 frente a un método econométrico clásico, regresión lineal.

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

Abstract:

The real estate market plays an important role in the economy and society, therefore, the downgrading of cadastral valuations, particularly urban land, has harmful effects on tax, territorial and housing public policies, property market, as in the stability of the finance system. For this reason, the cadastres face the challenge of developing massive valuations of a jurisdiction in order to provide updated and quality data, quickly and efficiently. Given the technological advance, the generation of large volumes of information and the progress associated with computer science, the ideas of massive appraisal of real estate by the catastres is increasingly taking hold. Under these needs and new situation, the results reflects the advantage of the predictive capacity in estimating the value of urban land by applying an algorithmic technique of machine learning, known as Random Forest, in combination with a geo-statistical technique called Ordinary Kriging for the treatment of error.

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

[1] El Artículo forma parte de la ponencia presentada en la 51 Jornada Internacionales de Finanzas Públicas. FCE-UNC, 2018.

[2] Magister en Políticas Públicas. Lic. en economía. Universidad Siglo 21. Secretaría de Investigación.

[3] Agrimensor. Universidad Nacional de Córdoba. Facultad de Ciencias Exactas, Físicas y Naturales, Centro de Estudios Territoriales

[4] Lic. en Economía. Universidad Nacional de Córdoba. Facultad de Ciencias Económicas

[5] Magister en Estadísticas. Lic. en economía. Universidad Nacional de Córdoba. Facultad de Ciencias Económicas, Instituto de Economía y Finanzas.

[6] Lic. en Economía. Universidad Nacional de Córdoba. Facultad de Ciencias Económicas, Instituto de Economía y Finanzas

[7] Dr. en Agronomía. Investigador CONICET, Universidad Nacional de Córdoba. Facultad de Ciencias Agropecuarias

1. Introducción

La desactualización de las valuaciones del suelo urbano 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 evolución urbana no registrada en las valuaciones catastrales generan una estructura de bases imponibles regresivas, que gravan de manera laxa las áreas urbanas más dinámicas 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 es también relevante para el proceso de captura de externalidades 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 el desarrollo socioeconómico de las ciudades (Reese, 2003).

Parte de este tipo de externalidades, aquellas que no descansan en decisiones administrativas de cambios en uso del suelo, se generan a través de la inversión real directa (IRD) que comprende un componente del prepuesto público, y abarca todas las erogaciones destinadas a la construcción de bienes de capital, como son las edificaciones, instalaciones, obras de transporte, vivienda, urbanismo, agua potable, alcantarillado y otros servicios urbanos. En la Provincia de Córdoba, sólo en 2017, la IRD fue de $ 21.804 millones. Siendo estas inversiones políticas públicas que impactan directamente en la valorización del suelo urbano.

Los valores catastrales de la tierra vigentes en la Provincia de Córdoba tienen una importante desactualización acumulada desde hace décadas[8]. Esta situación se materializa en valores fiscales que promedian un 10%[9] del valor de mercado de los inmuebles[10]. Esta situación representa una evidente pérdida de recursos para los gobiernos locales, no sólo en la recaudación potencial del impuesto inmobiliario, sino en la no captura de las externalidades generadas por la inversión pública. Serra, et al. (2005) detallan en su estudio realizado en tres importantes ciudades de Brasil (Brasilia, Curitiba y Recife) que la ejecución de obras viales, pavimento y agua potable pueden valorizar, en promedio y según la zona de la ciudad, hasta 4 veces la inversión realizada. Lo anterior sugiere que la inversión del Estado en el territorio no se estaría traduciendo en la actualización de las valuaciones fiscales como un instrumento para recuperar parte de la inversión realizada. Esta situación representa una fuente de inequidad en el gasto de capital realizado por los gobiernos locales, que pierden una excelente oportunidad para financiar el desarrollo local mediante el fortalecimiento del impuesto inmobiliario y la contribución por mejoras, lo cual requiere de la actualización de los catastros y la valuación permanente de los inmuebles, en especial la tierra.

Bonet, et al. (2014), en su capítulo para Argentina señala *"… en la búsqueda de mayores recursos fiscales, los gobiernos subnacionales podrían beneficiarse de una fuente aparentemente subutilizada y en franco declive: el impuesto a la propiedad inmobiliaria*" y destacan varios aspectos positivos, entre ellos que es menos distorsivo, potencialmente más progresivo y menos procíclico que otros ingresos fiscales. Por otra parte, también pueden indicarse como fortalezas: (i) que resulta un impuesto de larga tradición, por lo que existe un grado razonable de familiaridad, aceptación y conocimiento entre los ciudadanos, (ii) su capacidad extra fiscal como instrumento regulador de los mercados de suelo y con capacidad para influir en el uso del suelo, y (iii) la oportunidad para fortalecer la autonomía local (De Cesare, 2012).

Por lo tanto, la interrelación de instrumentos como el impuesto inmobiliario y un catastro moderno y actualizado es virtuosa y genera oportunidades para el financiamiento del desarrollo y para la mejora en la gestión de las políticas territoriales. Con este objetivo, el Gobierno de la Provincia decidió llevar adelante en 2017 un ambicioso proyecto de revalúo inmobiliario que está comenzando a dar sus primeros resultados. El presente trabajo de investigación se enmarca sólo

[8] El último revalúo urbana data de 1987, con un revalúo parcial realizado en 1994 que alcanzó sólo a 19 localidades.

[9] La mediana de este valor se ubica en un 4%.

[10] Esta información fue provista por el proyecto "Estudio territorial inmobiliario de la Provincia de Córdoba", llevado adelante por la Secretaría de Ingresos Públicos y la Dirección General de Catastro de la Provincia de Córdoba, y del cual forman parte los autores.

en la ciudad de San Francisco a los fines de exponer la metodología aplicada, pero se resalta el hecho de que su aplicación alcanza a todo el territorio provincial.

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 que comenzaron a incorporar métodos de valuaciones masivas 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 y, de manera más reciente, algoritmos de aprendizaje automático como árboles de regresión, k-vecinos cercanos o redes neuronales.

Entendiendo la valuación masiva como un mecanismo de estimación a gran escala, se procederá a generar un modelo estadístico que permita estimar el valor unitario de la tierra (VUT) urbana en la ciudad de San Francisco, Córdoba, en base a un conjunto de variables independientes ("inputs") que, a priori, definen el VUT ("output").

El método utilizado para la estimación es un algoritmo de aprendizaje automático (machine learning) conocido como Random Forest, junto con una técnica geoestadística, conocida como Kriging Ordinario, en el tratamiento de los residuos, lo cual constituye la principal innovación del presente trabajo de investigación, ya que mejora notablemente la capacidad predictiva del modelo. Se realiza, además, una comparación entre la técnica de estimación aplicada con otros métodos estadísticos alternativos y se evalúa el desempeño diferencial en las capacidades predictivas en cada caso.

Finalmente, los autores desean destacar que la información utilizada en el presente trabajo fue generada en el marco del Estudio Territorial Inmobiliario de la Provincia de Córdoba, financiado en conjunto por el Programa de Naciones Unidad para el desarrollo (Programa PNUD AR/16/005) 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 entre mediados de 2017 y fines de 2018 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 en la provincia. 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 septiembre de 2018 cuenta con más de 8.000 datos georreferenciados.

2. Área de estudio: Ciudad de San Francisco

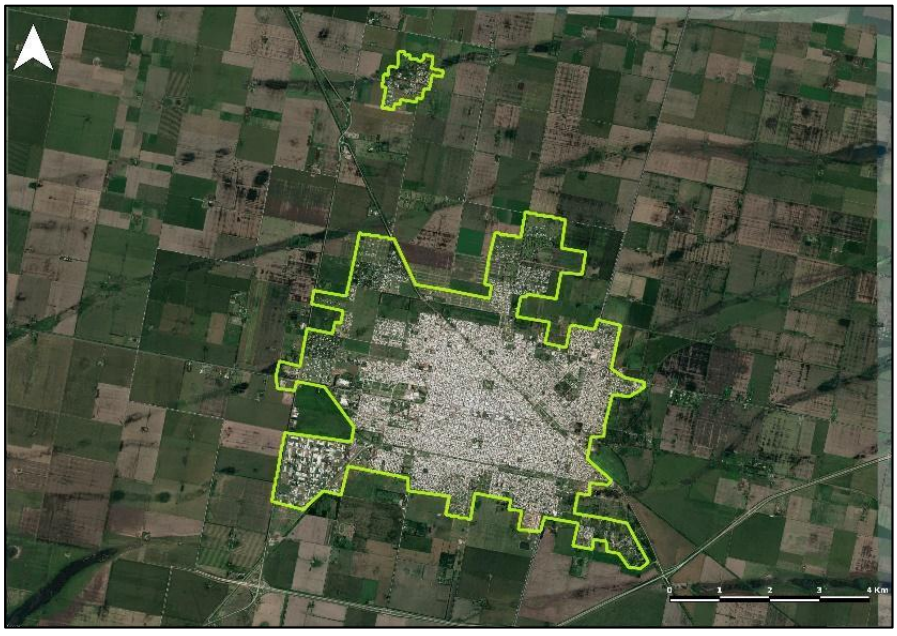
La ciudad de San Francisco se caracteriza por una estructura urbana monocéntrica de tejido ortogonal, ubicada a 206 km al este de la Ciudad de Córdoba, atravesado de NE a SE por el ferrocarril y contenida hacia el Norte y el Oeste por la ruta nacional N° 19 y ruta Provincial N° 158. Como se observa en el Figura 1, la ciudad se encuentra en plena expansión, cuenta con aproximadamente 62.000 habitantes (INDEC,2010) y en 2017 contaba con 29.036 parcelas urbanas registradas en la base catastral y 32.298 cuentas con valuación urbana. En el Figura 2 se muestra la mancha urbana de la ciudad, con mayoritaria expansión en sentido norte y oeste. La localidad se constituye en cabecera del departamento San Justo y con una impronta

agroganadera y con una importante trayectoria en el sector lechero. Asimismo, en los últimos años se ha constituido en un importante polo industrial de la provincia de Córdoba, con la localización de industrias metalmecánica, eléctrica, plástica entre otras.

Figura 1 - Mapa de San Francisco

Fuente: Elaboración propia.

Figura 2- Mapa de mancha urbana San Francisco



Fuente: Elaboración propia.

3. Enfoque Metodológico

Con el objeto de contextualizar la metodología de Random Forest, se parte analizando la los árboles de clasificación y regresión (CART) el cual es un método multivariado que permite describir la variable de estudio (generalmente llamada output), mediante medidas de posición y dispersión (ej. media, varianza), clasificarla y jerarquizarla en función de un conjunto de variables explicativas (también llamadas inputs), e inferir que valores puede asumir la variable dependiente en función de un conjunto de variables independientes.

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. Los 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 independientes, 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:

i. Nodo de raíz: Representa toda la población o muestra y esto se divide en dos o más conjuntos homogéneos.
ii. División (Split): Proceso de dividir la población en subconjuntos, llamados subnodos.
iii. Nodo de decisión: cuando un subnodo se divide en otros subnodos, se llama nodo de decisión.
iv. 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) el algoritmo CART está compuesto 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 $\sum \left(y_i - \underline{y}\right)^2$. En función de este esquema, cada uno de los dos nodos originados en la primera partición se vuelve a separar nuevamente si:

i. Hay suficiente heterogeneidad para producir una partición de observaciones, y
ii. 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), la construcción de un árbol de regresión consiste en desarrollar el siguiente procedimiento. Dada 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:

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

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

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

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:

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

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

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

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.

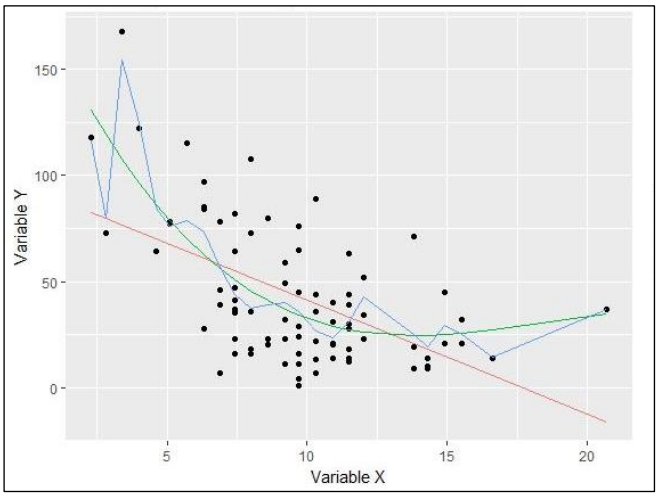
De esta manera, cada hoja de un árbol de regresión contiene aquellas observaciones lo suficientemente similares como para descartar la necesidad de generar nuevos nodos, representando un subconjunto homogéneo en función de los parámetros iniciales fijados en el algoritmo. El promedio del output para 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 la estimación de datos nuevos. Esta deficiencia, conocida como overfitting o sobre-ajuste, implica que el modelo presenta un excelente ajuste para los datos muestrales, como se observa en la Figura 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 – Overfitting y Underfitting



Fuente: Elaboración propia en base a Hastie, Tibshirani y Friedman (2008)

Por lo tanto, al desarrollar un árbol de regresión se obtiene un modelo con elevada 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 inconveniente 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 o Stacking, siendo el primero el método utilizado en este trabajo.

3.1. Combinación de diferentes árboles: Bagging.

Siguiendo a Breiman (1984), bajo un algoritmo de aprendizaje, la técnica de Bootstrap Aggretation, conocida como Bagging, 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 confecciona un CART para cada uno de estos subconjuntos muestrales y se promedian las estimaciones 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 (1984) y la selección aleatoria de inputs en cada nodo, 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 sobre la muestra, 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 Ti (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 genera 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, donde $m < M$. 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.
    vi. Se predice el conjunto de datos de testeo para evaluar la capacidad predictiva del modelo entrenado anteriormente, procediendo de la siguiente manera. 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.

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

3.3. Incorporación de la dependencia espacial

Si bien Random Forest es un método apropiado y ventajoso para hacer valuaciones masivas considerando una gran cantidad de variables, el modelo no tiene en cuenta la dependencia espacial de los datos, situación que resulta clave al estudiar un sistema complejo con características territoriales Subyace a este argumento la premisa de Tobler (1970) que indica que: "todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes". Resulta, entonces, de vital importancia incorporar al análisis algún método complementario que incorpore la estructura espacial al análisis para predecir el valor del suelo urbano.

3.3.1. Kriging Ordinario

Un método adecuado para incorporar la relación de dependencia espacial a los resultados de Random Forest es la técnica conocida como Kriging Ordinario, que es una técnica geoestadística univariada de interpolación lineal (Oliver y Webster, 2015). En otras palabras, se trata de 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 es aleatoria y dependiente del espacio, siendo el mismo proceso estacionario con media constante y varianza dependiente de la distancia y dirección.

Suponiendo una función aleatoria Z_{x_i}, siendo Z cada observación muestral y su localización $x_1, x_2, \dots, x_n$; para N datos, donde Z distribuye normal con media (estacionaria) , covarianza dependiente de la distancia y dirección representando la variable ouput. El estimador es una combinación lineal ponderada de los datos:

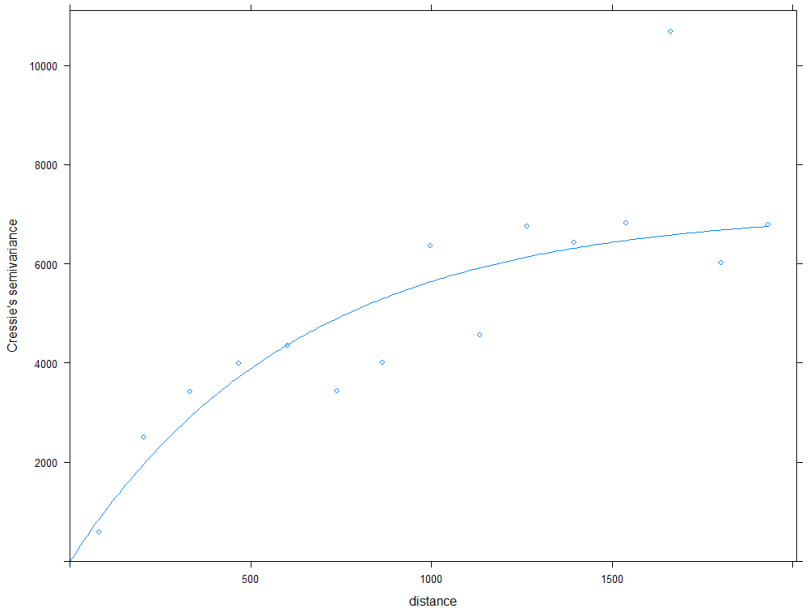
$$z(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i)$$

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 estas debe ser igual a 1. A su vez, x_0 es la localización donde se desea obtener la estimación del output bajo estudio.

A través de un semivariograma empírico, que recoge la estructura 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, 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).

En la Figura 4 puede apreciarse un semivariograma. El eje de las ordenadas mide la semivarianza, calculada como la varianza multiplicada por 0,5. En el eje de las abscisas se mide la distancia en metros. Cada punto en el plano representa la semivarianza promedio entre pares de observaciones que se encuentran en el radio medido por la distancia indicada. De esta manera, la varianza del output entre pares de observaciones que se encuentran a una distancia de 500 metros, como máximo, es menor que la varianza entre pares de observaciones que se encuentran separadas por 1.500 metros. Esta estructura de información es indicativa de la existencia de auto-correlación espacial, en donde las observaciones más próximas en el espacio tienden a tener outputs más parecidos entre sí que aquellas que se encuentran a distancias mayores. Para cuantificar la auto-correlación espacial detectada en el semivariograma, se procede a aproximar una función teórica que describa la estructura empírica observada en el gráfico.

Figura 4 – Semivariograma teórico y empírico



Fuente: Elaboración propia.

3.4. Combinación entre Random Forest y Kriging Ordinario

Si suponemos que los residuos obtenidos a partir del algoritmo Random Forest 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, se puede incorporar la auto-correlación espacial al análisis sumando la tendencia estimada mediante el algoritmo y la estimación de los residuos mediante un Kriging Ordinario para elevar la potencia predictiva del modelo utilizado. Es decir:

$$Estimación\ final = Tendencia\ estimada\ con\ Random\ Forest\ +\ Residuos\ estimados\ con\ Kriging$$

Es decir, si se detecta que los residuos obtenidos mediante el algoritmo Random Forest presentan auto-correlación espacial, no corregir este sesgo podría desembocar en la aplicación de un modelo que, en términos generales tenga un residuo normalmente distribuido con media igual a cero, pero que localmente arroje errores considerables. Para salvar este sesgo originado en la estructura de dependencia espacial de los datos, se procede a modelar los residuos y sumarlos a la estimación original en una segunda instancia.

4. Pre-procesamiento y descripción de la muestra

La elaboración de la base de datos para la predicción del valor de la tierra (VUT) de la ciudad de San Francisco fue realizada en el marco del proyecto de mejora en la eficiencia del Ministerio de Finanzas de la Provincia de Córdoba, Argentina, financiado por PNUD, con el objetivo de actualizar el valor fiscal de las propiedades.

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: en relación 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 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 urbana (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 se detallan a continuación:

Tabla 1 – Descripción de las variables utilizadas.

variable independite	Descripción	Tipo	Media	Desv. Est.	Mínimo	Máxima	Mediana	CV
d_depre	distancia a zonas depreciadas	Continua	2565.40	1121.60	472.97	7023.14	2539.15	0.44
perc_edif	% edificado en entorno	Continua	0.33	0.22	0.03	0.84	0.27	0.66
d_viasprin	distancia a vias principales	Continua	805.92	774.74	10.00	4965.41	597.89	0.96
d_indust	distancia a zona industrial	Continua	2238.74	1216.65	0.00	6016.84	2380.01	0.54
fragment	grado de fragmentacion urbana	Continua	0.86	1.01	0.00	3.00	0.00	1.18
d_baja	distancia a zona de meor categoría	Continua	2192.78	1362.64	0.00	6288.16	2022.77	0.62
ind_veg	promedio vegetación en un entrono 500 mts	Continua	0.58	0.29	0.11	0.98	0.59	0.50
bci	nivel de composición biofisica urbana	Continua	1.88	0.11	1.71	2.21	1.88	0.06
ndsi	identifica composición y cobertura del suelo	Continua	7.05	0.99	5.59	9.40	6.75	0.14
ui	indice de densidad edificada	Continua	-0.27	0.11	-0.51	-0.10	-0.25	-0.42
ndbi	identifica zonas con superficie edificada	Continua	-0.03	0.06	-0.20	0.07	-0.02	-2.04
prom_lote	Tamaño promedio parcela entorno de 500 m	Continua	427.71	336.94	254.22	1878.30	315.21	0.79
d_alta	distancia a zona de mayor categoría	Continua	1326.06	1000.93	0.00	6007.80	1229.81	0.75
d_lineadiv	distancia a lineas divisorias de valor	Continua	573.53	691.19	10.00	4691.02	405.03	1.21
d_viassec	distancia vias secundarias	Continua	258.46	261.41	10.00	1261.15	174.91	1.01
d_ffcc	distancia a ferrocarril	Continua	691.75	446.21	30.00	2001.12	571.44	0.65
prom_edif	promedio edificado en un entrono de 500 m	Continua	152.74	67.68	43.69	325.38	134.50	0.44
perc_baldm	porcentaje en metros cuadrados de parcelas baldias entorno 500 m	Continua	0.32	0.23	0.02	0.85	0.34	0.72
perc_bald	porcentaje de parcelas baldias en relacion al total de parcelas	Continua	0.31	0.23	0.02	0.87	0.33	0.73
sntl_pavim	nivel de desarrollo del territorio en un entorno 500 m	Continua	0.27	0.22	0.01	0.70	0.25	0.79
d_lineaqui	distancia a lineas de quiebre de valor	Continua	1304.40	1068.10	22.36	5864.00	1085.58	0.82
d_ruta	distancia a ruta	Continua	2850.57	1434.95	10.00	5342.53	3027.16	0.50

Fuente: Elaboración propia.

La base de datos cuenta con 174 observaciones, Dispersas dentro del ejido municipal de la ciudad de San Francisco. Como se observa en el mapa (Figura 5), cada punto representa una observación y el tamaño del punto hace referencia a su valor.

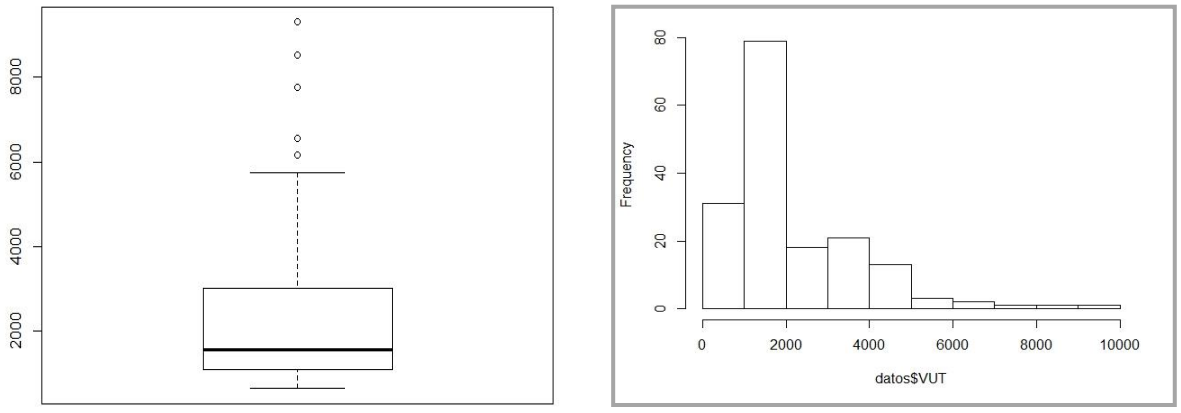
En el Figura 6 y a modo de observar la distribución del VUT, se presentan los gráficos de histograma y box plot. En ambos se aprecia que la distribución del VUT resulta ser asimétrica hacia la izquierda.

Por lo tanto, al aplicar la técnica para depurar los datos atípicos, el resultado obtenido aproximaría los valores del VUT a una distribución normal. Este resultado pareciera ser bueno ya que elimina los valores influyentes, aunque estos valores son aquellos que se encuentran en las zonas centrales 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. Dadas las características específicas del problema de estudio, de carácter netamente territorial, se decide respetar la dependencia espacial de la muestra, ya que se trabajará con modelos que no requieran normalidad en la base datos, como Random Forest.

Figura 5 – Mapa de San Francisco con observaciones muestrales

Fuente: Elaboración propia.

Figura 6 – Box Plot e Histograma del Valor del Suelos en San Francisco



Fuente: Elaboración propia.

A diferencia de los anterior, sí es necesario identificar y eliminar los puntos atípicos en su entorno espacial, también llamados inliers (Anselin, 2001). Estos son datos que difieren significativamente de los observados en su vecindario. Su identificación se realiza a partir del cálculo 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 calcula el índice de Moran local. Aplicando este procedimiento, mediante una matriz de 4 vecinos más cercanos sujeta a 1000 metros de distancia máxima, se reconocieron 22 observaciones como inliers, resultando así la base depurada con 170 observaciones.

En la Tabla N° 2 se muestran los estadísticos descriptivos de la muestra depurada. Siendo la media del valor unitario de la tierra urbana en San Francisco de $2.177, la mediana igual a $1.546 y el desvío estándar igual a $1.574, lo cual pone de manifiesto la elevada dispersión y la distribución sesgada que presenta la muestra utilizada.

Tabla 2 – Estadísticas descriptivas del Valor de Suelo

Min	Max	Mediana	Media	Desv. est.
$631	$9,308	$1,546.00	$2,177	$1,574.00

Fuente: Elaboración propia.

5. Resultados.

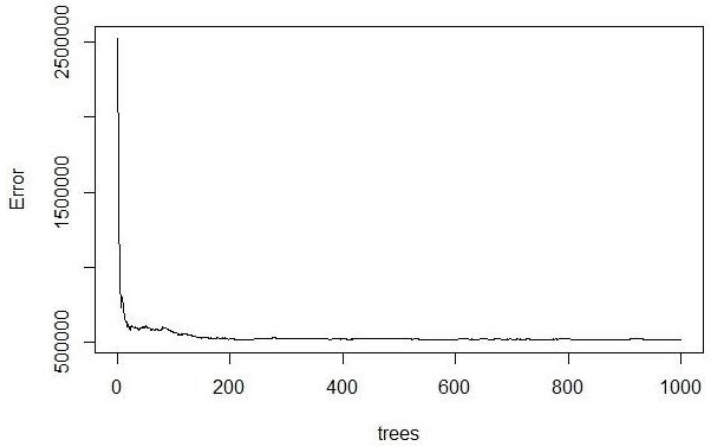
A los fines de llevar a cabo la estimación del VUT para cada cuadra de la ciudad, se utilizará el siguiente procedimiento, parte del cual se detalló en el apartado anterior:

i. 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).
ii. 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).
iii. 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.
iv. Se determinan los residuos de la estimación a partir de los valores predichos y los valores observados en cada una de las muestras
v. Se aplica sobre los residuos el método Kriging Ordinario, ajustando el semivariograma correspondiente.
vi. 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.

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 la Figura 7 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.

Figura 7 – Evolución del error de predicción en base a la cantidad de Arboles generados.



Fuente: Elaboración propia.

Del proceso de aprendizaje automático del método Random Forest, y a través de un procedimiento de validación cruzada en 10 grupos, surgió un subconjunto óptimo de inputs a considerar en cada nodo es igual 13, de un total de 27 variables independientes.

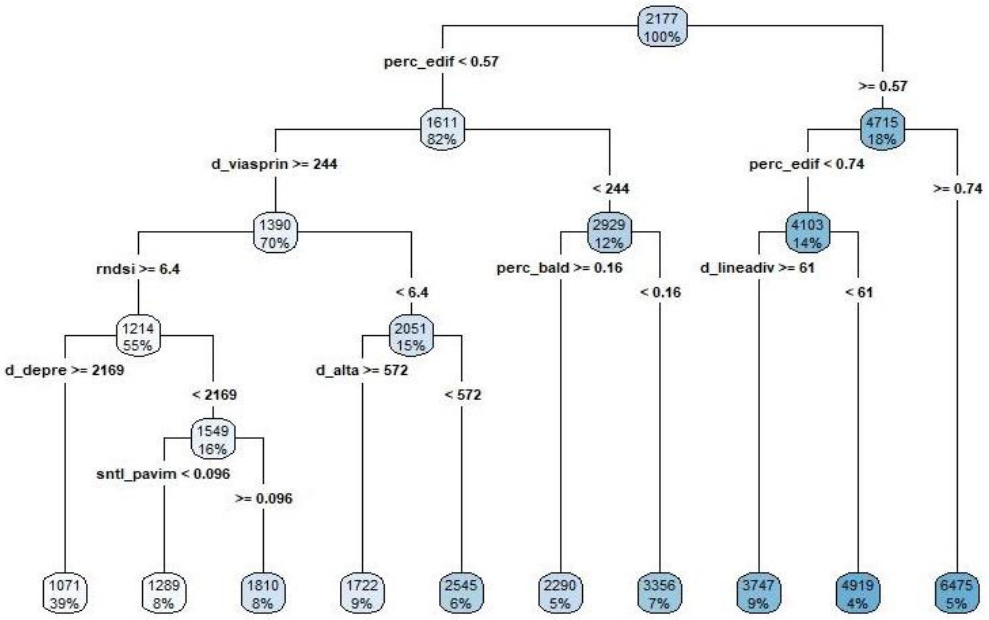
Una de las principales desventajas de los métodos basados en árboles es que no permiten cuantificar en una forma funcional la relación de cada variable input con el output bajo estudio, dada la gran cantidad de árboles aleatoriamente generados. Sin embargo, sólo a fines expositivos, se puede extraer del bosque de regresión un árbol aleatorio para analizar su composición, tal como se presenta en la Figura 8.

La variable más influyente para subdividir las observaciones en primera instancia “perc_edif” en un valor igual a 0,57 (porcentaje de metros cuadrados cubiertos en relación a la cantidad total de metros cuadrados de tierra en un entorno de 500 metros). Posteriormente, la siguiente división se realiza con la variable “d_viasprin” en un valor igual a 244 metros (distancia total hacia las vías principales de la ciudad). 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, en conjunto con el porcentaje de la muestra que cumple con dichos criterios.

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 la Figura 8 puede apreciarse, analizando los 500 árboles utilizados en la estimación, la importancia de las variables medida en términos porcentuales el incremento de la suma del error cuadrático medio[11] del modelo ante la ausencia de la variable analizada. Se observa que mantienen su importancia las mismas variables resaltadas en la Figura 9, a las cuales se suma la variable “d_lineadiv” (distancia a una línea trazada por especialistas en el mercado inmobiliario local a partir de la cual se supone que el valor desciende) y “d_indust” (distancia total a zona industrial).

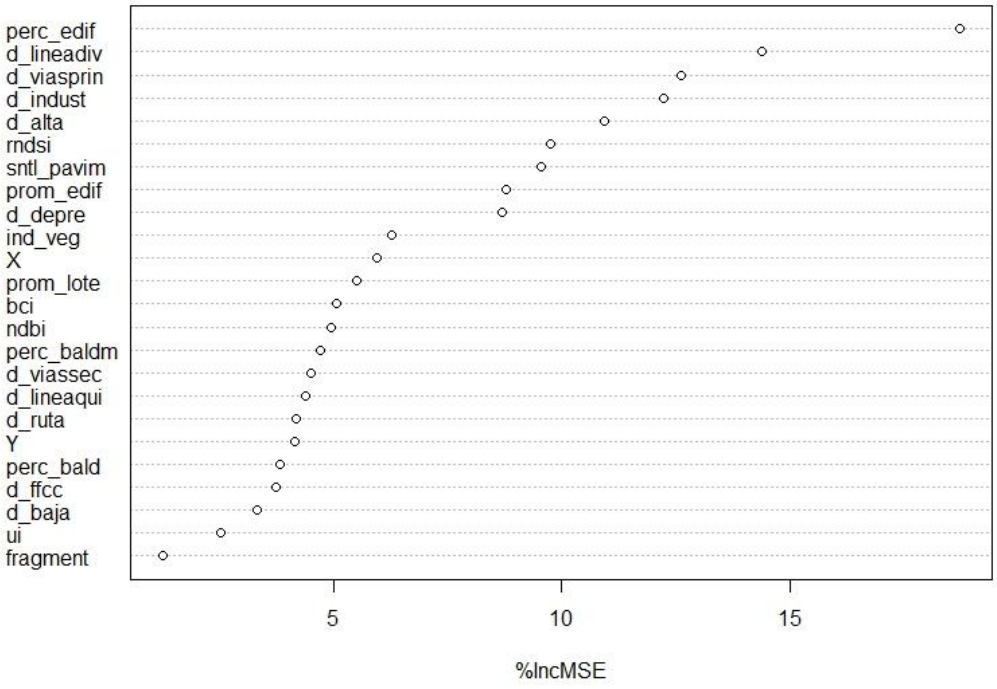
Figura 8 – Árbol de Regresión respecto a observaciones del San Francisco

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



Fuente: Elaboración propia.

Figura 9 – Importancia en el Error de las variables inputs.

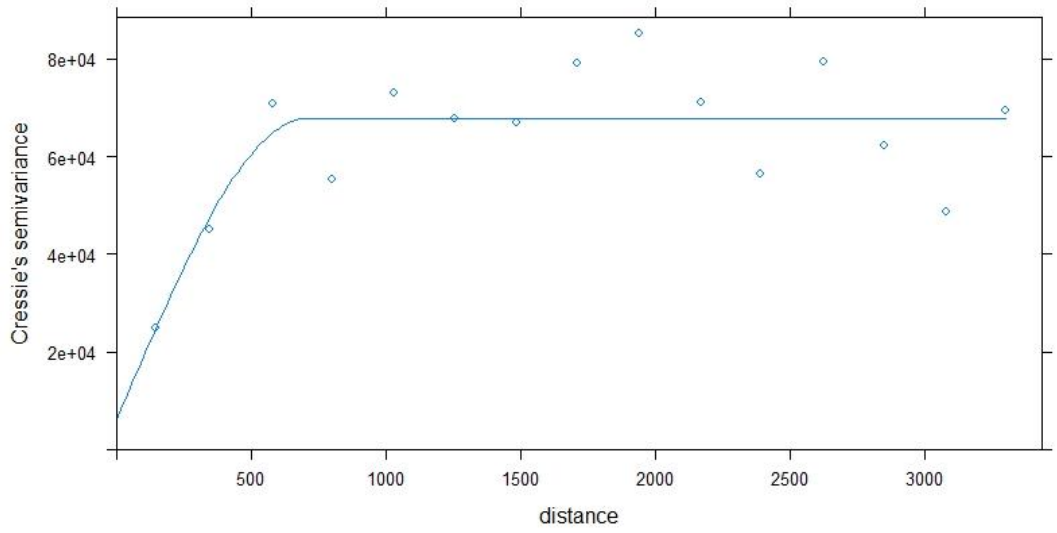


Fuente: Elaboración propia.

5.1. Incorporación de la dependencia espacial

Una vez obtenida la predicción, a través el método de Random Forest, se determinan los residuos (diferencia entre valor observado y predicho), sobre los que se aplicará la técnica Kriging Ordinario para interpolar la estimación al espacio continuo del mapa. Previo a esto, es necesario determinar el semivariograma adecuado para la interpolación. En función de la estructura de dependencia espacial de la muestra surge el modelo exponencial como el que mejor representa la auto-correlación espacial de los residuos, tal como puede apreciarse en la Figura 10:

Figura 10 – Semivariograma teórico



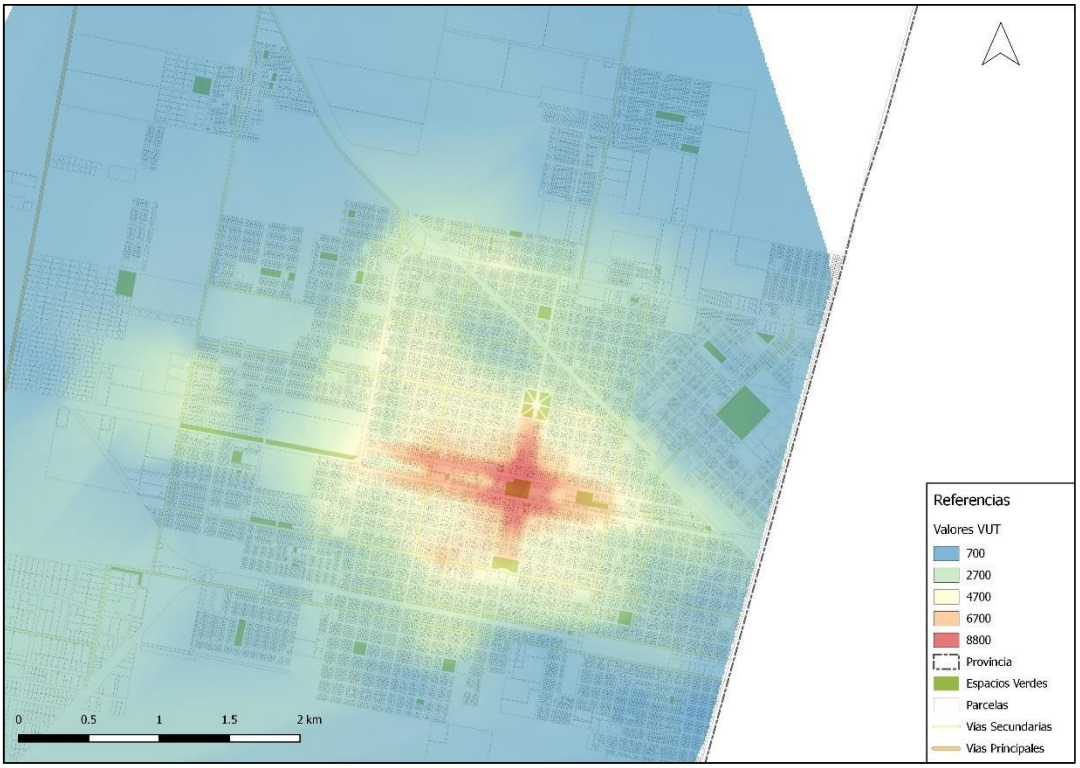
Fuente: Elaboración propia.

5.2. Ensamblaje de métodos para la estimación final

Para obtener la estimación final, en una primera instancia se estima la tendencia aplicando el modelo obtenido del algoritmo Random Forest sobre la base de datos de predicción, que cuenta sólo con los valores de las variables independientes o inputs. En una segunda instancia se aplicó el mismo método para la estimación de los datos muestrales para obtener una estimación de los residuos muestrales. Esta estimación de los residuos se interpoló al espacio continuo de la ciudad y, en una tercera instancia, se sumaron los residuos así obtenidos a la tendencia inicialmente estimada.

La Figura 11 muestra el mapa de valores del suelo urbano estimado, siendo el color rojo los valores más elevados y el celeste los valores más bajos.

Figura 11 – Mapa de San Francisco con raster de la predicción.



Fuente: Elaboración propia.

6. Comparación de métodos.

Para llevar a cabo el estudio comparativo entre modelos se procede, a través de un proceso de validación cruzada (“Cross Validation”), a evaluar la capacidad predictiva de los modelos en cuestión. Esta técnica consiste en dividir la base muestral de observaciónes en dos subconjuntos, uno de entrenamiento, que se utiliza para entrenar el modelo, y otro de prueba, cuya finalidad es realizar la validación del modelo. Generando 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 $n-1$ de veces, se denomina la técnica de validación cruzada $leave-one-out$. En otros términos, se extrae un dato de la muestra y con el resto se modela para la predicción. En esta sección se compararán el método Random Forest + Kriging Ordinario con la Regresión Linear Múltiple (OLS). En ambas especificaciones se utilizarán las mismas variables independientes.

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

i. Mean absolute percentaje error (MAPE):

$$MAPE = \sum_{i=1}^{n} \frac{\frac{|\widehat{y_i} - y_i|}{y_i}}{n}$$

ii. Median absolute percentaje error (MeAPE):

$$MeAPE = median\left(\frac{|\widehat{y_i} - y_i|}{y_i}\right)$$

En donde $\widehat{y_i}$ es el valor estimado o predicho, y_i es el valor observado, n es el tamaño de la muestra.

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:

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

$$CV = \frac{\frac{\widehat{y_i}}{y_i} - media(\frac{\widehat{y_i}}{y_i})}{media(\frac{\widehat{y_i}}{y_i})}$$

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

$$CD = \frac{\frac{\widehat{y_i}}{y_i} - mediana(\frac{\widehat{y_i}}{y_i})}{mediana(\frac{\widehat{y_i}}{y_i})}$$

Luego de realizar el proceso de validación cruzada para la elaboración de los estadísticos a través del software R, como puede apreciarse en la Tabla N°2, para el algoritmo Random Forest con Kriging el error esperado de la predicción medido por MAPE es igual al 20% (en términos de mediana es igual al 16%). A su vez, los coeficientes calculados para evaluar la homogeneidad de la estimación en relación a los datos de mercado permanecen por debajo del 19%, en línea con lo estipulado por el IAAO (2003).

Tabla 3– Comparación de métodos predictivos

Modelo	MAPE	MeAPE	CV	CD
RF-KO	0.20	0.16	0.19	0.19
OLS	0.24	0.19	0.24	0.24

Fuente: Elaboración propia.

Mientras que la exploración y comparación de métodos, resulta claro, a la vista de los resultados obtenidos que la incorporación de algoritmos de aprendizaje automático mejora el proceso de valuación masiva. Siendo Random Forest un método ofrece un progreso en el campo de la predicción al comparar los rendimientos con el modelo de regresión lineal espacial. Esto se observa en el MAPE donde el primer método presentó 0.20% mientras que el clásico método OLS 24%. Lo mismo sucede cuando se utilizan los parámetros propuesto por IAAO para la evaluación, es así que el coeficiente de dispersión para RF-KO son un 6% menor al obtenido en el otro método.

Por lo tanto, más allá de las ventajas asociadas a la incorporación de la dependencia espacial al análisis comentadas anteriormente, la capacidad predictiva del método híbrido de Random Forest + Kriging Ordinario muestra ser mayor que la del modelo lineal más simple.

7. Consideraciones finales.

El objetivo del presente artículo consistió en presentar la metodología de valuación masiva del valor unitario del suelo (VUT) y con lo cual reposicionar al impuesto inmobiliario, subutilizado actualmente, como un mecanismo de recaudación eficaz, equitativo y progresivo.

Para cumplir con el objetivo se procedió a través de una técnica algorítmica de aprendizaje automático denominada Random Forest, la cual fue combinada, 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 San Francisco, Provincia de Córdoba, se comprobó que la capacidad predictiva del algoritmo resultó adecuada. Debido a que los resultados obtenidos se enmarcan dentro de los parámetros de calidad establecidos por el IAAO. Esto es, el error relativo promedio en valor absoluto fue igual al 20%, mientras que la mediana de esta misma medida ascendió al 16%. En cuanto a la homogeneidad de los errores obtenidos, el coeficiente de variación fue igual a 0,19, siendo el coeficiente de dispersión igual a 0,19.

Asimismo, como resultado de la comparación del algoritmo Random Forest con la técnica de regresión lineal, se reflejó la ventaja del primero en términos del segundo. Ya que el primero permite contemplar las relaciones entra variables de forma no lineal, caracterizándose por ser un método no paramétrico, que no requiere el cumplimiento previo de fuertes supuestos, facilitando la aplicación del método a cualquier problema de la realidad. Además, los resultados arrojados por la regresión lineal presentaron un menor ajuste valuado en términos del error de predicción en relación al método de aprendizaje automático.

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. Además, la incorporación de la dependencia espacial al análisis, mediante la hibridación de la técnica algorítmica con una técnica geoestadística, tiene la ventaja de corregir potenciales errores de estimación locales que quedan ocultos detrás de las métricas globales usuales que suelen aplicarse para cuantificar la capacidad predictiva. 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.

Bibliografía

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.

Bonet J, Muñoz, A. y Pineda, C, Mannheim (2014). *El potencial oculto: factores determinantes y oportunidades del impuesto a la propiedad inmobiliaria en América Latina*. Banco Interamericano de Desarrollo.

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

Breiman, L.; Friedman, J.; Stone, C.; Olshen, R. (1984). *Classification and regression trees*. California, Wadsworth, Inc.

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

De Cesare, Claudia. (2012). Improving the Performance of the Property Tax in Latin America. Policy Focus *Report. Lincoln Institute of Land Policy*.

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(6).

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(3) 383-401.

International Association of Assessing Officers (2003). Standard on automated valuation.

Jeremy, M. (2006). Mapping the Results of Geographically Weighted Regression.*The Cartographic Journal*. 43(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(1) 94-112.

Lockwood, T. y Rossini, P. (2011). Efficacy in Modelling Location Valuation models (AVMs). Within the Mass Appraisal Process. Pacific Rim Property *Research Journal*, 17(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 al. (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(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.

Serra, M. V., David E. Dowall, Diana Meirelles da Motta, and Michael Donovan. (2005). An examination of three Brazilian cities: Brasilia, Curitiba, and Recife. In *Estudos estratéicos de apoio às politicas urbanas para os grupos de baixa renda no Brasil* (Enabling strategy for moving upgrading to scale in Brazil). *Urban land markets and urban land development*. Washington, DC: Cities Alliance.

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

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