UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA CIVIL

# PREDICCIÓN DE RENTAS EN SANTIAGO DE CHILE UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

MEMORIA PARA OPTAR AL TITULO DE INGENIERA CIVIL

ISABELLE CHARDON SCHIRM

PROFESOR GUÍA:
FRANCISCO MARTÍNEZ CONCHA

MIEMBROS DE LA COMISIÓN:
ALEXANDRE BERGEL
PEDRO DONOSO SIERRA

SANTIAGO DE CHILE

2019

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERA CIVIL
POR: ISABELLE CHARDON SCHIRM
FECHA: ABRIL 2019
PROF. GUÍA: FRANCISCO MARTÍNEZ CONCHA

# PREDICCIÓN DE RENTAS EN SANTIAGO DE CHILE UTILIZANDO ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

En este trabajo se aplican tres algoritmos de aprendizaje automático al problema de predicción de rentas de bienes inmobiliarios en la ciudad de Santiago de Chile por primera vez. Se dispone de una base de datos de 600.902 transacciones de casas y departamentos efectuadas entre los años 2007 y 2018, facilitada por la empresa TocToc.com, especialista en georreferenciación y tasación de bienes inmobiliarios.

Se comparan una red neuronal, un algoritmo de máquinas de vectores de soporte para la regresión (SVR) y un bosque aleatorio (conjunto de árboles de regresión) para la predicción de rentas de departamentos en la comuna de La Florida. Los errores absolutos porcentuales medios obtenidos son de 19,17 %, 14,69 % y 9,67 % respectivamente, por lo que el bosque aleatorio es el algoritmo más preciso para la predicción de rentas. Además, el bosque aleatorio funciona tanto con muestras pequeñas como con muestras grandes, mientras el poder predictivo de la red neuronal y del algoritmo SVR baja al reducir el tamaño de la muestra. También se observa que baja el error absoluto porcentual medio cuando la variable de precio a predecir es el valor del metro cuadrado y no el precio total del departamento, en concordancia con las conclusiones de Antipov y Pokryshevskaya (2012). Luego, se construyen distintas muestras según ingresos comunales y por región geográfica para predecir rentas de departamentos utilizando el bosque aleatorio, con lo que se calcula la importancia de los atributos de los bienes según el método de importancia de Gini (Breiman, 2001).

Las variables que más contribuyen en la determinación de los precios de los bienes son el año de venta, el ingreso promedio de los hogares por comuna y el índice de calidad calculado por el Servicio de Impuestos Internos, que representa la calidad estructural y la antigüedad de los departamentos. Cuando mejora el índice de calidad y aumenta el ingreso promedio de los hogares suben la rentas de los departamentos. Este último resultado también ha sido encontrado en Santiago de Chile por Figueroa (1992).

# Agradecimientos

Quiero darle las gracias a mis profesores guías Francisco Martínez y Alexandre Bergel por su confianza y su apoyo. Ha sido un proceso muy interesante, gratificante y lleno de enseñanzas. Gracias también a Pedro Donoso por sus pertinentes consejos y sugerencias.

Gracias a TocToc.com por colaborar con nosotros y en particular a Jonathan Orrego por su ayuda.

Gracias a mis compañeros del departamento de Transporte por alegrar los días.

También quiero agradecer a mis amigos chilenos con quién hemos compartido lindos momentos.

Agradecimientos especiales para mi familia y mis amigos de Francia que me transmitieron su energía y su cariño desde el otro lado del planeta.

Gracias a Diego por estar a mi lado, apoyarme y compartir tantos momentos de felicidad conmigo.

Finalmente, gracias a mis padres por su amor incondicional.

# Tabla de contenido

# Introducción

El mercado inmobiliario en Santiago de Chile ha evolucionado constantemente en la última década con una fuerte alza de las ventas de los departamentos, debido a los cambios en los hábitos de consumo de sus habitantes. De acuerdo con las observaciones de la Cámara Chilena de la Construcción, los santiaguinos valoran más la proximidad a los lugares de trabajo y la presencia de transporte público de calidad que la comodidad de una casa, generalmente ubicada en la periferia de la ciudad. Este cambio en los hábitos tuvo por consecuencia el aumento de los arriendos y de las compras de departamentos por inversionistas. Adaptándose a la evolución del mercado inmobiliario, empresas tales como TocToc.com se especializan en la recopilación de información y la georreferenciación de los bienes inmobiliarios de Santiago con el objetivo de proveer a los potenciales compradores toda la información necesaria para la adquisición de una vivienda.

Además de ser dictadas por un contexto económico, las rentas de los bienes inmobiliarios (que se define como los montos de las transacciones realizadas) dependen de sus atributos propios y también de atributos externos relacionados con su localización. Por esta razón, estos bienes son calificados de «casi únicos», puesto que dos viviendas estructuralmente idénticas son necesariamente espacialmente diferenciadas. El modelo hedónico ha sido la herramienta comúnmente utilizada para modelar la relación causa efecto entre los atributos de los bienes y su precio y para la predicción de rentas mediante una regresión lineal. En Santiago de Chile se ha aplicado el modelo hedónico para analizar las componentes del precio de las viviendas (Figueroa, 1992), valorar atributos de viviendas sociales (Quiroga, 2005) y estudiar el efecto de la presencia de infraestructuras para ciclistas sobre las rentas (Vega Flores, 2017).

En el principio de los años 90 se popularizó el uso de las redes neuronales para predecir las rentas de los bienes inmobiliarios y se mostró que representan una alternativa seria a la tradicional regresión hedónica. Sin embargo, sus detractores denuncian su dificultad de interpretación y la imposibilidad de obtener una función matemática que relacione los atributos de las viviendas con su precio. Luego surgieron otros algoritmos de aprendizaje automático para la predicción de rentas tales como el algoritmo de máquinas de vectores de soporte para la regresión (SVR por *Support Vector Regression*; Drucker, Burges, Kaufman, Smola y Vapnik, 1997) y el bosque aleatorio (Breiman, 2001), que al igual que las redes neuronales son capaces de tratar con problemas no lineales. En este trabajo se presenta la primera aplicación de estos algoritmos para la predicción de rentas en Santiago de Chile.

El objetivo de este trabajo es comparar tres algoritmos comúnmente usados en la área de la inteligencia artificial para la predicción de rentas y concluir sobre una posible alternativa

al modelo hedónico tanto en términos de poder predictivo como en la interpretación de los resultados. Se busca predecir el precio de los departamentos en Santiago de Chile utilizando los distintos algoritmos con una precisión comparable con la precisión obtenida en otras ciudades y analizar los resultados espacialmente. Para eso, se cuenta con una base de datos puesta a disposición por la empresa TocToc.com, especialista en georreferenciación y tasación de bienes inmobiliarios. Además, se busca entender la contribución de los atributos internos y externos de los bienes inmobiliarios en la determinación de su precio, de manera similar a la valoración monetaria de los atributos propuesta en el modelo hedónico.

En el primer capítulo se describe el funcionamiento de las redes neuronales, del bosque aleatorio y del algoritmo SVR y se recopilan los resultados obtenidos para la predicción de rentas de bienes inmobiliarios utilizando estos tres métodos. También se presenta el modelo hedónico y sus aplicaciones en ciudades del mundo y en Santiago de Chile. En el segundo captítulo se describe la base de datos puesta a disposición por TocToc.com. En el tercer capítulo se implementan los tres algoritmos y se analizan los resultados obtenidos para la predicción de rentas de departamentos en la comuna de La Florida, con el objetivo de comparar los métodos en términos de tiempo de ejecución, complejidad y resultados. En el cuarto capítulo se muestran los resultados obtenidos con el bosque aleatorio para la predicción de rentas en toda la ciudad de Santiago y se analizan espacialmente mediante un software de sistema de información geográfica (GIS). Para considerar la heterogeneidad espacial y socio-económica de la ciudad de Santiago, se aplica el algoritmo del bosque aleatorio a muestras elaboradas por zonas geográficas y por estrato de ingreso de los habitantes. Luego, se presentan los riesgos de sesgo en el cálculo de la importancia de las variables y cómo prevenirlos, a partir de los trabajos de Strobl, Boulesteix, Zeileis y Hothorn (2007) y Hothorn, Hornik y Zeileis (2006). Finalmente se concluye sobre la utilización del bosque aleatorio para la predicción de rentas y se presentan líneas de investigación futuras.

# Capítulo 1

# Revisión bibliográfica

El modelo hedónico (Rosen, 1974) ha sido el modelo de referencia por décadas para la modelación de las rentas de los bienes inmobiliarios y los métodos de regresión lineal han sido considerados como métodos tradicionales para la predicción de rentas. En el principio de los años 1990, esta área se ha abierto a nuevos métodos de predicción aplicando técnicas de aprendizaje automático. Dentro ellas, la red neuronal ha sido la más utilizada y se pudo mostrar que este algoritmo es capaz de competir con la regresión hedónica clásica (Tay y Ho, 1992). Sin embargo, también ha sido criticada por su falta de transparencia y de estabilidad. Años después se empezaron a utilizar el algoritmo de máquinas de vectores de soporte para la regresión (SVR) (Drucker et al., 1997) y el bosque aleatorio (Breiman, 2001) y a pesar de la escasez de estudios donde se aplican a la predicción de rentas, los resultados han sido prometedores por lo que estos dos algoritmos podrían representar una alternativa seria a la regresión hedónica y a la red neuronal.

En esta sección se describe el funcionamiento de la red neuronal, el bosque aleatorio y el algoritmo SVR y se presenta una recopilación de los principales estudios disponibles en la literatura donde se comparan estos métodos con la regresión hedónica y también entre ellos.

## 1.1. Redes neuronales

Las redes neuronales son modelos computacionales inspirados en el comportamiento de las neuronas del cerebro humano que se usan en problemas de regresión, clasificación (predicción de la categoría a la cual pertenece un objeto a partir de sus características), análisis de imágenes y de traducción. En los sistemas nerviosos, una neurona recibe información en forma de señal eléctrica o química a través de las dendritas, que luego es ponderada y procesada en el soma. La señal obtenida se propaga por el axón hasta llegar a las terminaciones sinápticas, cuyo propósito es transmitir la información a las otras neuronas del cerebro, formando una red.

McCulloch y Pitts (1943) fueron los primeros investigadores en proponer un modelo computacional para las neuronas a partir de funciones lógicas, llamado perceptrón. El per-

ceptrón recibe la información proveniente de otras unidades y la pondera aplicándole una función lineal $f(x) = ax + b$ donde $x$ es un vector de $\mathbb{R}^n$ que contiene los valores numéricos de la información y $a, b \in \mathbb{R}$. El valor de salida obtenido en la neurona artificial es 0 o 1, dependiendo del valor de la suma de las ponderaciones de las entradas. Si este valor supera cierto umbral, el valor de salida de la neurona vale 1 y 0 en el caso contrario.

Rosenblatt (1958) utiliza la neurona artificial de McCulloch y Pitts (1943) para desarrollar un algoritmo de aprendizaje para resolver problemas de clasificación, que consiste en una neurona con una capa de entrada y de salida que procesa variables binarias. Este algoritmo converge si el problema de clasificación es linealmente separable, es decir si existe un hiperplano que permite separar los datos de las distintas categorías en un espacio de cualquier dimensión.

El algoritmo de *Backpropagation* (Rumelhart y McClelland, 1986) extendió el uso de las redes neuronales a problemas no linealmente separables, y permite a una red aproximar cualquier función medible Borel de un espacio finito a otro con cualquier nivel de precisión (Hornik, Stinchcombe y White, 1989), por lo que se puede adaptar a cualquier tipo de problema.

Las redes utilizadas han sido utilizadas tanto para problemas de clasificación como para problemas de regresión en distintas áreas de investigación, tales como la biología (Barabasi y Oltvai, 2004), el reconocimiento de imágenes (Rowley, Baluja y Kanade, 1998) y la predicción de rentas inmobiliarias (Borst, 1991).

### 1.1.1. Funcionamiento

Una red neuronal se compone de neuronas artificiales conectadas entre ellas y organizadas por capas. La Figura 1.1 muestra una red neuronal totalmente conectada, es decir que cada neurona es conectada a todas las neuronas de la capa adyacente, y con un nodo de salida, aunque es posible construir una red neuronal con más de un nodo de salida si se requiere predecir varias variables. En la Figura 1.1 las conexiones entre las neuronas son representadas por flechas y las neuronas por círculos. Cuando se aplica una red neuronal a un problema de regresión, se busca la mejor configuración de la red para representar la relación causa-efecto entre un conjunto de variables ($X_i$) y un valor objetivo ($Y_i$) con el mayor nivel de precisión posible.

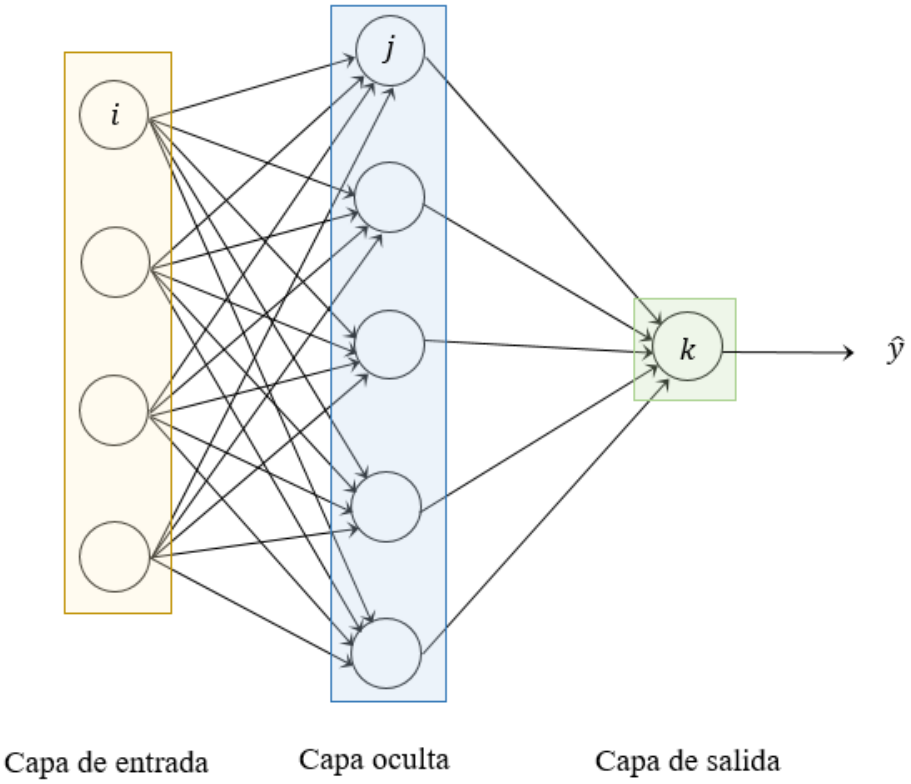


Figura 1.1: Representación gráfica de una red neuronal totalmente conectada con una capa de entrada, una capa de oculta y una capa de salida.

La entrada a un nodo es la suma ponderada de las salidas de los nodos conectados a este. Entonces la entrada $h_j$ a un nodo $j$ está descrita por la Ecuación 1.1, donde $w_{ij}$ (peso asociado a la conexión entre la neurona $i$ y la neurona $j$) y $b_j$ (constante asociada a la neurona $j$) son los parámetros a estimar, y $x_i$ es la variable de salida de una neurona $i$ que entra a la siguiente capa (excepto en el caso de la capa que sigue inmediatamente a la capa de entrada, donde $x_i$ es una variable predictora observada).

$$h_j = \sum_i w_{ij} x_i + b_j \tag{1.1}$$

En cada nodo se aplica una función - denominada función de activación - a $h_j$, que suele ser de tipo sigmoid, tangente hiperbólica o semi-lineal, también llamado ReLU por Rectified Linear Unit (Ecuaciones 1.2, 1.3 y 1.4). Esta última es la más utilizada en la actualidad por entregar mejores resultados que la sigmoid y la tangente hiperbólica (Glorot, Bordes y Bengio, 2011) además de tener validez biológica (Hahnloser, Sarpeshkar, Mahowald, Douglas y Seung, 2000).

$$f(x) = \frac{1}{1 + e^{-}x} \tag{1.2}$$

$$f(x) = tanh(x) \tag{1.3}$$

$$f(x) = max(0, x) \tag{1.4}$$

Entonces, el output de una neurona $j$ se puede escribir como:

$$v_j = f(h_j) = f(\sum_i w_{ij} x_i + b_j) \tag{1.5}$$

Este mecanismo de propagación hacia adelante se repite hasta el final, donde se obtiene el valor de salida de la red. En el ejemplo de la Figura 1.1, el valor de la salida de la red es:

$$\widehat{y} = g(d_k) \tag{1.6}$$

donde $d_k$ es la ponderación de las salidas obtenidas en la capa oculta, es decir:

$$d_k = \sum_j w_{jk} + b_k \tag{1.7}$$

Los parámetros $w_{ij}$ y $b_j$ son estimados mediante un algoritmo denominado algoritmo de propagación hacia atrás (Rumelhart y McClelland, 1986), cuyo propósito es ajustar los pesos asociados a las conexiones y las constantes de las neuronas para minimizar el error entre la salida de la red neuronal y el valor objetivo (observado exógenamente). La medición de error más común es la suma de los errores cuadrados, definida como:

$$E = \frac{1}{2} \sum_p (y_p - \widehat{y})^2 \tag{1.8}$$

donde $p$ es el número de observaciones utilizadas para la predicción, $y_p$ es el valor observado de la variable e $\widehat{y}_p$ es el valor entregado por la red.

La búsqueda de los parámetros de la red se efectúa de manera iterativa mediante un algoritmo de minimización de gradiente estocástico. Usualmente se utiliza el algoritmo Adam (*Adaptative Momentum Estimation*) por su rapidez de convergencia (Ruder, 2016). El término de propagación hacia atrás viene de la utilización de la regla de la cadena para calcular la derivada de la función objetivo con respeto a los pesos de la red. En el ejemplo de la Figura 1.1, la derivada de la función objetivo con respecto a $w_{ij}$ se puede escribir como:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial \widehat{y}} \frac{\partial \widehat{y}}{\partial d_k} \frac{\partial d_k}{\partial v_j} \frac{\partial v_j}{\partial h_j} \frac{\partial h_j}{\partial w_{ij}} \tag{1.9}$$

Se aplica el mismo proceso de minimización del error para la actualización de las constantes neuronales. El algoritmo de propagación hacía atrás termina cuando se encuentran los valores $(w, b)$ que minimizan la función objetivo.

### 1.1.2. Inicialización de los pesos y de las constantes neuronales

La inicialización de los coeficientes impacta de manera significante en la rapidez de convergencia del algoritmo de minimización y su resultado. He, Zhang, Ren y Sun (2015) proponen una inicialización adaptada a la función de activación semi-lineal, que consiste en mantener

los coeficientes en un rango que depende del tamaño de la capa de neuronas donde tienen efecto. Con este método se alcanza el mínimo global del error más rápido y de manera más eficiente que con las otras inicializaciones existentes. Los coeficientes tienen una distribución normal centrada en 0 y de varianza $\sqrt{\frac{2}{n_l}}$:

$$W \sim N(0, \sqrt{\frac{2}{n_l}}) \tag{1.10}$$

donde $n_l$ es el número de nodos que se encuentran en las capas conectadas por $W$.

### 1.1.3. Problemas de sobreajuste

Se utiliza el término de sobreajuste cuando una red neuronal adapta sus parámetros a características específicas no predictoras de la muestra de entrenamiento. La consecuencia es que la red no puede representar la relación causa-efecto que existe entre las entradas y las salidas en una base de datos nueva. En este caso, se dice que la red es sobre entrenada. Dentro de las técnicas más comunes para resolver el problema de sobreajuste esta la técnica *LASSO* por *Least Absolute Shrinkage and Selection Operator* (Tibshirani, 1996), que busca estabilizar la estimación de los parámetros de la red. Este método consiste en sumar a la función objetivo del problema de minimización un término que depende del valor de los pesos ($w_{ij}$) y entonces privilegia los pesos pequeños, que son menos sensibles al ruido que pueden tener los datos. La nueva función objetivo es:

$$E = \frac{1}{2}\sum_p (y_p - \widehat{y_p})^2 + \lambda \sum |w| \tag{1.11}$$

donde $\lambda$ es el término de regularización, que es determinado por el modelador.

La regularización propuesta por Tikhonov (1963), llamada *Ridge Regression*, funciona de manera similar, pero penaliza la suma de los coeficientes al cuadrado. En este caso la función objetivo a minimizar es:

$$E = \frac{1}{2}\sum_p (y_p - \widehat{y_p})^2 + \lambda \sum w^2 \tag{1.12}$$

La diferencia entre los dos métodos es que, al minimizar la sumatoria del valor absoluto de los pesos, el método *LASSO* aumenta el número de coeficientes nulos, mientras que el método de Tikhonov reduce el valor de los coeficientes y así reduce el impacto de cada peso sobre el resultado final. Estas dos técnicas son complementarias, por lo que es conveniente combinarlas para mejorar el efecto de la regularización sobre el comportamiento de la red (Zou y Hastie, 2005). El problema de sobreajuste también se puede resolver utilizando el método de *Dropout* (Srivastava, Hinton, Krizhevsky, Sutskever y Salakhutdinov, 2014), que consiste en ignorar una parte de los coeficientes a cada iteración del algoritmo de propagación hacia atrás, lo que significa que a cada iteración se entrena una red reducida distinta.

### 1.1.4. El modelo de atención

El modelo de atención es una herramienta que permite extraer de una red neuronal una medida de la importancia de cada variable en la relación causa efecto existente entre las variables de entradas y de salida. Para una red neuronal simple, se agrega una capa adicional con una función de activación de tipo *softmax* entre la capa de entrada y la primera capa oculta. Esta capa se llama capa de atención (*atention layer*) y contiene el mismo número de neuronas que la capa de entrada. Este método, descrito por Vaswani et al. (2017) permite extraer el porcentaje de participación de las variables aplicando un producto escalar entre la capa de entrada y la capa de atención.

## 1.2. Bosques aleatorios

### 1.2.1. Los árboles de regresión

El Bosque Aleatorio es una combinación de árboles de decisión o de regresión desarrollada por Breiman (1996), que se utiliza para problemas de regresión o de clasificación y cuyo objetivo es predecir el valor de una variable continua $Y$ o la pertenencia a una categoría a partir de un conjunto de variables predictoras ($X$).

Un árbol de regresión es un árbol binario que se compone de un nodo raíz, de nodos internos y de nodos terminales llamados hojas, representados por círculos en la Figura 1.2. Cada nodo interno representa un subconjunto de las observaciones y un test binario que resulta en la generación de dos nodos hijos $R_i$ y $R_d$. El algoritmo de regresión consiste en dividir el espacio de las variables ($X_i$) en particiones homogéneas para obtener en la salida del árbol una predicción de la variable $Y$ en función de los valores de ($X_i$). Por ello, en cada nodo del árbol se buscan una variable $X_j$ y un umbral $S$ para dividir el nodo en dos nodos hijos $R_i$ y $R_d$:

$$R_i(j, S) = \{X | X_j < S\} \tag{1.13}$$

$$R_d(j, S) = \{X | X_j \geq S\} \tag{1.14}$$

El objetivo es encontrar los valores de $j$ y $S$ que minimizan la heterogeneidad resultante de la división de un nodo $R$, es decir maximizar la cantidad

$$\triangle i(s, R) = i(R) - P_i(R_i) - P_d(R_d) \tag{1.15}$$

donde $P_i$ y $P_d$ son la proporción de observaciones que pertenecen a los nodos $R_i$ y $R_d$ respectivamente, y $i(R)$ es una función que mide la «impureza» del nodo $R$. Las funciones más comunes para $i$ son la entropía de Shannon, el índice de Gini o la varianza.

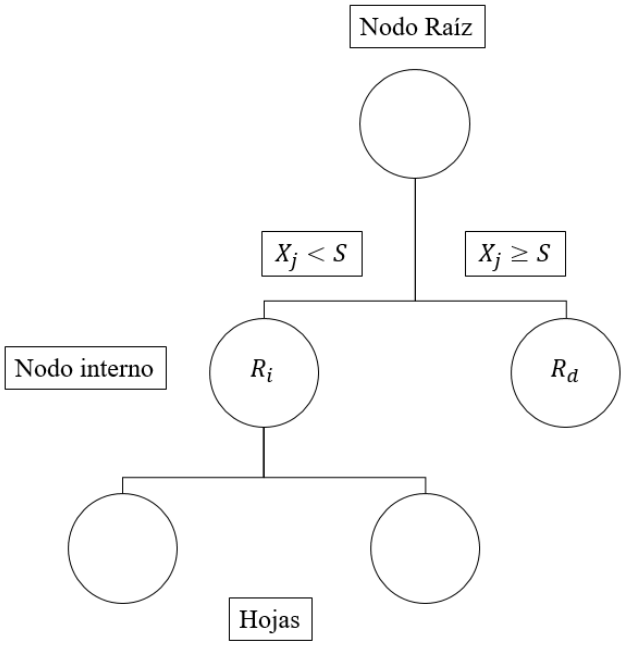


Figura 1.2: Representación gráfica de un árbol de regresión

El proceso de segmentación termina cuando no es posible dividir las observaciones que pertenecen a un nodo en conjuntos más homogéneos, o cuando la cantidad de observaciones en el nodo alcanza un límite inferior definido por el modelador.

### 1.2.2. Del árbol de regresión al bosque aleatorio

El árbol de regresión presenta la desventaja de ser inestable cuando existe ruido en las variables. Para proveer estabilidad al algoritmo se aplica un método de *bagging*, que consiste en promediar los resultados de muchos árboles de regresión para obtener un modelo muy estable y que además no presenta problemas de sobreajuste. A partir de la base de datos original se generan $n$ muestras aleatorias elaboradas según el método *bootstrap* (Efron, 1982) destinadas a construir $n$ árboles de regresión. Al dividir los nodos de los árboles en nodos hijos se utiliza un conjunto de $m < p$ variables predictoras en vez de las $p$ disponibles y finalmente se promedian los resultados entregados por todos los árboles para obtener la predicción final.

A partir de un bosque aleatorio es posible obtener la importancia de las variables $(X_i)$ en la predicción. Para medir la importancia de una variable $X_m$, Breiman (2001) propone promediar sobre los $N_T$ árboles del bosque las impurezas $\triangle i(R)$ de los nodos donde se utiliza la variable $X_m$ para dividir el nodo $R$:

$$Imp(X_m) = \frac{1}{N_T} \sum_{T} \sum_{R \in T: v(R) = X_m} p(R) \triangle i(R) \tag{1.16}$$

donde $v(R)$ es la variable utilizada para dividir el nodo $R$.

## 1.3. Máquinas de vectores de soporte para la regresión (SVR)

Las máquinas de vectores de soporte (Drucker et al., 1997) son un conjunto de algoritmos utilizados en problemas de clasificación y de regresión no lineal. Se dispone de una base de datos de entrenamiento $(x_1, y_1), ..., (x_n, y_l) \in \mathbb{R} \times \mathbb{R}^d$, donde $\mathbb{R}^d$ es el espacio de las variables predictoras y $\mathbb{R}$ es el espacio de una variable continua a predecir. El objetivo del algoritmo de regresión es encontrar los parámetros $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ que permitan aproximar la variable objetiva tolerando una desviación máxima $\epsilon$ y castigando los errores mayores a $\epsilon$ proporcionalmente a un factor $C$. En el caso lineal, el problema se puede escribir de la siguiente manera :

$$\text{minimizar} \quad \frac{1}{2}||w||^2 + C\sum_i \xi_i \tag{1.17}$$

$$\text{s.a} \quad \begin{cases} y_i - \langle w, x_i \rangle - b & \leq 0 \\ \langle w_i, x_i \rangle + b - y_i & \leq 0 \\ \xi & \geq 0 \end{cases} \tag{1.18}$$

donde la variable $\xi$ representa el error a castigar, es decir :

$$\xi = \begin{cases} 0 & \text{si } |\xi| \leq \epsilon, \\ |\xi| - \epsilon & \text{si no.} \end{cases}$$

El problema de minimización se puede resolver más fácilmente en su formulación dual, lo que además permitirá resolver casos no lineales. Cabe descatar que en la formulación del problema dual interviene el producto vectorial entre los distintos vectores de variables predictoras $\langle x_i, x_j \rangle$, como se muestra en la Ecuación 1.19 :

$$\text{maximizar} \quad \begin{cases} -\frac{1}{2}\sum_{i,j}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle \\ -\epsilon\sum_i(\alpha_i + \alpha_i^*) + \sum_i y_i(\alpha_i - \alpha_i^*) \end{cases} \tag{1.19}$$

$$\text{s.a} \quad \sum_i (\alpha_i - \alpha_i^*) = 0 \quad y \quad \alpha_i, \alpha_i^* \in [0, C]$$

En el caso no lineal, se busca procesar las variables mediante un mapeo no lineal $\phi$ desde el espacio $\mathbb{R}^d$ hacia un espacio de dimensión mayor, para aplicar el problema dual a las variables predictoras mapeadas $\phi(x_i)$. Sin embargo, este mapeo implica en el problema dual calcular el producto $\langle \phi(x_i), \phi(x_j) \rangle$, lo que representa un costo computacional importante. Para disminuir el número de operaciones necesarias, se busca una función $k$ que permita calcular directamente este valor sin pasar por un mapeo, es decir una función $k$ «implicita» tal que :

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle \tag{1.20}$$

Este método llamado «truco del kernel» permite efectuar regresiones no lineales, las más comunes siendo la regresión polinomial de grado $d$ y la regresión de tipo RBF (*Radial Basis Function*), para la cuales las funciones implícitas son $k(x_i, x_j) = (\langle x_i, x_j \rangle + c)^d$ y $k(x_i, x_j) = \exp(-\gamma||x_i - x_j||^2)$.

Por ejemplo, se considera un mapeo cuadrático $\phi$ desde un espacio de dimensión hacía un espacio de dimensión 3 tal como descrito en la Ecuación (1.21),

$$\phi : \mathbb{R}^2 \to \mathbb{R}^3$$

$$\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \tag{1.21}$$

El producto vectorial $\langle \phi(x), \phi(x') \rangle$ resultando de la formulación dual se puede calcular a partir de la expresión de la función de mapeo, lo que representa un número de operaciones importantes y resulta en un costo computacional elevado (ver Ecuación 1.22)

$$\begin{aligned} \langle \phi(x), \phi(x') \rangle &= \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2), (x_1'^2, \sqrt{2}x_1'x_2', x_2'^2) \rangle \\ &= x_1^2x_1'^2 + 2x_1x_1'x_2x_2' + x_2^2x_2'^2 \\ &= (x_1x_1' + x_2x_2')^2 \\ &= \langle x, x' \rangle^2 \end{aligned} \tag{1.22}$$

El método del kernel consiste en calcular directamente la función $k = \langle x, x' \rangle^2$ para reducir el número de operaciones necesarias y resolver el problema de minimización del algoritmo.

En el Cuadro 1.1 se muestran las ventajas y desventajas de la red neuronal, del bosque aleatorio y del algoritmo SVR para la regresión.

| Método | Ventajas | Desventajas |
|---|---|---|
| Red neuronal | Se pueden aproximar funciones complejas | Número elevado de parámetros a ajustar |
| | Capacidad de generalización | Riesgo de sobreajuste |
| | | «Caja negra » |
| | | Funciona bien con muestras grandes |
| | | No existe metodología para la elección de la arquitectura |
| | | Alto costo computacional |
| Bosque aleatorio | Se puede calcular la importancia de las variables | No se obtiene una función que relaciona la entrada y la salida |
| | No hay riesgo de sobreajuste | |
| | Fácil de implementar | |
| | Funciona con muestras más pequeñas | |
| | Se puede visualizar los árboles | |
| SVR | Se pueden representar funciones no lineales | No existe metodología para la elección de los hiperparámetros |
| | Capacidad de generalización | Alto costo computacional |
| | Robusto | El número de kernels es limitado |

Cuadro 1.1: Ventajas y desventajas de la red neuronal, del bosque aleatorio y de las máquinas de vectores de soporte para la regresión

## 1.4. Utilización de técnicas de inteligencia artificial para la predicción de rentas

En esta sección se recopilan los principales resultados obtenidos en los problemas de predicción de rentas mediante el uso de los tres algoritmos descritos y se compara la regresión hedónica con estas técnicas en término de poder predictivo.

### 1.4.1. El modelo hedónico

El modelo hedónico (Rosen, 1974) se basa en la teoría del consumidor y postula que un hogar, al valorar un bien inmobiliario, considera atributos propios del bien (número de

piezas, año de construcción) y factores externos (barrio, accesibilidad a los centros de salud, transporte), atribuyendo a cada una de estas características un valor monetario. De esta manera, el valor de una vivienda es obtenido agregando el valor de todas las características que la definen. El método más común para la aplicación del modelo hedónico es la regresión lineal de mínimos cuadrados ordinarios ($OLS$). La función de precios hedónicos se puede escribir de la siguiente forma :

$$y = f(\beta S, \alpha N) + \epsilon \tag{1.23}$$

donde $y$ es un vector de rentas, $S$ es la matriz de los atributos propios de los bienes, $N$ es la matriz de los factores externos de los bienes, $\beta$ y $\alpha$ son los parámetros que corresponden a $S$ y $N$ y $\epsilon$ es el término de error.

La función de regresión más utilizada para modelar la función de precios hedónicos es la lineal :

$$y = \delta X + \epsilon \tag{1.24}$$

donde $y$ es el precio de un bien, $X$ el vector de sus características internas y externas, $\beta$ el vector de los coeficientes de la regresión y $\epsilon$ el término de error.

El modelo hedónico lineal a sido ampliamente utilizado para la modelación de precios y permitió entender cómo los atributos de las viviendas y sus entornos contribuyen al valor de la renta. En términos de localización, se han calculado tradicionalmente medidas de accesibilidad que toman por referencia el centro de negocios (CBD por *Central Business District*) y que consisten en distancia o tiempos de viajes hacía el CBD en transporte privado y público. Se demostró que la accesibilidad al centro de negocios tiene efecto sobre el valor de la renta (Ridker y Henning, 1967) y que en particular la presencia de un servicio público de calidad contribuye a aumentar el valor de este (So, Tse y Ganesan, 1997).

En términos estructurales, la superficie habitable es el atributo más importante, junto con el número de habitaciones y baños. Garrod y Willis (1992) mostraron que la presencia de una habitación o un baño adicional sube la renta de un bien de 7 % y 14 % respectivamente, fenómeno que se puede explicar por la necesidad de las familias de tener suficientes espacios funcionales. También muestra que la renta aumenta con la presencia de patio, agua caliente y/o aire acondicionado. Por otro lado, la edad de un departamento afecta negativamente su precio porque se estima que que los costos de mantención aumentan con la antiguedad de este (Clark y Herrin, 2000).

El impacto de las características socio-económicas se estudia considerando bienes inmuebles estructuralmente parecidos que se encuentran en vecindades distintas. El estrato social de los habitantes de una zona afecta el valor de la renta de las viviendas (Richardson, Vipond y Furbey, 1974), junto con la calidad de la enseñanza pública (que se puede medir en términos de presupuesto o de resultados), la criminalidad (Clark y Herrin, 2000), y la accesibilidad a los centros de compras (Des Rosiers, Lagana, Thériault y Beaudoin, 1996).

Las aplicaciones del modelo hedónico a la ciudad de Santiago de Chile han permitido llegar a las siguientes conclusiones. Figueroa (1992) muestra que la superficie de terreno, la presencia de una pieza de servicio, el nivel socio-económico del vecindario y una alta densidad de construcción tienen un efecto positivo sobre los precios, mientras la densidad poblacional del barrio, la proximidad al centro de la ciudad y el número de habitaciones tienen un efecto

negativo. Este último resultado contradice la idea que una habitación adicional contribuye a aumentar el valor de una vivienda. Figueroa (1992) lo interpreta como una señal de que en Santiago predomina la correlación entre esta variable y la superficie de la vivienda, que al aumentar hace disminuir el precio del metro cuadrado. Altas tasas de criminalidad y de contaminación también contribuyen a disminuir el valor de un bien raíz (Lavın, Dresdner y Aguilar, 2011). Vega Flores (2017) encuentra un efecto positivo en el precio de las viviendas ubicadas cerca de ciclovías, parques, plazas y estaciones de metro. Por el contrario, muestra que las viviendas ven su valor afectado por la proximidad de corredores de buses, trazados de tren o Metro en superficie o elevado y de autopistas en superficies o elevadas.

### 1.4.2. Predicción de rentas con redes neuronales

La regresión hedónica ha sido por mucho tiempo el método de predicción tradicional, a pesar de restringir la relación causa efectos entre los atributos de las viviendas y su precio a ciertas funciones por lo que en algunos casos no se obtienen los resultados esperados (Tay y Ho, 1992). Al principio de los años 90 se empezó a utilizar como alternativa a la regresión hedónica las redes neuronales, que presentan la ventaja de poder representar fenómenos no lineales sin limitarse a un número finito de formas funcionales. A partir de los errores absolutos porcentuales medios obtenidos por las redes neuronales y las técnicas de regresión se formaron conclusiones contradictorias. El error absoluto porcentual medio (MAPE por *Mean Absolute Percentage Error*) se define como :

$$MAPE = \frac{100}{n} \sum_i \frac{y_i - \hat{y}_i}{\hat{y}_i} \tag{1.25}$$

donde $n$ es el número de observaciones, $y_i$ es el valor real de la renta o del precio del metro cuadrado de los bienes inmuebles e $\hat{y}_i$ es su valor predicho por el bosque aleatorio.

Tay y Ho (1992) y Do y Grudnitski (1992) muestran que las redes neuronales tienen mejor poder predictivo que la regresión, y en el último caso se obtiene un error absoluto porcentual medio para la red neuronal y la regresión de 6,9 % y 11,3 % respectivamente a partir de una muestra de 105 transacciones. Los resultados obtenidos por Worzala, Lenk y Silva (1995) y Allen y Zumwalt (1994) sugieren lo contrario.

Además de la heterogeneidad de los resultados obtenidos por los distintos equipos de investigación, se ha cuestionado el uso de la red neuronal para la predicción de rentas, en particular por Worzala et al. (1995). A partir de una muestra de 288 transacciones efectuadas en Colorado EE.UU, comparan los resultados obtenidos por una regresión y dos redes neuronales implementadas en dos lenguajes de programación distintos, y concluyeron que la redes neuronales superan la regresión para la predicción de rentas. Además, advirtieron que los resultados obtenidos por las redes neuronales fueron distintos según el software utilizado y que siendo poco sensibles a valores singulares, las redes neuronales son poco adaptadas a las muestras de datos pequeñas.

Por otra parte, las redes neuronales son criticadas por su dificultad de interpretación en comparación con el modelo de regresión hedónica. En efecto, el gran número de parámetros de

la red y su estructura compleja han sido citados como motivos para considerarlas como caja negra, a pesar de la existencia de varias técnicas para interpretarlas, recopiladas por Gevrey, Dimopoulos y Lek (2003). En este contexto, las nuevas técnicas de regresión elaboradas en los años noventas tales como los bosques aleatorios (Breiman, 2001) y las máquinas de soporte para la regresión (Drucker et al., 1997) representaron unos serios competidores a las redes neuronales para la predicción de rentas.

### 1.4.3. Predicción de rentas con el algoritmo SVR y el bosque aleatorio

La utilización de los bosques aleatorios y de las máquinas de vectores de soporte para la predicción de rentas es novedosa por lo que la literatura al respecto es más escasa. Antipov y Pokryshevskaya (2012) compararon once métodos distintos para la predicción de rentas en la ciudad de San Petersburgo a partir de una muestra de 2848 departamentos de dos habitaciones vendidos en el año 2010, y en particular compararon el error absoluto porcentual medio obtenido con una regresión hedónica, un bosque aleatorio y una red neuronal con funciones de activación de tipo sigmoid.

Mostraron que los algoritmos tienen mejor poder predictivo cuando la variable a predecir es el valor del metro cuadrado y no el valor total del bien raíz. Los resultados obtenidos en el primer caso, presentados en el Cuadro 1.2, permiten concluir que el bosque aleatorio tiene el mejor poder predictivo, seguido por la red neuronal y la regresión hedónica.

| Algoritmo | MAPE ( %) |
|---|---|
| Bosque Aleatorio | 14,86 |
| Red Neuronal | 16,9 |
| Regresión hedónica | 18,33 |

Cuadro 1.2: Error absoluto porcentual medio obtenido con la red neuronal, el bosque aleatorio y la regresión hedónica obtenido por Antipov y Pokryshevskaya (2012).

Čeh, Kilibarda, Lisec y Bajat (2018) compararon los resultados obtenidos con el bosque aleatorio y la regresión hedónica para la predicción de rentas en Ljubljana, Slovenia, a partir de una muestra de 7404 transacciones de departamentos efectuadas entre 2008 y 2013. Los valores de los errores absolutos porcentuales medios obtenidos (Cuadro 1.3) muestran que el algoritmo de bosque aleatorio es más apropriado que la regresión hedónica para la predicción de rentas y confirman que este problema es no lineal.

| Algoritmo | MAPE (%) |
|---|---|
| Bosque Aleatorio | 7,27 |
| Regresión hedónica | 17,48 |

Cuadro 1.3: Error absoluto porcentual medio obtenido con el bosque aleatorio y la regresión hedónica obtenido por Čeh, Kilibarda, Lisec y Bajat (2018).

Lin y Chen (2011) aplicaron el algoritmo de máquinas de vectores de soporte y una red neuronal al problema de predicción de rentas en Taiwan, y mostraron que el método de máquinas de vectores de soporte para la regresión supera la red neuronal (errores absolutos porcentuales medios de 4,46 % y 5,80 % respectivamente).

En los dos casos, la red neuronal no logra superar el bosque aleatorio y las máquinas de vectores de soporte para la regresión, lo que se podría explicar por el tamaño de las muestras utilizadas. La muestra utilizada por Čeh et al. (2018) contiene 7404 observaciones, un tamaño pequeño para una red neuronal pero adaptado a los otros algoritmos.

El uso de los bosques aleatorios y de las redes neuronales permite deshacerse de la necesidad de definir una forma funcional que relacione un vector de atributos ($X$) con un valor objetivo $Y$, lo que presenta ventajas cuando se busca representar un fenómeno complejo y no lineal tal como la relación entre una vivienda y su precio. El modelo hedónico, a pesar de ser fácilmente interpretable, se limita a un número finito de formas funcionales para modelar la compleja relación causa efecto entre los atributos de una vivienda y su precio. Los algoritmos de aprendizaje automático se pueden aplicar a la predicción de rentas inspirándose del modelo hedónico, es decir relacionando el precio de una vivienda $Y$ con el vector ($X$) de sus atributos internos y externos. Aplicando por primera vez en Santiago de Chile los algoritmos de bosque aleatorio, red neuronal y máquinas de vectores de soporte para la regresión a una base de datos que contiene transacciones de viviendas junto con sus atributos, se busca obtener resultados comparables con los resultados presentados en la literatura y analizar la importancia de los atributos de manera similar al análisis de los coeficientes de la regresión hedónica.

# Capítulo 2

# Recolección y procesamiento de los datos

La base de datos necesaria para desarrollar este trabajo fue obtenida mediante un convenio con la empresa TocToc.com. Esta empresa ofrece un servicio que, a través de una página web, permite acceder a la información de cualquier bien inmobiliario de la ciudad de Santiago, además de una estimación de la renta y del arriendo de dicho bien. La estimación de la renta de los bienes es calculada mediante un modelo de comparación entre bienes similares y espacialmente cercanos y no provee resultados aceptables cuando las observaciones son escasas en alguna localización, lo que indujo la empresa a buscar nuevas alternativas para la predicción de las rentas.

La base de datos contiene 600.902 observaciones que corresponden a transacciones de casas o departamentos efectuadas entre el año 2007 y el año 2018, junto a una lista de características para cada bien vendido. Las transacciones se realizaron en la Región Metropolitana en las comunas de Cerrillos, Cerro Navia, Conchalí, Estación Central, Huechuraba, Independencia, La Cisterna, La Florida, La Granja, La Pintana, La Reina, Las Condes, Lo Barnechea, Lo Espejo, Lo Prado, Macul, Maipú, Ñuñoa, Pedro Aguirre Cerda, Padre Hurtado, Peñalolén, Providencia, Pudahuel, Puente Alto, Quinta Normal, Recoleta, Renca, San Bernardo, San Joaquín, San Miguel, San Ramón, Santiago Centro y Vitacura, lo que representa un total de 33 comunas. Cada vivienda es localizada por sus coordenadas geográficas, además de tener asociada una manzana, una comuna y una zona homogénea en términos del mercados, delimitada por TocToc.com.

La base de datos cubre todas las viviendas vendidas y registradas por el Servicio de Impuestos Internos, por lo que busca representar la realidad del mercado inmobiliario Santiaguino en los once últimos años. Sin embargo, el valor de la renta para un mismo bien puede depender del contexto de la transacción en que se realiza. Por ello, es necesario tomar en cuenta que existe un fenómeno masivo de transacciones de bienes entre personas de una misma familia, y que el precio de una casa puede variar dependiendo si está destinada a ser habitada o destruida para la construcción de un proyecto inmobiliario. Además, las informaciones de superficies de construcción y de terreno pueden ser no actualizadas en la base de datos del Servicio de Impuestos Internos cuando se efectúan extensiones sin permisos, hecho que generalmente ocurre en las comunas de bajo recursos.

A pesar de desconocer los motivos de las compras y los casos de construcción sin permiso, TocToc.com elaboró una metodología para identificar las transacciones que no representan el mercado basada en el análisis de la distribución de los valores de las transacciones efectuadas en una misma zona homogénea, sobre periodos temporales de un semestre. Este método consiste en calcular el promedio del valor del metro cuadrado para cada zona diferenciando las casas y los departamentos, y calcular la desviación estándar de este valor para cada tipo de propiedad. Se generan límites de aceptabilidad en función de la desviación estándar del valor del metro cuadrado para cada zona de la siguiente forma:

- Si la desviación estándar de la zona se encuentra entre 0 y 20 UF/$m^2$, entonces el límite es el promedio $\pm$ la desviación estándar.
- Si la desviación estándar de la zona se encuentra entre 20 y 40 UF/$m^2$ entonces el límite es el promedio $\pm\frac{1}{2}$ la desviación estándar.
- Si la desviación estándar de la zona es más de 40 UF/$m^2$ entonces el límite es el promedio $\pm\frac{1}{4}$ la desviación estándar.

Se considera que las propiedades fuera de rango presentan singularidades y no representan el mercado, resultado que luego es revisado en base a criterio experto.

Después de aplicar esta metodología, la base de datos se reduce a 350.127 observaciones verificadas. En el Cuadro 2.1 se muestran las comunas de Santiago que aparecen en la base de datos, junto al número de transacciones verificadas disponibles para cada comuna.

| Comuna | Dptos | Casas | Comuna | Dptos | Casas |
|---|---|---|---|---|---|
| Cerrillos | 925 | 1689 | Maipú | 3926 | 31287 |
| Cerro Navia | 251 | 2120 | Ñuñoa | 25504 | 2958 |
| Conchalí | 626 | 2236 | P. Aguirre Cerda | 429 | 1377 |
| Estación Central | 2781 | 1998 | Peñalolén | 1584 | 8518 |
| Huechuraba | 788 | 6319 | Providencia | 19916 | 849 |
| Independencia | 6299 | 1123 | Pudahuel | 2194 | 8429 |
| La Cisterna | 2474 | 1272 | Puente Alto | 1831 | 27621 |
| La Florida | 8529 | 14425 | Quinta Normal | 4981 | 1551 |
| La Granja | 432 | 2080 | Recoleta | 3707 | 2249 |
| La Pintana | 883 | 3629 | Renca | 932 | 5921 |
| La Reina | 1131 | 5200 | San Bernardo | 2151 | 6272 |
| Las Condes | 26762 | 8704 | San Joaquín | 1231 | 1370 |
| Lo Barnechea | 2281 | 6236 | San Miguel | 6693 | 752 |
| Lo Espejo | 268 | 981 | San Ramón | 293 | 988 |

| | | | | | |
|---|---|---|---|---|---|
| Lo Prado | 940 | 983 | Santiago Centro | 62377 | 872 |
| Macul | 5701 | 2539 | Vitacura | 6782 | 3230 |

Cuadro 2.1: Número de transacciones verificadas efectuadas entre 2007 y 2018 de departamentos y casas por comunas

## 2.1. Variables consideradas

A partir base de datos disponible, se busca estimar un precio para cada vivienda a partir de sus características, sino también considerándola parte de un contexto social y espacial en la ciudad. Para eso, es necesario, además de considerar atributos propios, elegir variables de entorno que podrían afectar el valor de un bien y explicar las diferencias de precio que existen a lo largo de la ciudad. La información disponible permite describir los bienes a distintas escalas que se pueden definir como escalas del bien, escala de manzana, escala de zona y escala de ciudad, detalladas a continuación.

### 2.1.1. Escala de la vivienda

Los atributos cada vivienda están recopilados por el Servicio de Impuestos Internos. Se dispone del año de construcción, de la superficie de construcción y de terreno en el caso de las casas, de una variable que permite saber si el bien es nuevo o no y de un índice de calidad. Este índice es calculado por el Servicio de Impuesto Interno y depende del material de construcción, de la estructura y del tamaño de las piezas, entre otros atributos. Un índice de calidad de 1 corresponde al mejor puntaje, mientras un índice de calidad de 5 representa el puntaje más bajo.

### 2.1.2. Escala de la manzana

Para cada manzana se calculó la distancia al centro de salud privado y al colegio privado más cercano. Además, cada manzana tiene asociada el resultado de la evaluación Simce (Sistema de Medición de la Calidad de la Educación) elaborada por la Agencia de la Calidad de la Educación (A.C.E), obtenido en los colegios privado, particular pagado y público más cercano. El valor del Simce utilizado para la predicción es el promedio de los resultados obtenidos en los tres tipos de colegios.

En términos de áreas verdes, el Centro de Políticas UC definió en la mesa de áreas verdes 2017 un Indicador de Accesibilidad Urbana a Áreas Verdes (IAUAV), que considera dos variables generalmente medidas por separado: la capacidad de las áreas verdes, que se definen por la superficie de áreas verdes por habitantes a nivel comunal o regional, y una medida de accesibilidad. La utilización de estas dos variables para la construcción del indicador permite

trabajar a escala de manzana y capturar las desigualdades existentes dentro de una misma comuna.

Primero se distribuyen todas las áreas verdes de la ciudad en radios de influencia determinados según su tamaño, como es presentado en el Cuadro 2.2. Luego se dividen los metros cuadrados de áreas verdes distribuidos por el total de población a la que benefician. El resultado del cálculo es la accesibilidad de esa manzana. Para este análisis se utilizó la información pública existente del año 2018 para los parques (Ministerio de Vivienda y Urbanismo) y del 2013 para las plazas (Ministerio de Vivienda y Urbanismo, Comisión Asesora de Estudios Habitacionales y Urbanos).

| Tipología de área verde | Superficie ($m^2$) | Modo de desplazamiento | Tiempo de desplazamiento |
|---|---|---|---|
| Plaza menor | 500 - 5000 $m^2$ | Caminata | 5 min |
| Plaza mayor | 5000 - 20 000 $m^2$ | Caminata | 10 min |
| Parque menor | 2 - 10 ha | Transporte público | 10 min |
| Parque mayor | 10 ha o más | Transporte público | 20 min |

Cuadro 2.2: Consideración de las areas verdes. Fuente: Centro de Políticas UC

A partir de los resultados obtenidos se graficó el indicador de accesibilidad urbana a áreas verdes obtenido con el método anterior (Figura 2.1). El Gran Santiago cuenta con 4,7 $m^2$ de áreas verdes por habitantes, con grandes disparidades entre las comunas: mientras los residentes de la comuna de Lo Barnechea disponen de 11,7 $m^2$ de áreas verdes por habitantes, esta cifra cae a 2,3 $m^2$ por habitantes en El Bosque.

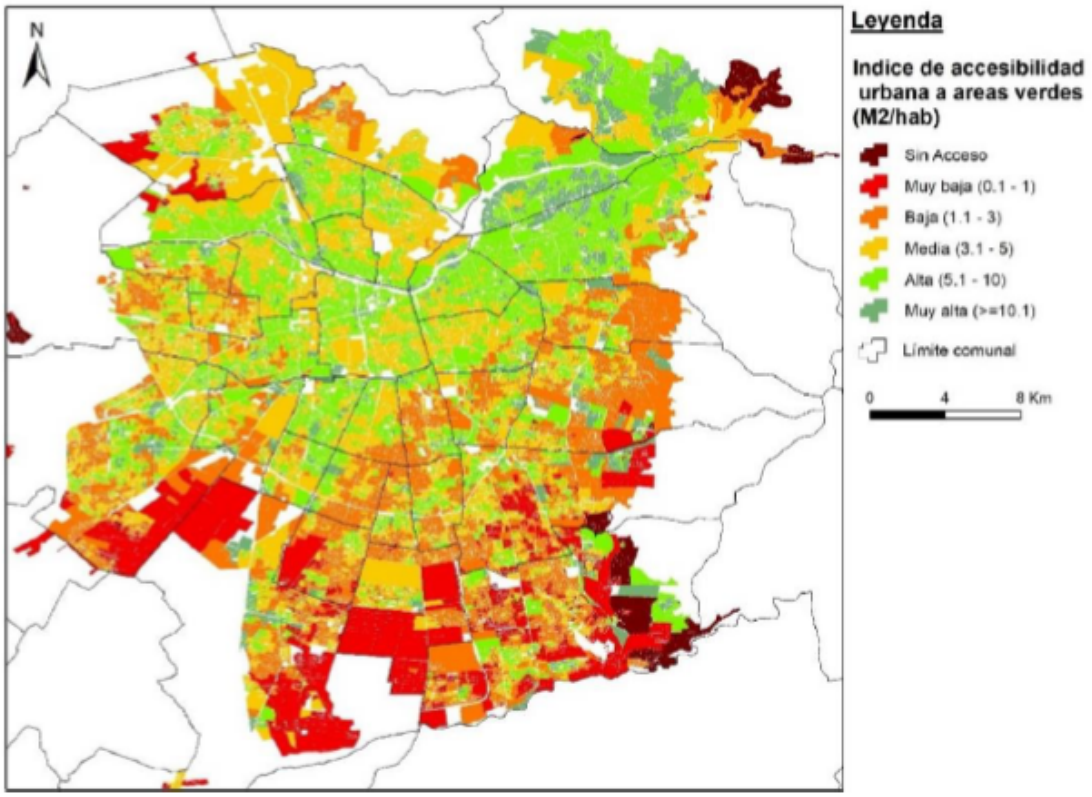


Figura 2.1: Representación gráfica del IAUAV en el Gran Santiago. Fuente : Centro UC Políticas Públicas

### 2.1.3. Escala de zona

Para cada zona homogénea se promediaron las variables Simce, áreas verdes y delincuencia, con el objetivo de tener una medida más global de estas variables.

### 2.1.4. Escala de ciudad

Para cada vivienda se calcularon las distancias basadas en la red vial de la ciudad a nueve puntos considerados centros urbanos por proveer servicios a la comunidad tales como bancos, centros de salud y colegios, o por la presencia de empresas. Los nueve centros urbanos escogidos son Plaza de Maipú, Plaza de Quilicura, Plaza de San Bernardo, Plaza de Puente Alto, Bellavista de la Florida, Metro Las Mercedes, Plaza de Armas, Metro Tobalaba y Metro Manquehue, cuya ubicación se muestra en la Figura 2.2.

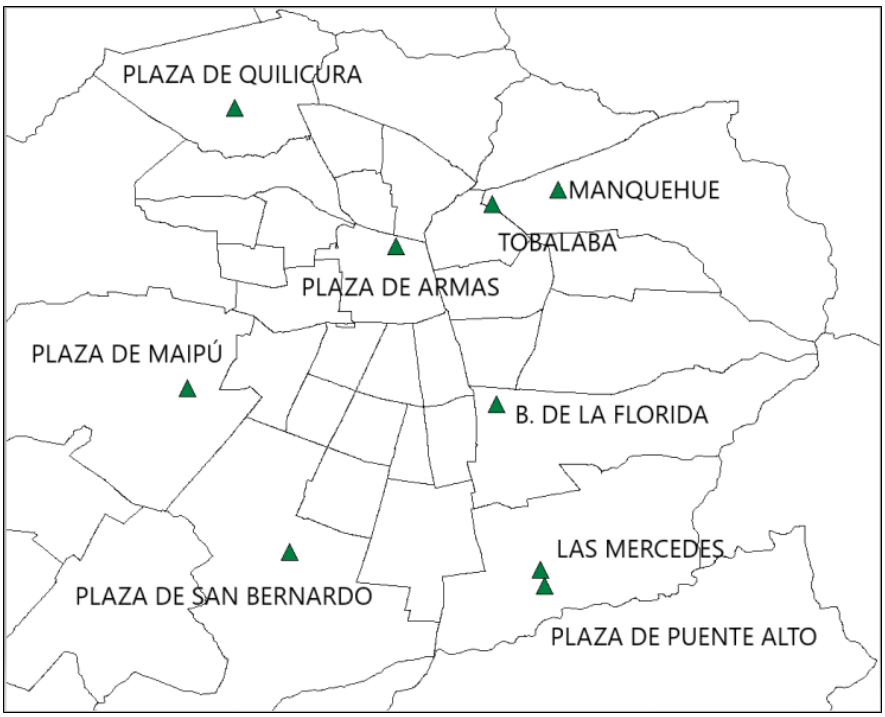


Figura 2.2: Ubicación de los centros urbanos considerados

Además se calculó para cada vivienda la distancia basada en la red vial de la ciudad a la estación de Metro más cercana, tomando en cuenta los distintos cambios que se han efectuado en la red de Metro entre 2007 y 2018: extensión de la línea 1 de Escuela Militar a Los Dominicos en 2009, extensión de la línea 5 de Plaza de Maipú a Quinta Normal en 2011, la construcción de la estación San José de la Estrella en 2009 y la inauguración de la línea 6 en el año 2017. Además se consideró la inauguración del MetroTren Nos en el año 2017.

En el Cuadro 2.3 se resumen las variables utilizadas para la predicción de las rentas.

| Variable | Promedio | Min | Max | Desv. Est | Unidad | Origen Dato |
|---|---|---|---|---|---|---|
| Año de venta | 2012,4 | 2007,0 | 2018,0 | 3,4 | - | SII |
| Precio | 2752,2 | 303,0 | 171666,0 | 2711,4 | UF | SII |
| Año construcción | 2003,5 | 1915,0 | 2017,0 | 13,5 | - | SII |
| Id calidad | 2,8 | 1,0 | 5,0 | 0,7 | - | SII |
| Superficie | 61,9 | 22,0 | 744,0 | 33,3 | m2 | SII |
| Pl. de Armas | 6402,2 | 34,0 | 43785,3 | 4528,3 | m | TocToc |
| Tobalaba | 7120,5 | 88,3 | 48371,0 | 4600,4 | m | TocToc |
| Manquehue | 9237,2 | 39,9 | 50848,1 | 5252,1 | m | TocToc |
| B. de la Florida | 10985,2 | 187,3 | 37077,6 | 3921,2 | m | TocToc |
| Pl. Puente Alto | 20529,7 | 130,7 | 35632,8 | 4540,0 | m | TocToc |

| | | | | | | |
|---|---|---|---|---|---|---|
| Metro las Mercedes | 19457,8 | 274,5 | 34544,5 | 4492,5 | m | TocToc |
| Pl. San Bernardo | 19883,8 | 146,8 | 38742,0 | 5023,2 | m | TocToc |
| Pl. Maipú | 16467,5 | 582,9 | 38983,8 | 4846,5 | m | TocToc |
| Pl. Quilicura | 16816,8 | 3712,7 | 56012,9 | 4522,5 | m | TocToc |
| Distancia Metro | 1409,8 | 1,2 | 29753,9 | 1515,2 | m | TocToc |
| A. Verdes Manzana | 5,1 | 0,0 | 320,0 | 5,3 | m2 | C. P. UC |
| Delinc. Manzana | 6,3 | 1,0 | 10,0 | 2,9 | - | Carabineros |
| Dist. Colegio Privado | 1148,5 | 0,0 | 8625,6 | 1068,8 | m | TocToc |
| Simce Manzana | 270,1 | 200,0 | 322,7 | 16,5 | - | A.C.E |
| Edad Manzana | 1993,4 | 1917,0 | 2012,0 | 17,7 | - | SII |
| Calidad Manzana | 3,4 | 1,5 | 5,0 | 0,3 | - | SII |
| Dist. Salud Privada | 2769,9 | 18,5 | 21428,2 | 3121,6 | m | TocToc |
| Áreas Verdes Zonas | 5,0 | 0,0 | 49,3 | 3,8 | m2 | C. P. UC |
| Simce Zona | 270,1 | 217,6 | 309,4 | 11,6 | - | A.C.E |
| Delincuencia Zona | 6,3 | 1,0 | 10,0 | 2,3 | - | Carabineros |

Cuadro 2.3: Descripción de las variables utilizadas

La ciudad de Santiago no es homogénea en término de precios del suelo, como se puede ver en la Figura 2.3. En el año 2017, el precio del suelo más alto se observó en las comunas de Vitacura y lo Barnechea, y el precio del suelo más bajo en La Pintana y el Espejo, de acuerdo con la heterogeneidad espacial que caracteriza la ciudad en términos de ingresos. La riqueza se concentra en la zona oriente de la ciudad, en las comunas de Providencia, Las Condes, Vitacura, Ñuñoa, La Reina y Lo Barnechea, mientras las comunas más pobres, tales como Lo Espejo, San Joaquín y Renca, se ubican en las zonas sur y norte poniente, por lo que existe segregación espacial en la ciudad. Vargas (2006) explica este hecho por las políticas urbanas aplicadas al final de los años 90 que consistían en concentrar en las periferias de la ciudad las viviendas adquiribles con subvenciones del estado, es decir las viviendas sociales.

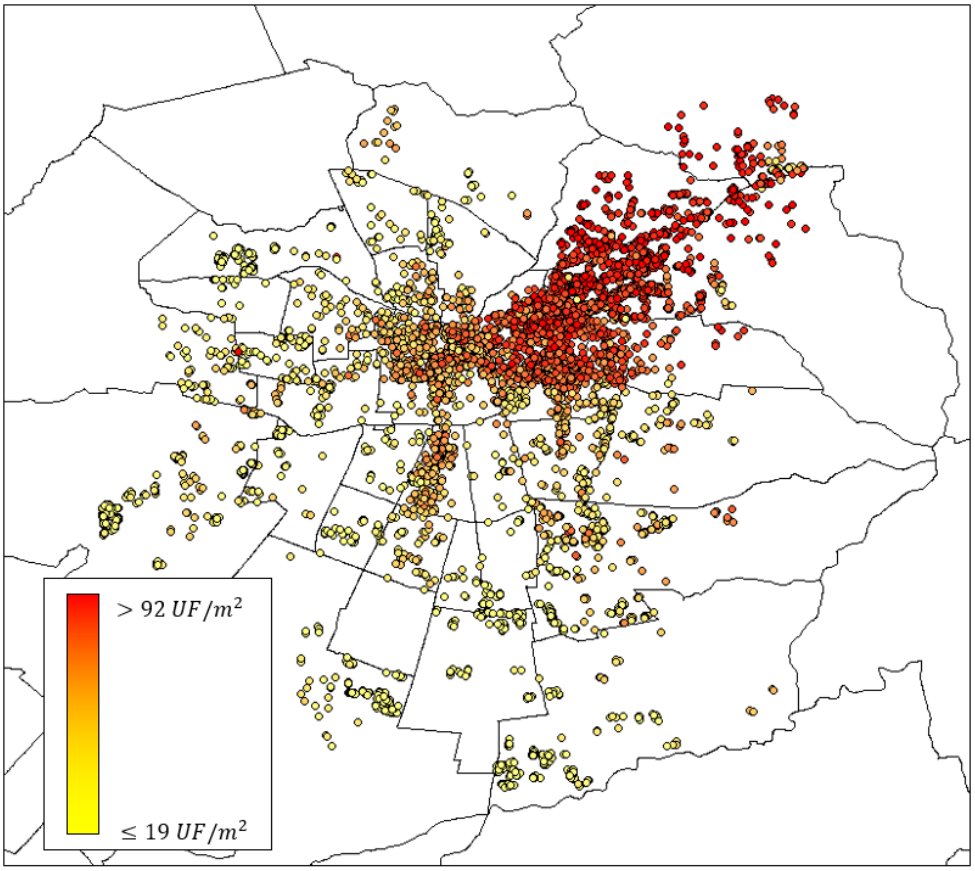


Figura 2.3: Representación espacial de los precios del suelo en Santiago en el año 2017

Le heterogeneidad de precios del suelo en la ciudad se relaciona con la heterogeneidad socio-económica de sus habitantes. Las comunas donde el precio del suelo es más bajo se caracterizan por la escasez de áreas verdes y de centros de salud privada, viviendas de baja calidad y departamentos más antiguos y más pequeños que el promedio de la ciudad, y coinciden con las comunas de menor ingreso por hogar. Las comunas donde el precio del suelo es más alto disponen de más áreas verdes, presentan mejores resultados a la prueba Simce que el promedio de la ciudad y coinciden con las comunas de mayor ingreso. En estas zonas, las superficies construidas por bien inmuebles son dos veces mayor al resto de la ciudad.

### 2.1.5. Observaciones

La base de datos contiene elementos que pueden ser fuente de errores en la muestra. Por ejemplo, la metodología para verificar las transacciones se basa en la desviación estándar de las rentas para determinar y eliminar los valores singulares mientras estos podrían ser informativos para los algoritmos de predicción de rentas. Además, es posible que las informaciones recopiladas por el Servicio de Impuestos Internos no representen la realidad en terreno cuando existen construcciones informales, en particular en las zonas de bajos ingresos. Finalmente, la falta de algunas variables descriptivas tales como información sobre los compradores y los habitantes de las distintas zonas, para agregar a los atributos de los departamentos el nivel

socio-económico del vecindario.

A partir de las variables recopiladas, se espera predecir el monto de las transacciones efectuadas en Santiago entre 2007 y 2018 con un error absoluto porcentual medio aceptable, es decir en los rangos de los obtenidos por Antipov y Pokryshevskaya (2012), Čeh et al. (2018) en el caso del bosque aleatorio, y por Lin y Chen (2011) y Tay y Ho (1992) para el algoritmo SVR y la red neuronal. Además, se espera capturar efectos de entorno tales como la influencia del acceso al transporte público o la presencia de áreas verdes en el valor de la renta. Estos resultados son posibles de obtener mediante los métodos descritos en las secciones 1.1.1 y 1.2.1 de este trabajo.

# Capítulo 3

# Comparación del poder predictivo de los algoritmos

En esta sección se describe la metodología de comparación de los tres algoritmos y su implementación mediante el lenguaje de programación *Python*. Se trabaja con una muestra de 8529 transacciones de departamentos efectuadas entre 2007 y 2018 en la comuna de La Florida, cuya distribución geográfica esta representada en la Figura 3.1. Esta comuna aloja un centro de servicios y de comercios potenciado por la llegada de la Linea 5 del Metro de Santiago y tiene un carácter principalmente residencial.

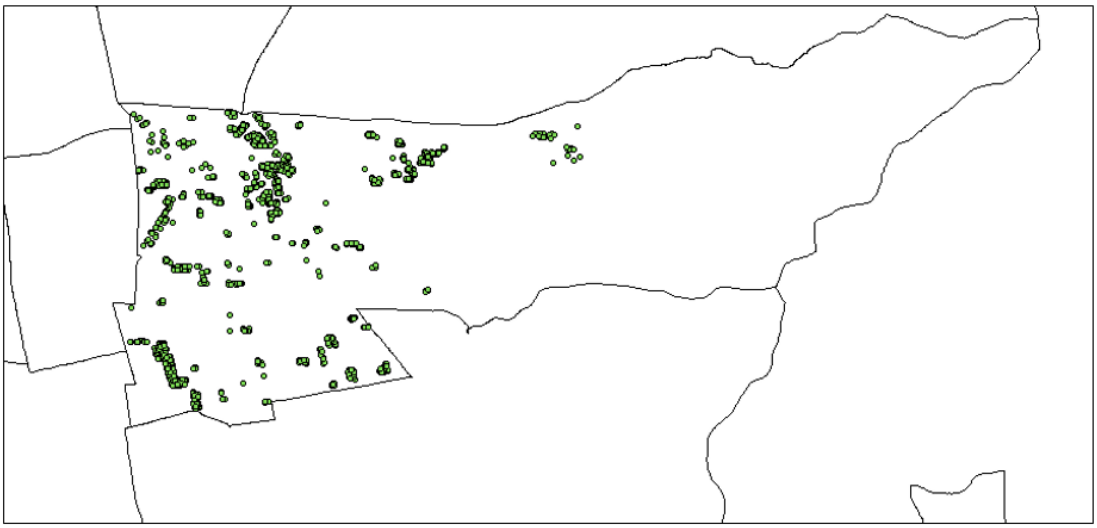

Figura 3.1: Representación espacial de las transacciones efectuadas en la florida entre 2007 y 2018

## 3.1. Medición del poder predictivo de los algoritmos

A partir de una muestra de observaciones, los métodos utilizados se implementan ajustando los parámetros de una función en el caso de las máquinas de vectores de soporte y de la red neuronal, o construyendo árboles de decisión en el caso del bosque aleatorio. Para comprobar los resultados obtenidos con los algoritmos, es necesario separar la base de datos en dos subconjuntos de observaciones. Una parte de la base de datos constituye el set de aprendizaje, con el cual los modelos ajustan sus parámetros, y la otra parte constituye el set de verificación desconocido de los modelos. Se utiliza un 80 % de las observaciones para la fase de aprendizaje, y el 20 % restante para la fase de verificación. La medición del error se obtiene corriendo los algoritmos entrenados con la base de verificación y comparando la predicción de la renta obtenida con la renta observada.

En este trabajo se usa como medición del error absoluto porcentual medio, que se define como:

$$\epsilon = \frac{1}{n} \sum_{i} \frac{y_{real} - y_{pred}}{y_{real}} * 100 \tag{3.1}$$

donde $y_{real}$ es el valor de la renta observado, $y_{pred}$ es la renta es el valor predicho por el algoritmo y $n$ es el tamaño de la muestra de validación.

Existe un riesgo de dependencia entre la partición de la base de datos y los resultados entregados por los distintos modelos si la base de aprendizaje no es representativa de la base completa. Para evitar este problema se aplica el método de validación cruzada, que consiste en aplicar los modelos a particiones distintas de la muestra y asegurarse de que el poder predictivo de los modelos no cambie al correrlos con las distintas particiones generadas. En la práctica, se elaboran cinco particiones de la base de datos completa de tamaño $n$, detalladas en el Cuadro 3.1.

| set n° | Obs. muestra de aprendizaje | Obs. muestra de validación |
|---|---|---|
| set n°1 | $[1; \frac{4}{5}n]$ | $[\frac{4}{5}n; n]$ |
| set n°2 | $[\frac{1}{5}n; n]$ | $[1; \frac{1}{5}n]$ |
| set n°3 | $[\frac{2}{5}n; n] \bigcup [1; \frac{1}{5}n]$ | $[\frac{1}{5}n; \frac{2}{5}n]$ |
| set n°4 | $[\frac{3}{5}n; n] \bigcup [1; \frac{2}{5}n]$ | $[\frac{2}{5}n; \frac{3}{5}n]$ |
| set n°5 | $[\frac{4}{5}n; n] \bigcup [1; \frac{3}{5}n]$ | $[\frac{3}{5}n; \frac{4}{5}n]$ |

Cuadro 3.1: Estructura de las cinco particiones de la muestra para la validación cruzada

El error porcentual que representa el poder predictivo de los modelos se define como el promedio de los errores absolutos porcentuales medios obtenidos a partir de las cinco particiones.

## 3.2. Implementación de los métodos

Los métodos de inteligencia artificial requieren una fase de ajuste de sus parámetros para alcanzar su funcionamiento óptimo, además de una etapa de procesamiento de la base de datos en el caso de la red neuronal y del algoritmo de máquinas de vectores de soporte. En esta sección se describe este proceso.

### 3.2.1. Red Neuronal

La red neuronal propuesta fue implementada mediante la librería *TensorFlow*, utilizando el lenguaje de programación *Python*. A continuación se describe el proceso de construcción de la red neuronal y el método que permite extraer de la red la importancia de cada variable en la predicción de la renta de los bienes.

#### Normalización

como fue descrito en la sección 1.1.1, una función de activación de tipo semi lineal, también llamada $ReLU$, es aplicada a la salida de cada nodo y le asigna su propio valor si la salida es positiva, o el valor nulo sino. Para no perder información al entrenar la red neuronal con una base de datos, es necesario normalizar las observaciones primero, por dos principales razones. Si no se normalizan los datos y la base contiene valores negativos, entonces la respuesta de la red a estos valores es nula, por la forma de la función de activación. Por otro lado, las variables tienen que estar todas en el mismo rango de valores ya que en el caso contrario las variables con menor amplitud serían consideradas menos importantes por la red neuronal que las variables de mayor amplitud. En efecto, dado que el algoritmo de minimización del error de la red funciona ajustando los parámetros $w_j$ y $b_j$, que ponderan las salidas de las variables de la forma $h_j = \sum_i w_{ij}x_i + b_j$ (ver Ecuación 1.1), entonces a mayor $x_i$ mayor el peso de esta variable en el resultado final. En la práctica, a cada variable se le aplica la siguiente normalización:

$$x_{inorm} = \frac{x_i - x_{imin}}{x_{imax} - x_{imin}} \tag{3.2}$$

donde $x_i$ es el valor de la variable $i$, $x_{imin}$ y $x_{imax}$ el mínimo y el máximo de la variable $i$ respectivamente y $x_{inorm}$ el valor normalizado.

#### Arquitectura de la red neuronal

No existen métodos para elaborar la arquitectura de la una red neuronal, ya que esta depende del número de variables explicativas y de la complejidad de la relación causa efecto que une las variables del problema de predicción. La red utilizada contiene una capa de entrada, una capa de atención, dos capas ocultas de 60 neuronas seguidas por dos capas ocultas de 12 neuronas y una capa de salida. Además, se utiliza un método de regularización de tipo L1, descrito en la sección 1.1.1.

**Inicialización de los parámetros**

De acuerdo a la sección 1.1.2, los pesos iniciales de la red siguen una distribución normal centrada en 0 y de varianza igual a $W \sim N(0, \sqrt{\frac{2}{n_l}})$, donde $n_l$ es el número de nodos de las capas donde tienen efecto los coeficientes, mientras las constantes iniciales son nulas. En la red neuronal propuesta, los coeficientes siguen la distribución siguiente :

| Capa | Nº de neuronas | Inicialización | Función de activación |
|---|---|---|---|
| Entrada | 25 | $W \sim N(0, \sqrt{\frac{2}{25}})$ | ReLU |
| Atención | 25 | $W \sim N(0, \sqrt{\frac{2}{25}})$ | Softmax |
| Oculta 1 | 60 | $W \sim N(0, \sqrt{\frac{2}{60}})$ | ReLU |
| Oculta 2 | 60 | $W \sim N(0, \sqrt{\frac{2}{60}})$ | ReLU |
| Oculta 3 | 12 | $W \sim N(0, \sqrt{\frac{2}{12}})$ | ReLU |
| Oculta 4 | 12 | $W \sim N(0, \sqrt{\frac{2}{12}})$ | ReLU |
| Salida | 1 | $W \sim N(0, \sqrt{2})$ | ReLU |

Cuadro 3.2: Estructura de la red neuronal utilizada

### 3.2.2. Maquinas vectores de soporte

El algoritmo de máquinas de vectores de soporte fue implementado utilizando el lenguaje de programación *Python* y la librería para el aprendizaje de máquinas *Sklearn.* como fue descrito en la sección 1.3, el problema de minimización a resolver con el algoritmo de máquinas de vectores de soporte se puede escribir :

$$\text{Minimizar} \quad \frac{1}{2}||w||^2 + C\sum_i \xi_i \tag{3.3}$$

$$\text{s.a} \quad \begin{cases} y_i - \langle w, x_i \rangle - b & \leq 0 \\ \langle w_i, x_i \rangle + b - y_i & \leq 0 \\ \xi & \geq 0 \end{cases} \tag{3.4}$$

donde los parámetros C y epsilon son escogidos por el modelador. No existe un método para la elección de los parámetros por lo que tras varios corridas con distintas combinaciones de $C$ y $\epsilon$ se eligió $C = 1,8$ y $\epsilon = 0,01$.

Al igual que en el caso de la red neuronal, es necesario normalizar todas la variables del problema dentro del mismo rango de valores, en este caso $(0, 1)$, para que sean todas tratadas con la misma importancia al correr el algoritmo.

### 3.2.3. Bosque aleatorio

El algoritmo de Bosque Aleatorio fue implementado utilizando el lenguaje *Python* y la librería *Sklearn*. Los parámetros del algoritmo son el número de árboles del bosque y la profundidad de estos, mientras la arquitectura de los árboles es generada por el algoritmo a partir de la muestra de entrenamiento, maximizando la heterogeneidad entre los nodos "hijos" al dividir el nodo "padre", como fue descrito en la sección 1.2.1. El bosque propuesto comporta 600 árboles y no tiene restricción de profundidad.

Aunque un árbol de regresión pueda generar problemas de sobreajuste porque su arquitectura es basada en los datos de la muestra de aprendizaje, no es el caso del bosque aleatorio. En efecto, el problema de sobreajuste desaparece al generar un gran número de árboles y promediando el resultado de estos para obtener el resultado final de la regresión. Por la misma razón no hace falta restringir la profundidad de los árboles, lo que sí es necesario cuando se resuelve un problema con un solo árbol de regresión.

## 3.3. Resultados

En esta sección se comparan los resultados obtenidos con los tres algoritmos en términos de error absoluto porcentual medio y de tiempo de ejecución. De manera similar al protocolo empleado por Antipov y Pokryshevskaya (2012), se comparan los resultados obtenidos para la predicción del valor total de las viviendas y del valor del metro cuadrado de estas. Los resultados obtenidos están presentados en el Cuadro 3.3 y permiten extender la conclusión obtenida por Antipov y Pokryshevskaya (2012) a las redes neuronales y al modelo SVR, puesto que al predecir el precio del metro cuadrado de una vivienda en vez de su precio total se mejora el poder predictivo de todos los algoritmos utilizados. Por esta razón se utilizará esta última variable como variable de precio en el resto de este trabajo.

El error absoluto porcentual medio más bajo es obtenido por el bosque aleatorio, seguido por los algoritmos SVR no lineales, la red neuronal y el algoritmo SVR lineal. En términos de tiempo de ejecución, la red neuronal se posiciona como el algoritmo más lento, y el algoritmo SVR como el más rápido.

| Algoritmo | Tiempo de ejecución (s) | MAPE promedio ( %) | |
|---|---|---|---|
| | | Renta | UF/m2 |
| Red Neuronal | 1320 | 23.85 | 19.17 |
| Bosque Aleatorio | 120 | 9.73 | 9.65 |
| SVR Lineal | 73 | 32.04 | 22.94 |
| SVR Polinomial | 442 | 17.86 | 14.73 |
| SVR Rbf | 26 | 17.55 | 14.69 |

Cuadro 3.3: Resultados obtenidos con la red neuronal, el bosque aleatorio y las máquinas de vectores de soporte para la regresión.

Para los tres algoritmos, se representó espacialmente la distribución de los errores absolutos porcentuales medios en la comuna (Figura 3.2). Existen tres zonas en el norte de la comuna donde los algoritmos presentan dificultades para predecir un valor de la renta cercano al valor observado, que se ubican en los alrededores de las estaciones de metro Santa Julia (Línea 4a) y Macul (Línea 5), y en Villa Santa Teresa. Estas zonas corresponden a condominios sociales adquiridos con subsidios del MINVU, por lo que su valor no es representante del mercado inmobiliario de la ciudad. En la zona Sur, el bosque aleatorio logra predecir la renta de la mayoría de los bienes con un error absoluto porcentual medio menor a 8 %, mientras una gran parte de estos ven su renta predicha con más de 50 % de error por la red neuronal y el algoritmo SVR.

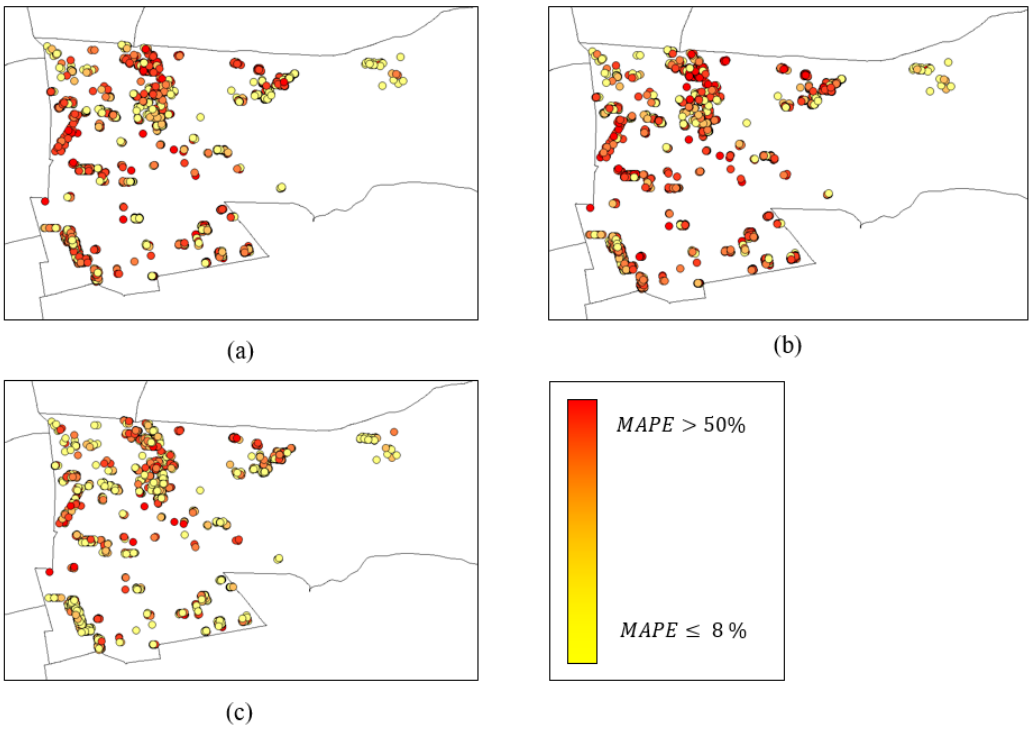


Figura 3.2: Distribución del error absoluto porcentual medio para el algoritmo SVR (a), la red neuronal (b) y el bosque aleatorio (c)

De manera general, la distribución espacial del error absoluto porcentual medio en la comuna es parecida para los tres algoritmos, por lo que se puede concluir que las dificultades encontradas para predecir las rentas podrían ser causadas por los datos utilizados, o por falta de variables explicativas adicionales, como fue el caso para las viviendas sociales.

La importancia de las variables se pueden obtener para la red neuronal y el bosque aleatorio mediante los métodos de atención y de impureza descritos en las secciones 1.1.4 y 1.2.2 respectivamente. En la Figura 3.3 se muestra la importancia calcula en porcentaje de las variables que representan el 90 % de participación en la predicción para el bosque aleatorio y la red neuronal.

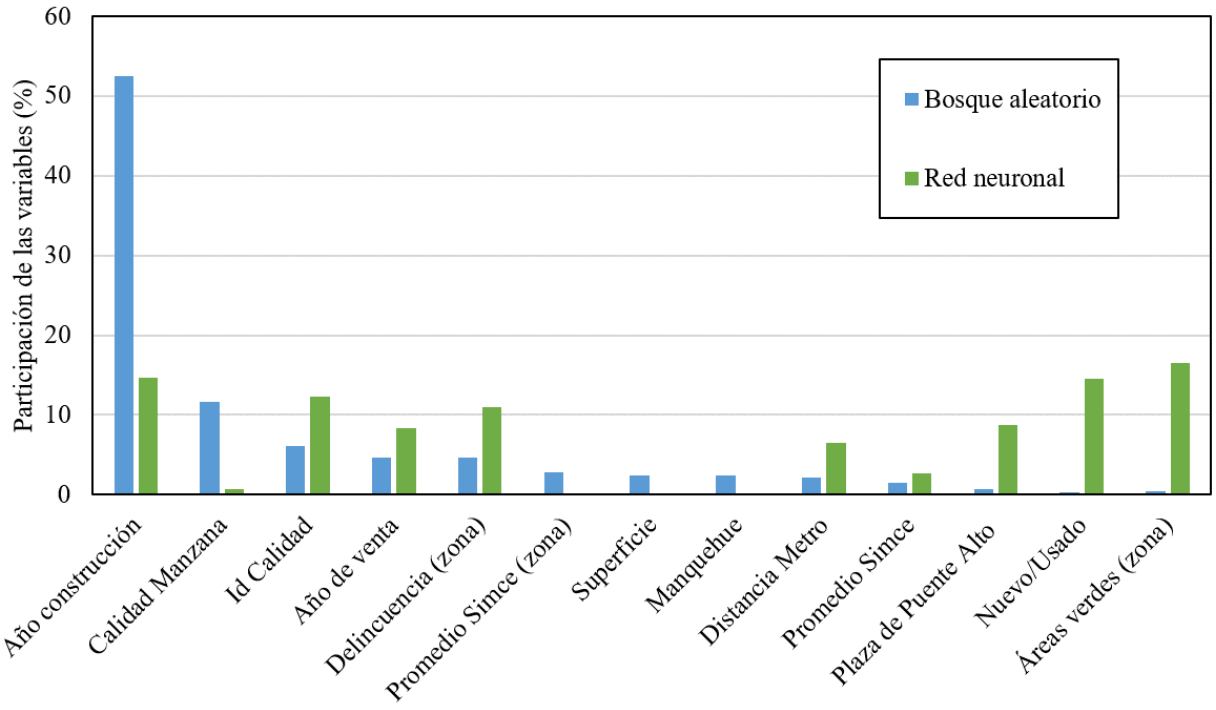


Figura 3.3: Distribución espacial del error absoluto porcentual medio obtenido con el algoritmo SVR (a), la red neuronal (b) y el bosque aleatorio (c)

Mientras el bosque aleatorio ocupa dos variables principales que son el año de construcción y la calidad de la manzana, el porcentaje de importancia de las variables es repartido de manera más equitativa en el caso de la red neuronal. Además, se calculó el error absoluto porcentual medio por año de venta de los departamentos, para los tres algoritmos (Cuadro 3.4). El número de venta no es constante a lo largo de los años y se observa una disminución de las ventas en 2009 producto de la crisis económica, y un fuerte aumento en el año 2016 debido al desarrollo urbano que conoció la comuna en este periodo, como se muestra en la Figura 3.4. Al igual que la cantidad de transacciones realizadas, el error absoluto porcentual medio no es constante a lo largo de los años en particular para la red neuronal y el algoritmo SVR. El poder predictivo de los dos últimos algoritmos disminuye a medida que disminuye el número de observaciones. Este fenómeno no ocurre con el bosque aleatorio. En el año 2009, cuando se registró la menor cantidad de transacciones, el error absoluto porcentual medio de la red neuronal y del algoritmo SVR superó los 43,3 % y 30 % mientras este número bajó a 14,2 % y 12,9 %, acercándose al error obtenido con el bosque aleatorio.

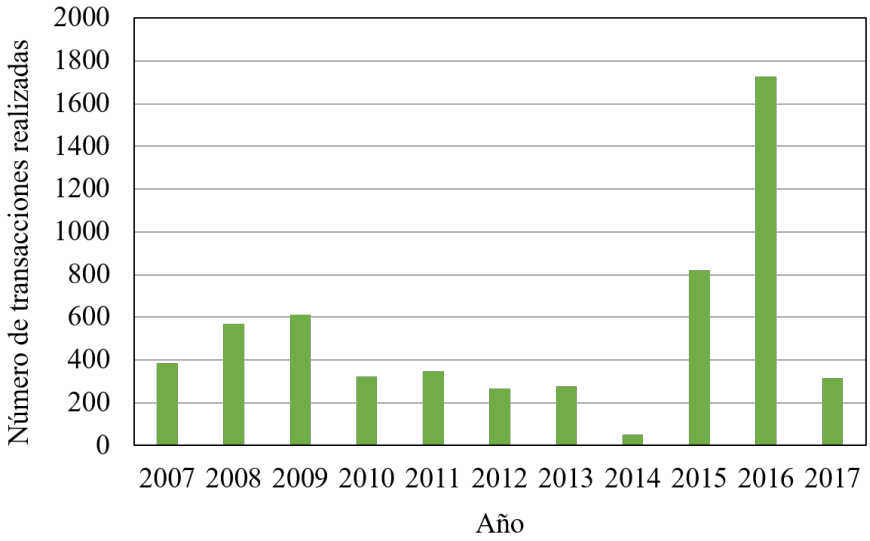


Figura 3.4: Número de transacciones realizadas por año en La Florida entre 2007 y 2018

| Año de venta | Nº de observaciones | Error porcentual pomedio (%) | | |
|---|---|---|---|---|
| | | Red Neuronal | SVR (Rbf) | Bosque Aleatorio |
| 2007 | 247 | 25,8 | 17,3 | 10,1 |
| 2008 | 541 | 28,8 | 19,9 | 11,4 |
| 2009 | 285 | 43,3 | 30,0 | 9,8 |
| 2010 | 620 | 18,0 | 14,4 | 9,6 |
| 2011 | 395 | 18,4 | 15,3 | 11,7 |
| 2012 | 611 | 15,6 | 13,6 | 9,6 |
| 2013 | 805 | 13,1 | 11,7 | 8,2 |
| 2014 | 488 | 17,5 | 14,2 | 9,1 |
| 2015 | 379 | 21,6 | 20,9 | 14,2 |
| 2016 | 2312 | 14,2 | 12,9 | 8,5 |
| 2017 | 1354 | 13,9 | 12,3 | 9,4 |
| 2018 | 489 | 16,4 | 15,8 | 11,4 |

Cuadro 3.4: Error absoluto porcentual medio obtenido por año de venta

# Capítulo 4

# Predicción de las rentas de los departamentos con el bosque aleatorio

En esta sección se analizan los resultados obtenidos con el bosque aleatorio en la predicción de rentas de departamento en la ciudad de Santiago entre los años 2007 y 2018, después de comprobar distintas especificaciones de la variable temporal con el objetivo de considerar la evolución no lineal de los precios del suelo a lo largo de los años. Dado el carácter fuertemente heterogéneo de la ciudad, se construyen distintas muestras de observaciones que representan sectores geográficos y sectores de ingresos de los compradores.

## 4.1. Especificación de la variable temporal

Se busca mejorar la especificación de la variable temporal del año de venta de las viviendas para considerar la evolución no lineal del precio del suelo a lo largo de los años y su heterogeneidad dentro de la ciudad. La Figura 4.1 muestra la evolución temporal de las rentas en las comunas de Las Condes, San Joaquín, Providencia y La Florida. El precio del suelo en estas comunas no conoció un aumento lineal en los diez últimos años pero dobló su valor, por lo que se construyó una variable temporal específica a cada comuna, descrita en el siguiente párrafo.

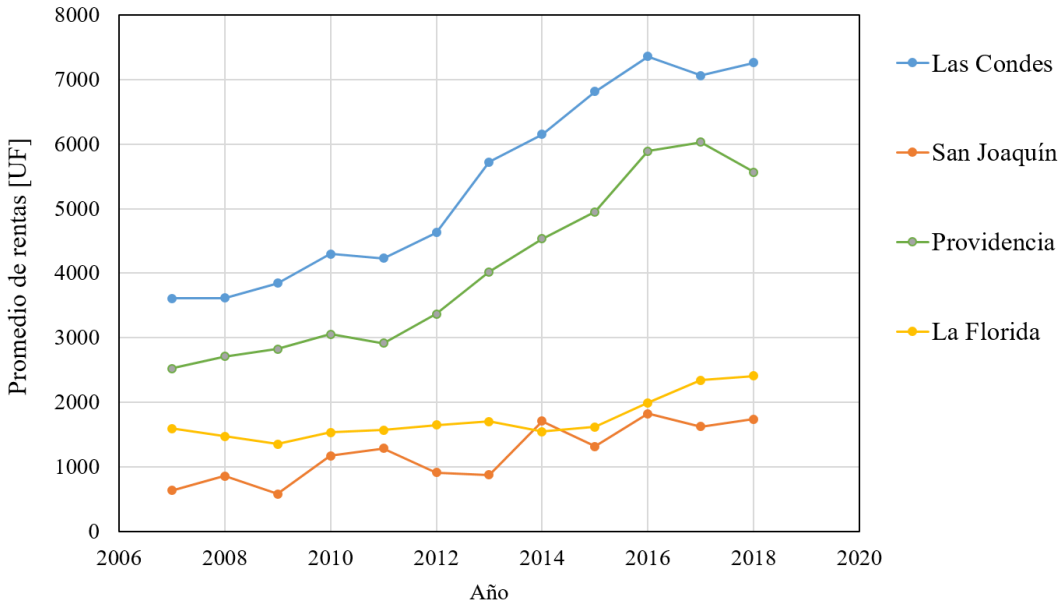


Figura 4.1: Evolución del promedio de las rentas en Las Condes, Providencia, San Joaquín y La Florida

Se comparan los resultados obtenidos considerando cuatro especificaciones de la variable temporal. La primera variante considera como variable temporal el año de transacción de la vivienda. La segunda considera las variaciones existentes entre las comunas, cuyos precios del suelo siguen diferentes dinámicas según su localización, el uso del suelo y el ingreso de sus habitantes. La variable temporal correspondiente es $p_{promedio}$, que a cada transacción asigna el promedio de las ventas efectuadas el mismo año en la comuna donde ha sido efectuada. La tercera y la cuarta variante consideran una nueva variable de precio $Y$ definida como $Y = \frac{p}{p_{promedio}}$ donde $p_{promedio}$ representa el promedio de las ventas en la comuna donde se efectuó la transacción o el promedio de las ventas en la ciudad completa. La nueva variable obtenida reemplaza las variables precio y año de transacción en el modelo y mide la variación entre el valor del bien vendido y el valor promedio de las transacciones efectuadas en la misma comuna o en toda la ciudad.

### 4.1.1. Resultados

En el Cuadro 4.1 se muestran los resultados obtenidos por el bosque aleatorio a parir de una muestra de 82.242 transacciones efectuadas en la ciudad de Santiago, para las cuatro variantes mencionadas.

El error absoluto porcentual medio obtenido para cada variante muestra que el algoritmo pierde en poder predictivo al fusionar la variable temporal y la variable de precio en una sola variable como se hizo en las variantes 3 y 4. Por otro lado, el error absoluto porcentual medio promedio obtenido para las variantes 1 y 2 es el mismo, por lo que se puede utilizar el año de transacción o el promedio de las ventas comunales para predecir la renta en este caso. En el resto de este trabajo se utilizará la variable año de transacción como variable temporal y la variable precio del suelo en $UF/m^2$ como variable de precio.

| Variante | Variable temporal | Error porcentual promedio (%) |
|---|---|---|
| 1 | Año de transacción | 10,99 |
| 2 | $p_{promedio}$ (comuna) | 10,99 |
| 3 | $\frac{p}{p_{promedio}}$ (comuna) | 16,93 |
| 4 | $\frac{p}{p_{promedio}}$ (ciudad) | 13,08 |

Cuadro 4.1: Error porcentual promedio obtenido por las cuatro variantes de variable temporal

### Participación de las variables

A continuación se muestra la importancia de las variables que representan 90 % de la participación total.

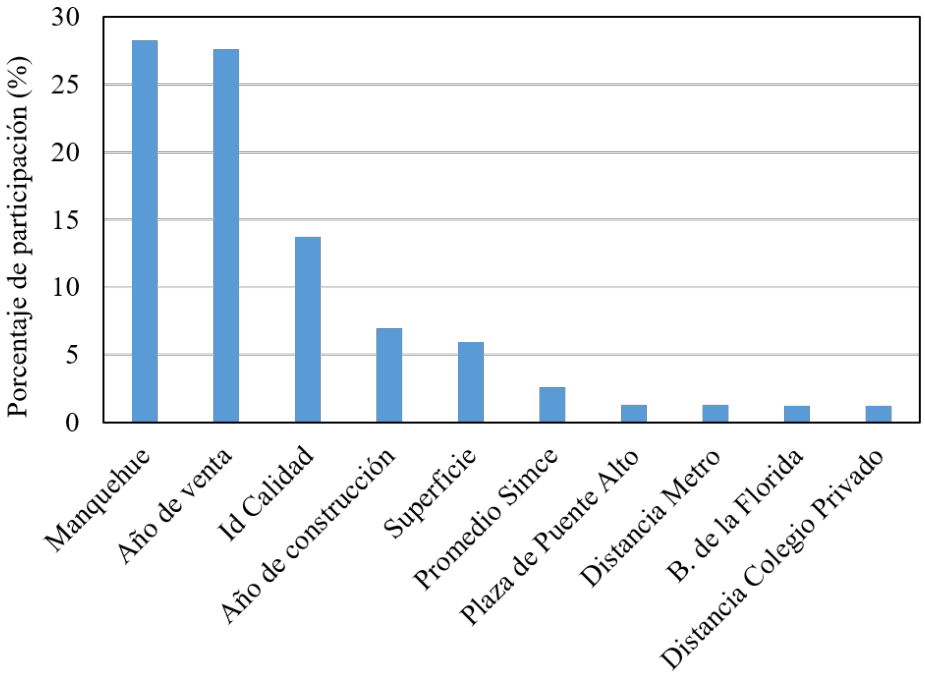


Figura 4.2: Importancia de las variables en la ciudad completa (90 % de participación total)

Las importancias de las variables obtenidas muestran que la distancia a la estación de metro Manquehue y el año de venta son los dos atributos que predominan en la atribución de precios de las viviendas. En efecto, en el sector oriente de la ciudad se concentran los departamentos de precios más altos, mientras el valor del suelo decae a hacía el poniente y el sur de la ciudad como se muestra el la Figura 2.3, sección 2.3. De está manera, el atributo de localización cobra más importancia que los atributos propios de las viviendas tales como el índice de calidad y el año de construcción. A pesar de predecir el valor del metro cuadrado de los departamentos, la variable superficie cobra importancia en la predicción por la non-linealidad que existe entre el precio de una vivienda y su superficie.

La distribución geográfica del error absoluto porcentual medio (Figura 4.3) no permite concluir sobre la existencia de un patrón espacial para la predicción de las rentas. De la misma manera, las viviendas para las cuales el error absoluto porcentual medio es el más alto no presentan atributos distintos al resto de ellas, por lo que el error absoluto porcentual medio es equitativamente repartido espacialmente y dentro de los distintos tipos de departamentos.

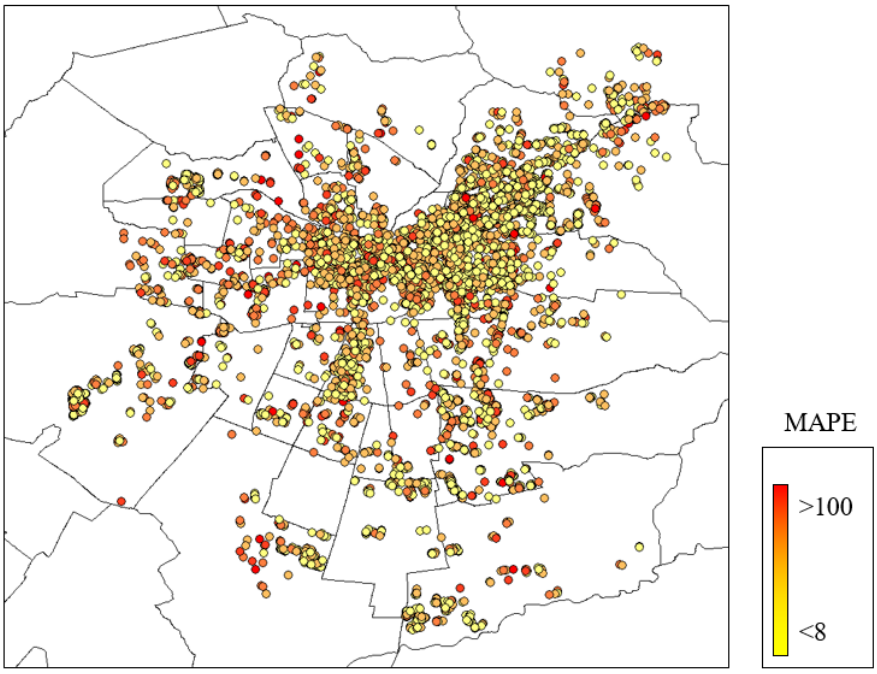


Figura 4.3: Representación espacial del error absoluto porcentual medio en la ciudad completa

## 4.2. Predicción de la renta por región geográfica

En esta sección se delimita la ciudad de Santiago en seis zonas geográficas que representan zonas de mercado laboral local, definidas por Fuentes, Mac-Clure, Moya y Olivos (2017) a partir de los resultados de la Encuesta Origen y Destino de Viajes (SECTRA, 2012) para los traslados hogar - trabajo. Las zonas definidas se denominan zona norte (Quilicura, Conchalí, Huechuraba, Independencia, Renca, Recoleta), zona sur (Peñalolén, La Florida,

Macul, La Granja, San Joaquín, San Miguel, Pedro Aguirre Cerda), zona surponiente (San Bernardo, El Bosque, La Cisterna, San Ramón), zona suroriente (La Pintana, Puente Alto), zona centrooriente (Santiago Centro, Providencia, Ñuñoa, Vitacura, Las Condes, La Reina, Lo Barnechea) y zona poniente (Cerro Navia, Quinta Normal, Lo Prado, Lo Espejo, Estación Central, Pudahuel, Cerrillos, Maipú) (Figura 4.4).

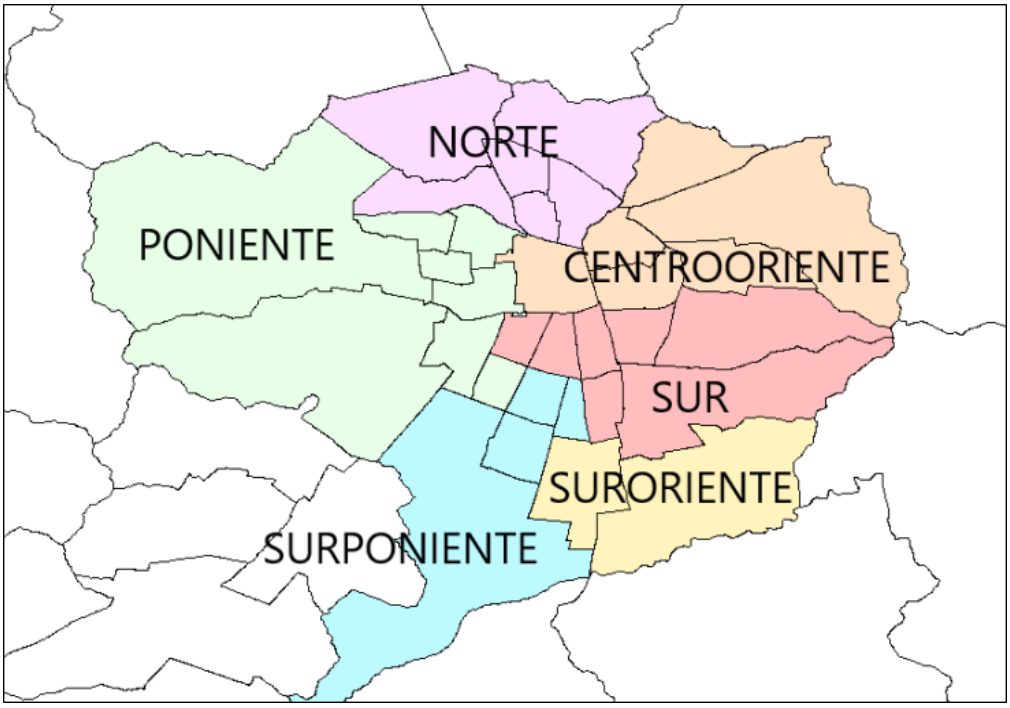


Figura 4.4: Zonas de mercado laboral definidas por Fuentes, Mac-Clure, Moya y Olivos (2017)

Según el mismo estudio, la zona centrooriente está habitada principalmente por un estrato medio-alto con educación superior universitario o técnica y trabajando en los servicios de alta categoría, pero también existe desigualdad social entre sus habitantes. Los habitantes de la zona poniente y sur pertenecen al estrato medio con educación media en la primera y media-superior en la segunda. En las zonas surponiente, norte y suroriente residen en mayoria trabajadores manuales, con ingresos bajos. Las diferencias entre las distintas zonas se puede apreciar promediando atributos de la muestra de transacciones por zonas geográficas, como se muestra en el Cuadro 4.2. La zona centrooriente se caracteriza por la alta presencia de áreas verdes, altos resultados en la prueba Simce, departamentos con superficies altas y cercanía a los centros de salud privados. En las zonas surponiente, suroriente y poniente se ubican departamentos más pequeños, más antiguos y de menor calidad en promedio que en el resto de la ciudad, y carecen de centros de salud privados. En las zonas sur y norte se úbican departamentos de superficie y de calidad media.

| Zona | Centrooriente | Norte | Surponiente | Suroriente | Sur | Poniente |
|---|---|---|---|---|---|---|
| Año de construcción | 2002,0 | 2004,6 | 1997,4 | 1996,8 | 2002,2 | 1996,0 |
| Id Calidad | 2,39 | 3,29 | 3,71 | 4,01 | 3,29 | 3,74 |
| Superficie ($m^2$) | 85,3 | 56,2 | 47,4 | 47,0 | 55,0 | 48,7 |
| Áreas verdes | 5,5 | 3,7 | 2,2 | 2,1 | 2,9 | 3,5 |
| Delincuencia | 6,0 | 6,5 | 7,0 | 6,1 | 5,5 | 3,5 |
| Distancia Metro (m) | 2293 | 2905 | 2681 | 3777 | 1298 | 2523 |
| Simce | 276 | 263 | 265 | 251 | 261 | 262 |
| Distancia salud privada (m) | 1684 | 4904 | 9242 | 9671 | 3870 | 9451 |

Cuadro 4.2: Promedio de atributos propios y externos de los departamentos vendidos entre 2007 y 2018 según zona geográfica

En términos de evolución temporal, el valor de la renta aumenta de manera desigual en las distintas zonas geográficas, en particular en la zona centrooriente, como se puede ver en la Figura 4.5. En esta zona las rentas de los departamentos aumentaron a mayor tasa que en las otras comunas, por lo que el valor un bien es más sensible a su año de venta que en las otras zonas. Por lo contrario, las rentas se mantuvieron casi constantes en las zonas suroriente y norte.

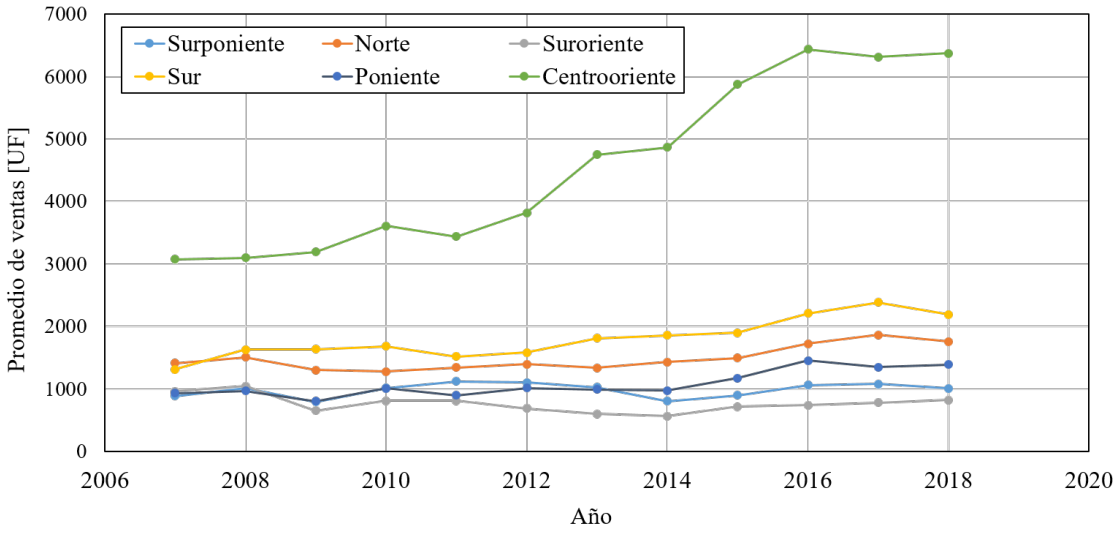


Figura 4.5: Evolución temporal del promedio de las rentas según zona geográfica

### 4.2.1. Resultados

En el Cuadro 4.3 se muestra el error porcentual promedio obtenido según zona geográfica. El error relativa porcentual varia entre 8,49 % y 11,08 %.

| Zona | Número de observaciones | error absoluto porcentual medio ( %) |
|---|---|---|
| Norte | 12351 | 11,08 |
| Sur | 24598 | 9,91 |
| Centrooriente | 82366 | 11,00 |
| Poniente | 15325 | 9,67 |
| Surponiente | 4918 | 8,79 |
| Suroriente | 2713 | 8,49 |

Cuadro 4.3: Error absoluto porcentual medio obtenido según zona geográfica

### 4.2.2. Participación de las variables

#### Atributos propios de las viviendas

En la Figura 4.6 se grafica la importancia de los atributos propios de las viviendas según zona geográfica. Los atributos de las viviendas que afectan el precio del suelo son principalmente el índice de calidad y el año de construcción, en particular en las zonas sur y poniente. Cabe destacar que el año de construcción es una variable considerada en el cálculo del índice de calidad realizado por el Servicio de Impuestos Internos, por lo que estas dos variables están correlacionadas. En efecto, los edificios con poca antigüedad suelen tener mejores infraestructuras tales como ascensores, más dispositivos de seguridad y mejor diseño estructural que los edificios antiguos. La zona sur, donde el índice de calidad cobra más importancia que en las otras zonas, es la zona donde la calidad de las viviendas es la más heterogénea.

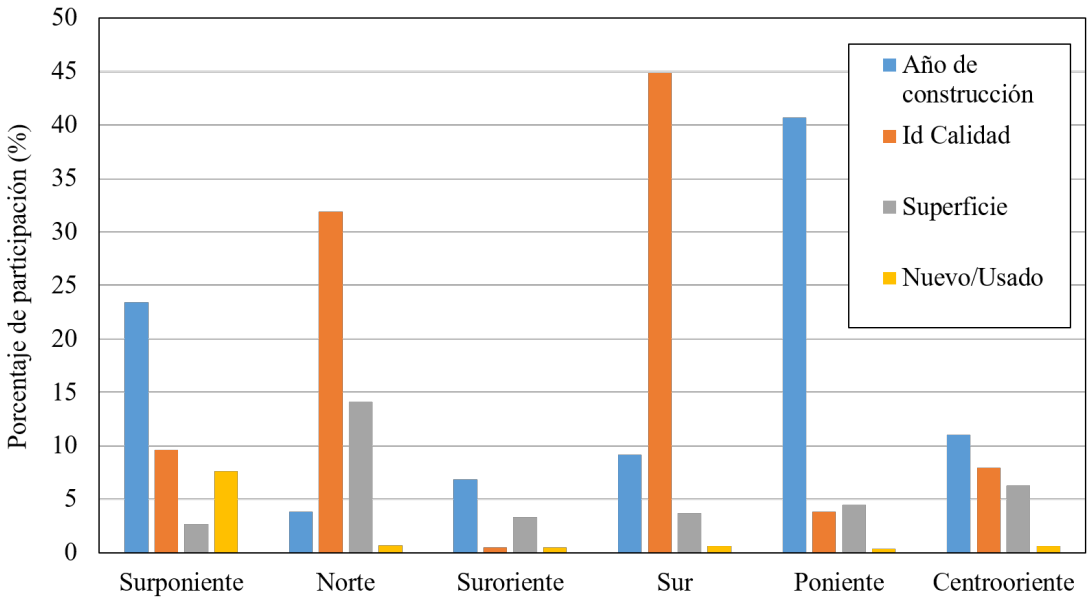


Figura 4.6: Importancia de los atributos propios de los departamentos según zona geográfica

### Atributos de entornos

En la Figura 4.7 se grafica el porcentaje de participación de los atributos de entorno que suman 90 % de la participación total según zona geográfica.

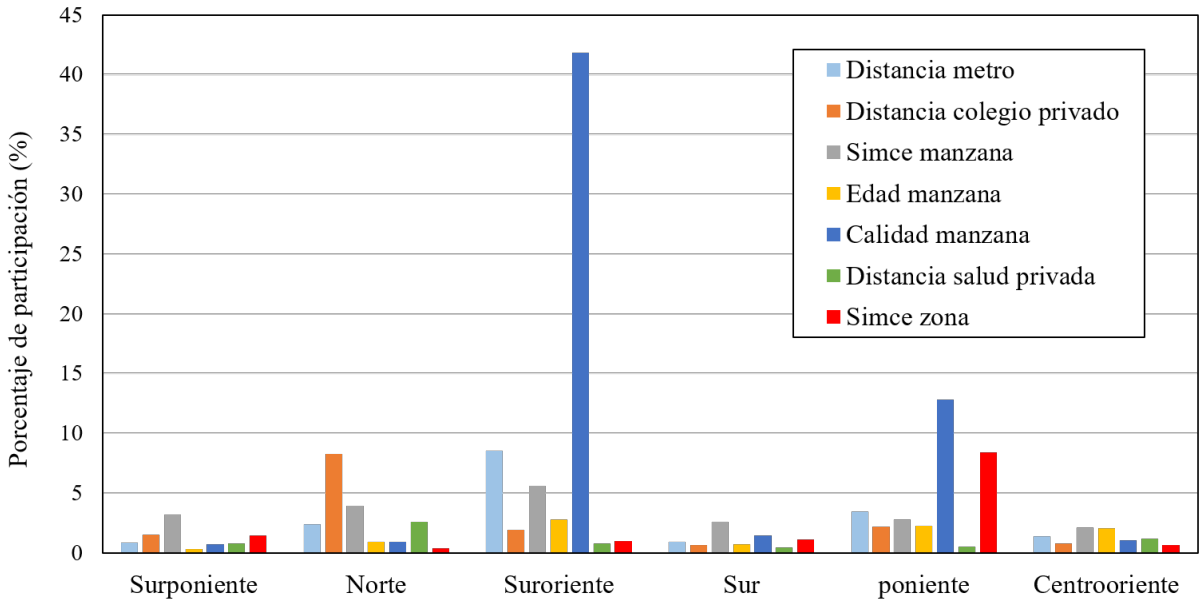


Figura 4.7: Importancia de los atributos de entorno de los departamentos según zona geográfica

En la zona suroriente predomina la variable índice de calidad de la manzana, cuyo valor es fuertemente correlacionado con el índice de calidad de los departamentos, por lo que el índice de calidad de los departamentos de la zona suroriente también es una variable explicativa de la renta de los departamentos. El mismo fenómeno se puede observar en la zona poniente. En esta misma zona cobra importancia el promedio de los resultados Simce por zona homogénea, y alcanza 8 % de participación. Un análisis espacial de la relación entre precio del suelo y promedio de los resultados simce (Figura 4.8) permite concluir que las zonas con alto promedio Simce coinciden con zonas recientemente construidas donde se ubican departamentos de mejor calidad.

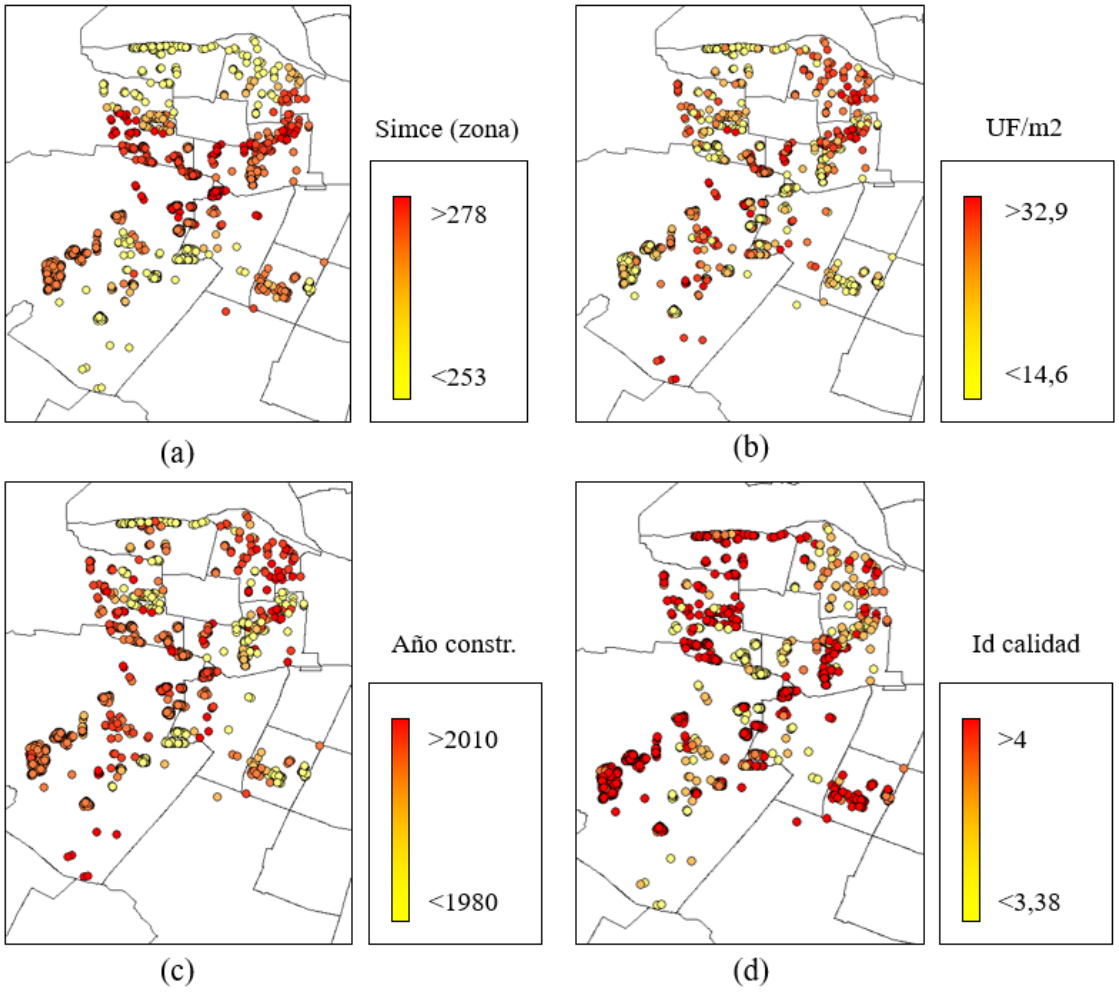


Figura 4.8: Representación espacial de las variables promedio de los resultado Simce por zona TocToc.com (a), precio del suelo (b), año de construcción (c) e índice de calidad (d) en la zona poniente.

En la zona norte, la distancia al colegio más cercano alcanza el porcentaje de participación más elevado dentro de los atributos de entorno. En la Figura 4.9 se representan espacialmente las variables distancia al colegio privado más cercano, precio del suelo, año de construcción y índice de calidad en la zona norte de la ciudad.

Las comunas de Independencia, Recoleta Sur y Huechueraba poniente son las zonas donde el precio del suelo es lo más alto, y también donde se ubican los colegios privados San Francisco Javier de Huechuraba y Academia de las Humanidades. La representación espacial del año de construcción y del índice de calidad muestran que las zonas cercanas a estos colegios son zonas construidas después de los años 2000 con un buen índice de calidad, por lo que las variables promedio Simce e índice de calidad podrían estar correlacionadas en esta zona. Entonces, la importancia de la variable promedio Simce también traduce la importancia del índice de calidad. En la sección 4.5 se presentan los distintos riesgos de sesgo en el cálculo de la importancia de las variables, en particular en el caso de las variables correlacionadas entre ellas. Los edificios úbicados en Renca poniente y Huechuraba oriente tienen un valor del suelo bajo a pesar de ser recién construidos, lo que contradice el hecho de que en general, los bienes recién construidos tiene un precio más alto que los bienes antiguos por ser de mejor calidad. Sin embargo, estos edificios pertenecen a la categoría de las viviendas sociales por lo que, además de no representar el mercado, no tienen un buen índice de calidad.

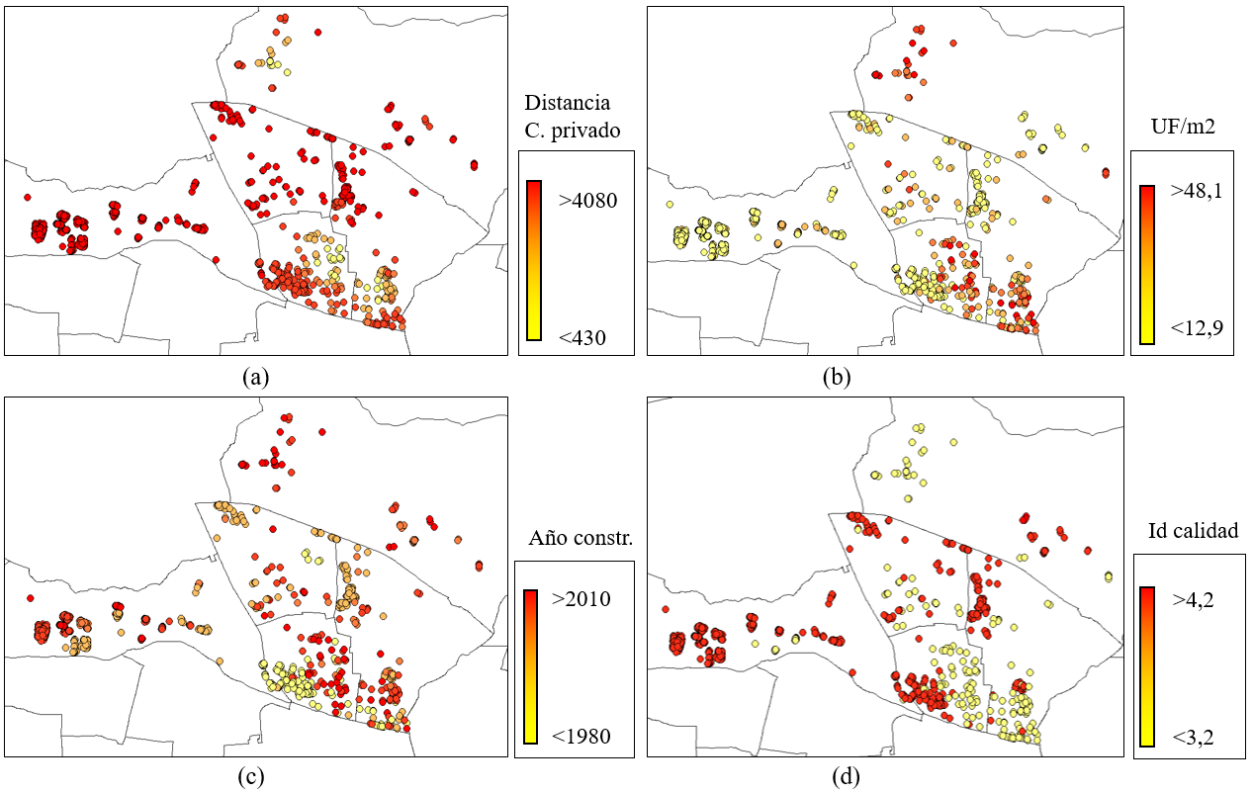


Figura 4.9: Representación espacial en la zona norte de las variables distancia al colegio privado más cercano (a), precio del suelo (b), año de construcción (c) e índice de calidad (d).

En la zona suroriente, donde pasa la Línea 4a del metro en la comuna de Puente Alto, la distancia a la estación de metro más cercana alcanza los 8 % de participación. La construcción de la línea 4a entorno al eje Concha y Toro tuvo por consecuencia el desarrollo de la actividad comercial y la llegada de servicios gubernamentales, convirtiendo esta zona en un subcentro urbano de la ciudad. En este sector, la renta de los departamentos es más elevada que en el resto de la comuna.

### Año de venta

El año de venta cobra más importancia en la zona centrooriente que en el resto de la ciudad (Figura 4.10), conformemente a la evolución temporal del precio del suelo comentada en la sección 4.2. En las zonas suroriente, surponiente y poniente, donde el precio del suelo se mantiene casi estable en el tiempo, la importancia del año de venta no supera los 12 % de participación. Las zonas norte y sur conocieron un aumento más importante del valor del suelo y las rentas de los departamentos son más sensible al año de venta en estas zonas.

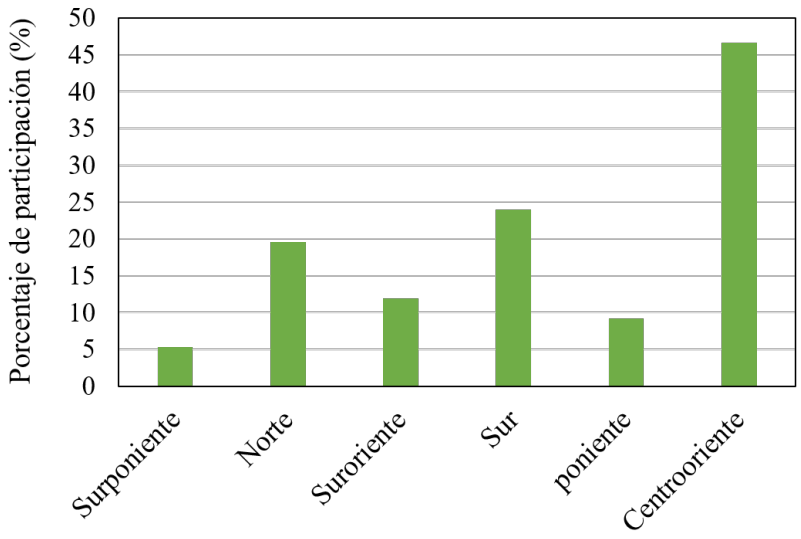


Figura 4.10: Importancia de la variable año de venta de los departamentos según zona geográfica.

### Distancia a los centros suburbanos

A continuación se muestra la importancia de las variables distancia a los nueve centros suburbanos según zona geográfica (Figura 4.11).

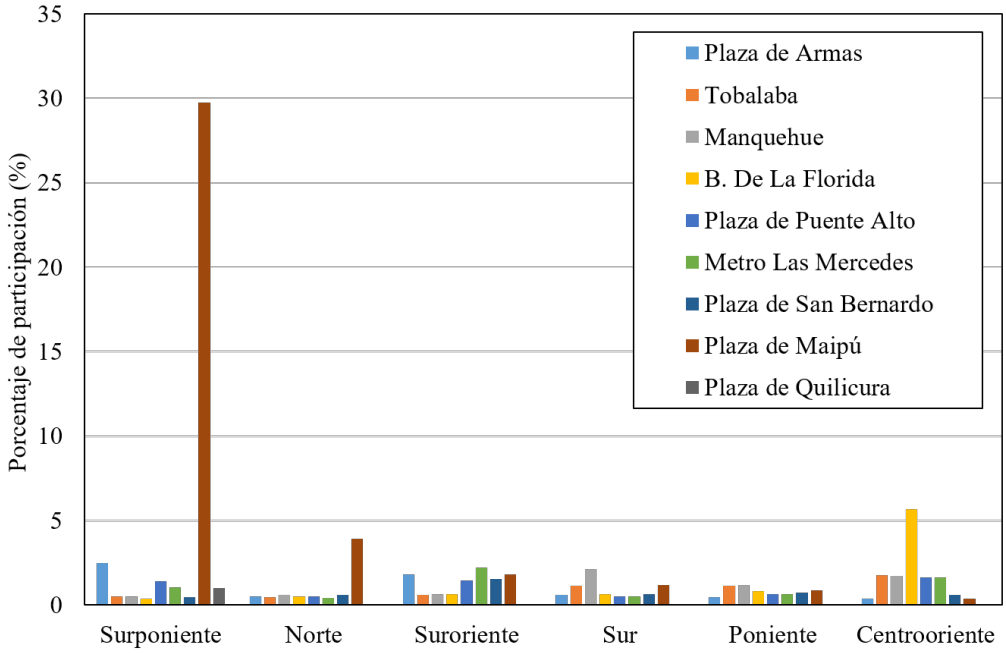


Figura 4.11: Importancia de las variables distancia a los centros suburbanos según zona geográfica

La variable distancia a los centros suburbanos afecta principalmente a las zonas surponiente, norte y centrooriente. En la zona norte, este fenómeno se puede explicar por la presencia de edificios recientes con bajo índice de calidad en el sector sur de la zona, en las comunas de Recoleta e Independencia (Figura 4.9). En la zona centrooriente, la participación de la variable distancia a Bellavista de La Florida refleja la disminución del precio del suelo al alejarse del centro de negocios representado por la estación de metro Manquehue. En efecto, como se puede ver en la Figura 4.12, el valor del suelo es más bajo en las comunas de Ñuñoa y la Reina que en las comunas de Las Condes, Vitacura y Lo Barnechea.

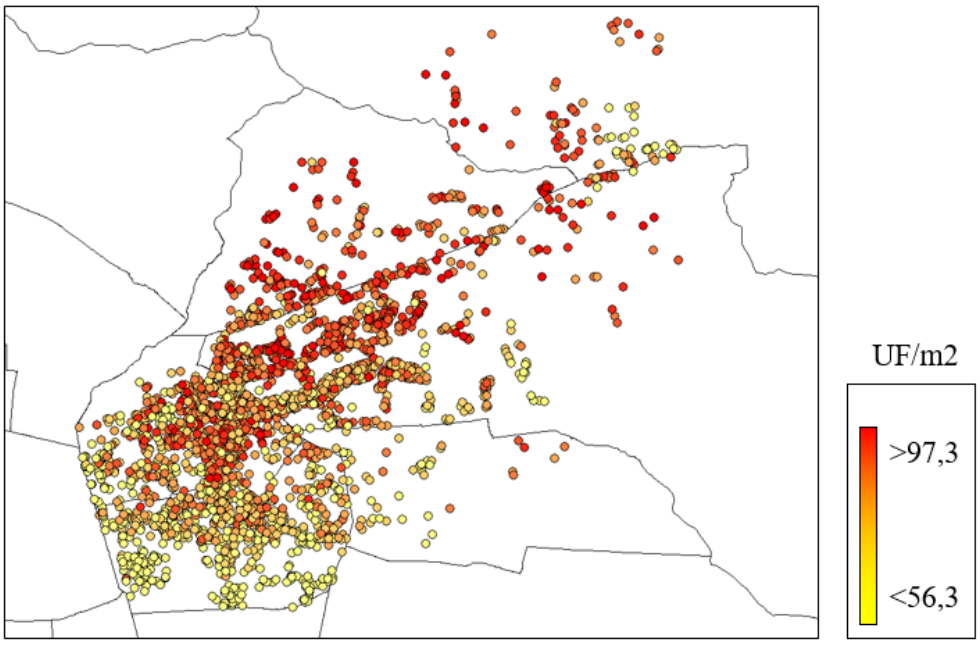


Figura 4.12: Representación espacial del valor del suelo en la zona centrooriente

En la zona surponiente, la importancia de la variable distancia a la plaza de Maipú alcanza los 30 % de participación y traduce una desigualdad entre los distintos sectores de esta zona. como lo muestra la Figura 4.13, existe una separación este-oeste en términos de calidad de la construcción. La comuna de San Ramón y el este de San Bernardo se caracterizan por edificios antiguos con de baja calidad, mientras el índice de calidad de los departamentos mejora en la comuna de La Cisterna y al oeste de San Bernardo. Por otro lado, el sector norte de la zona cuenta con mejor accesibilidad al servicio de metro y la comuna de La Cisterna cuenta con edificios más nuevos que en el resto de la zona.

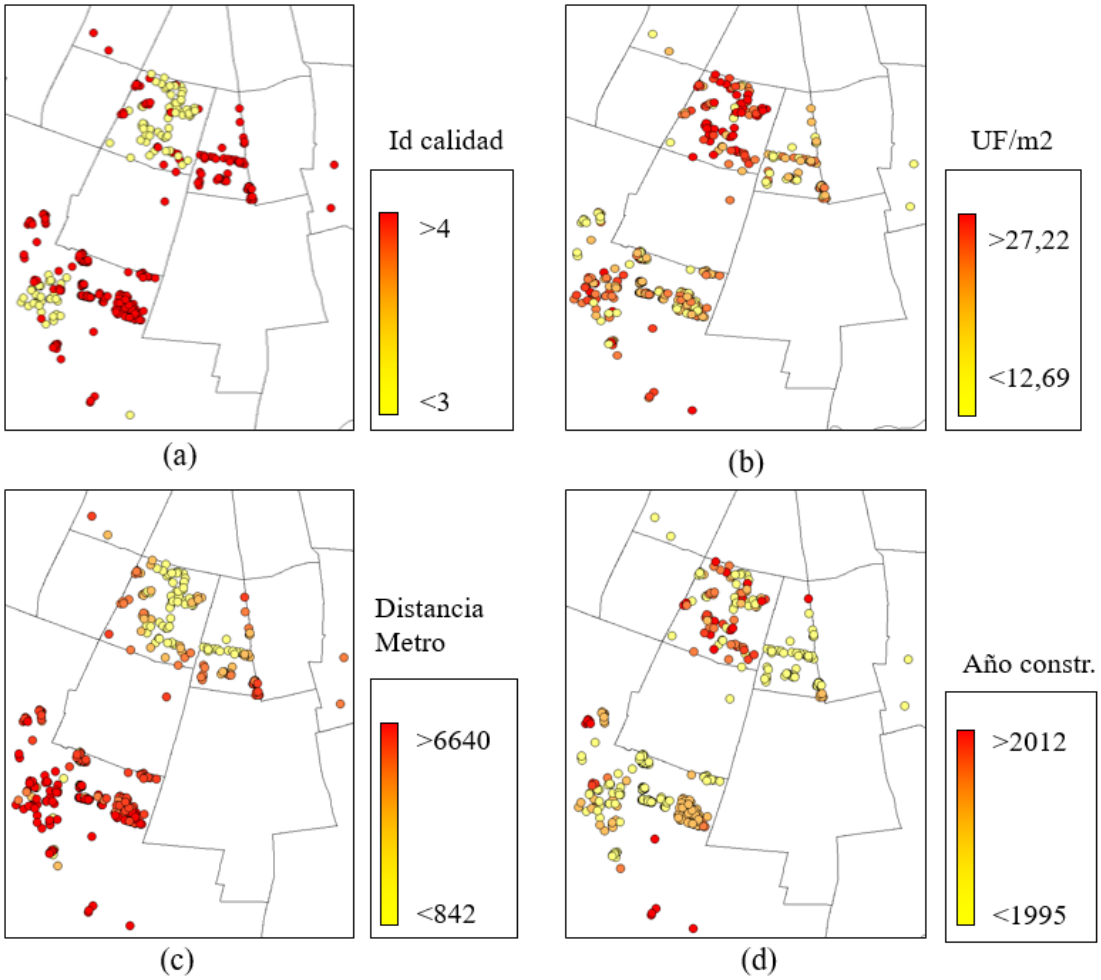


Figura 4.13: Representación espacial en la zona surponiente de las variables índice de calidad (a), precio del suelo (b), distancia a la estación de metro más cercana (c) y año de construcción (d).

## 4.3. Predicción de la renta por sector de ingresos

En esta sección se analizan los resultados obtenidos por el bosque aleatorio a partir de dos muestras que representan comunas de ingresos altos y comunas de ingresos bajos, elaborados a partir de las tazas de pobreza por ingreso por comunas (Encuesta Casen, 2013). La primera muestra representa comunas de ingresos altos y contiene 82.377 transacciones efectuadas entre los años 2007 y 2018 en las comunas de Providencia, Las Condes, Ñuñoa, Vitacura, Lo Barnechea y La Reina. La segunda muestra representa a comunas de ingresos bajos y contiene 18135 transacciones efectuadas en las comunas de Cerrillos, Cerro Navia, Conchalí, Independencia, La Cisterna, La Granja, La Pintana, Lo Espejo, Lo Prado, Pedro Aguirre Cerda, Renca, San Bernardo, San Joaquín y San Ramón.

En el Cuadro 4.4 se muestran el promedio y la desviación estándar de algunos atributos en las zonas de altos y bajos ingresos. En el grupo de comunas de ingresos bajos se ubican

departamentos de baja calidad y baja superficie, que cuentan en general con mal acceso a los centros de salud y a los colegios privados. El promedio Simce y el índice de acceso a las áreas verdes son menores en estas comunas que en las comunas de altos ingresos, y la tasa de delincuencia es mayor. En el grupo de altos ingresos se ubican departamentos de mayor calidad y con superficies más altas en promedio. El acceso a las áreas verdes representa el doble del acceso del primer grupo, y hay una alta presencia de colegios privados y de centros de salud privados.

| | Ingresos bajos | | Ingresos altos | |
|---|---|---|---|---|
| Atributo | Promedio | Desv. estándar | Promedio | Desv. estándar |
| Año de construcción | 2003 | 21 | 2001 | 16 |
| Id Calidad | 3,47 | 0,53 | 2,36 | 0,58 |
| Superficie ($m^2$) | 46,7 | 10,7 | 82,5 | 41,0 |
| IAUAV ($m^2/hab$) | 2,5 | 3,6 | 5,2 | 5,2 |
| Decil delincuencia | 5,3 | 2,4 | 6,1 | 2,4 |
| Distancia Metro ($m$) | 2414 | 1954 | 1615 | 1549 |
| Distancia colegio privado ($m$) | 2378 | 1690 | 588 | 364 |
| Promedio Simce | 263 | 14 | 276 | 16 |
| Distancia salud privada | 6093 | 4378 | 1351 | 702 |

Cuadro 4.4: Promedio y desviación estándar de nueve atributos propios y externos de los departamentos según sector de ingresos

En términos de evolución temporal, el precio de los departamentos ubicados en las comunas de alto ingreso es más sensible al año de venta que en las comunas de bajo ingreso, como se muestra en la Figura 4.14.

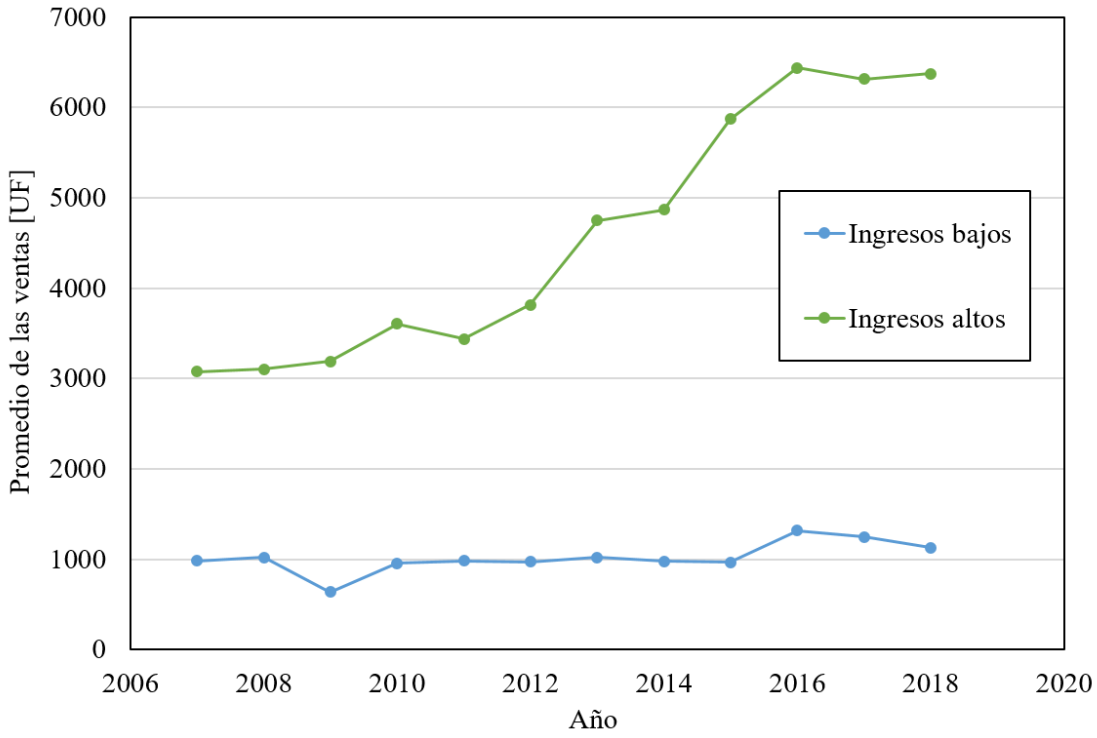


Figura 4.14: Evolución temporal de las rentas por grupo de comunas

### 4.3.1. Resultados

En el Cuadro 4.5 se muestra el error absoluto porcentual medio obtenido por grupos de ingresos, y en la Figura 4.15 se muestra la importancia de las variables por grupo de ingresos. De acuerdo a la evolución temporal de las rentas, la variable año de transacción cobra más importancia en las comunas de ingreso alto que en las comunas de ingresos bajos, donde predomina la variable índice de calidad.

| Grupo de comunas | Número de observaciones | Error absoluto porcentual medio (%) |
|---|---|---|
| Ingresos Altos | 82366 | 11 |
| Ingresos Bajos | 18134 | 9,67 |

Cuadro 4.5: Error absoluto porcentual medio según grupo de comunas

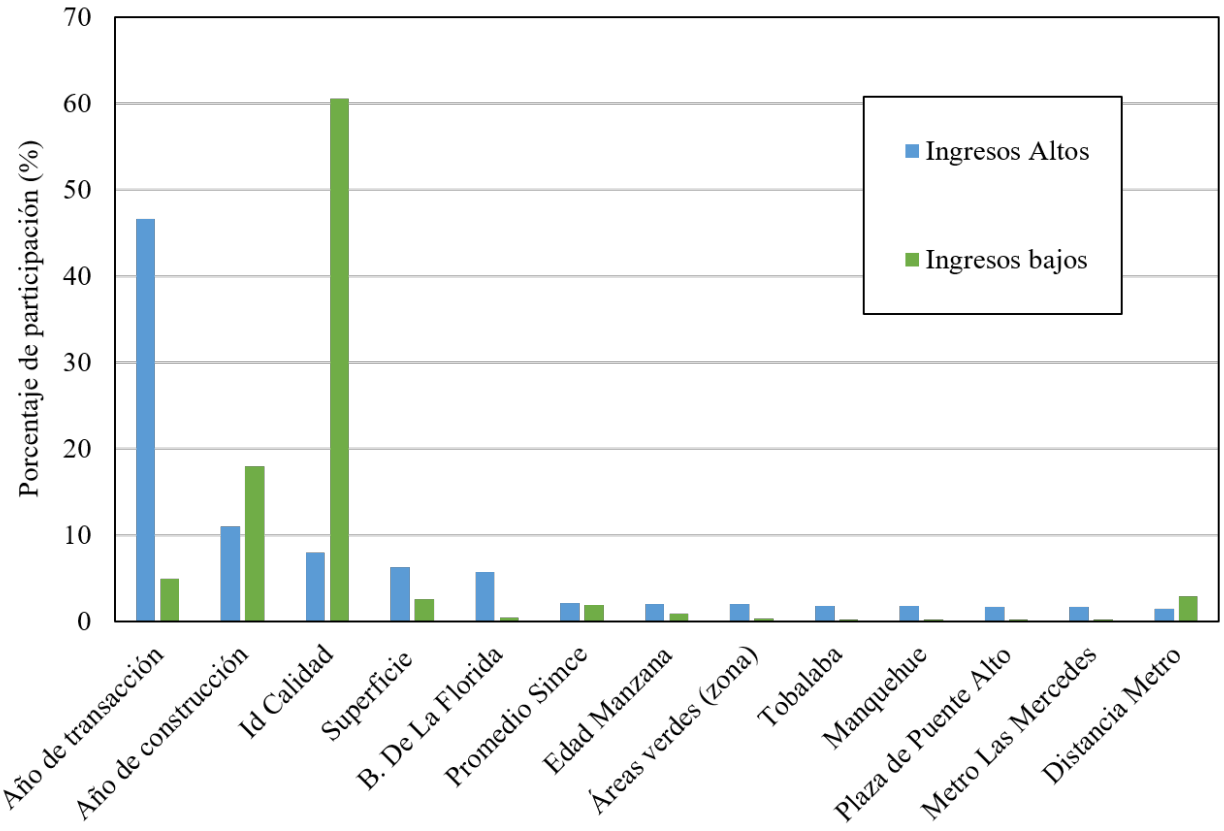


Figura 4.15: Importancia de las variable según grupo de ingresos

## 4.4. Predicción de la renta agregando una variable de ingreso por comuna

Le resultados anteriores muestran que las importancias de las variables son distintas si se predicen los precios de los departamentos en las comunas de ingresos altos o de ingresos bajos. En las comunas de altos ingresos predomina la variable año de transacción porque el precio del suelo aumenta a una tasa mayor que en las comunas de bajos ingresos, donde predominan las variables año de construcción e índice de calidad. Esto señala que el ingreso de los hogares de las comunas debe poder ser utilizado por el bosque aleatorio para predecir los precios de los bienes.

En esta sección se agregan a las variables predictoras el ingreso monetario promedio de los hogares según comuna (resultado de la encuesta CASEN 2013), cuya representación espacial (Figura 4.16) muestra que las comunas donde se ubican los hogares con más recursos se ubican en la zona oriente y en la comuna de Santiago. Los hogares con menos recursos están ubicados en las periferias sur, norte y poniente. En el Cuadro 4.6 se muestran los ingresos promedios de los hogares en las comunas consideradas en este trabajo.

| Comuna | Ingreso (CLP) | Comuna | Ingreso (CLP) |
|---|---|---|---|
| Cerrillos | 716.905 | Maipú | 857.958 |
| Cerro Navia | 531.362 | Ñuñoa | 2.052981 |
| Conchalí | 540.747 | P. Aguirre Cerda | 580.6613 |
| Estación Central | 650.501 | Peñalolén | 645.237 |
| Huechuraba | 607.102 | Providencia | 1.976.061 |
| Independencia | 502.782 | Pudahuel | 538.390 |
| La Cisterna | 918.453 | Puente Alto | 614.688 |
| La Florida | 849.390 | Quinta Normal | 698.624 |
| La Granja | 492.649 | Recoleta | 468.284 |
| La Pintana | 563.001 | Renca | 639.504 |
| La Reina | 1.930.327 | San Bernardo | 561.577 |
| Las Condes | 2.275.015 | San Joaquin | 781.546 |
| Lo Barnechea | 1.582.340 | San Miguel | 1.045.134 |
| Lo Espejo | 660.107 | San Ramón | 591.425 |
| Lo Prado | 595.267 | Santiago Centro | 1.279.990 |
| Macul | 978.705 | Vitacura | 2.649.750 |

Cuadro 4.6: Ingreso monetario promedio por hogar según comuna. Fuente: encuesta CASEN 2013

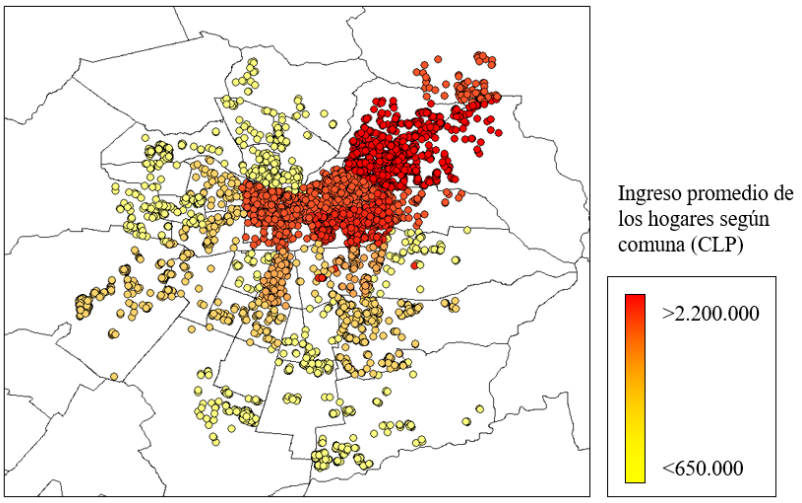


Figura 4.16: Representación espacial del ingreso promedio por hogar según comunas. Fuente: encuesta CASEN 2013

### 4.4.1. Resultados

El error absoluto porcentual medio obtenido es de 10,99 %, por lo que no mejora la predicción del bosque aleatorio. Sin embargo, se produce un cambio en la importancia de las variables, como se muestra en la Figura 4.17. En la ausencia de la variable ingreso, la variable más importante es la distancia a la estación Manquehue, que representa el aumento del precio del suelo al acercase de la zona oriente. Cuando se agrega la variable de ingreso, esta se convierte en la variable más importante, mientras la importancia de la distancia a la estación Manquehue se acerca a cero. El rol importante de la variable ingreso era esperable dado la diferencia entre la importancia de las variables en los barrios de bajos y altos ingresos (ver sección 4.3.1) y la presencia de las comunas de ingresos más altos a proximidad de la estación Manquehue. Sin embargo, la preferencia del árbol aleatorio por la variable de ingreso podría significar que esta variable describe mejor el precio del suelo en la ciudad que la distancia a la estación Manquehue.

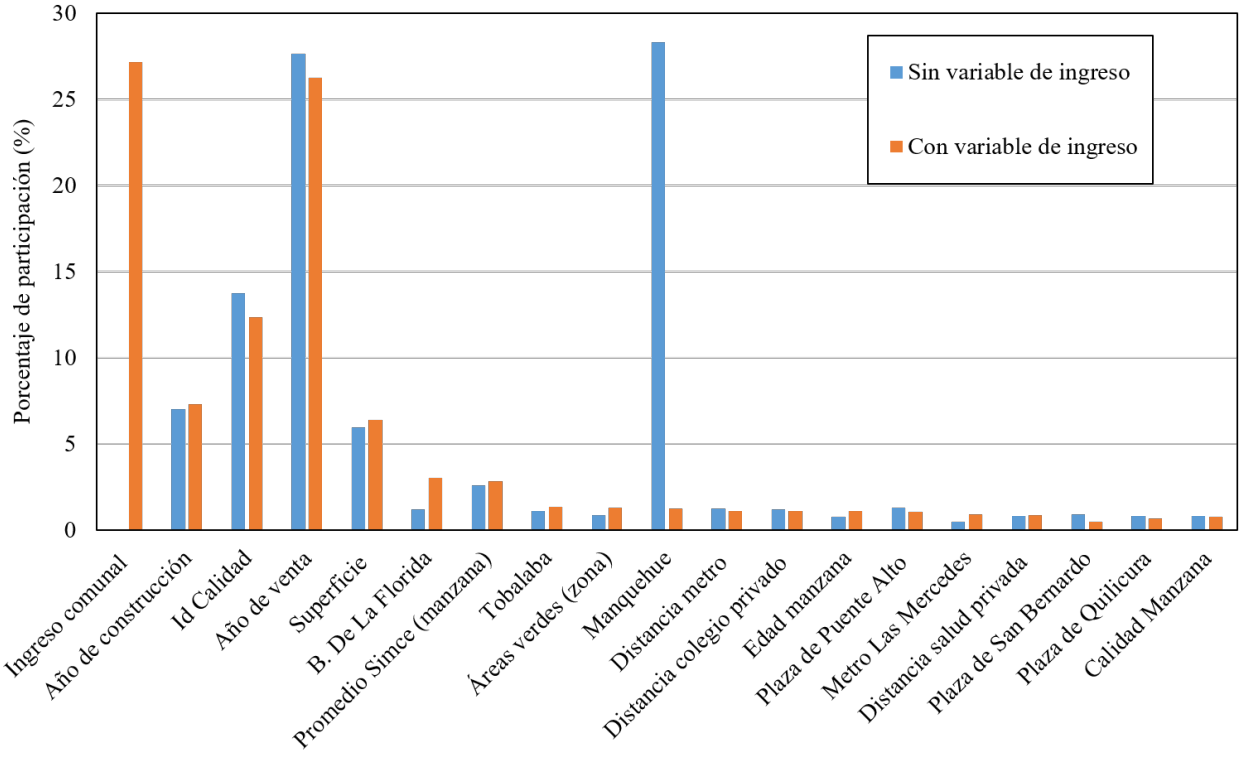


Figura 4.17: Importancia de las variables antes y después de agregar la variable de ingreso

## 4.5. Existencia de sesgo en la importancia de las variables

El error porcentual obtenido utilizando el bosque aleatorio es comparable con los resultados obtenidos por Antipov y Pokryshevskaya (2012) y Čeh et al. (2018) y muestra que el bosque aleatorio puede ser una herramienta poderosa para la predicción de rentas, incluso

cuando las muestras son pequeñas. Sin embargo, el cálculo de las importancias de las variables muestra que en algunos casos, variables no relevantes correlacionadas con otras que lo son pueden ser elegidas por el bosque aleatorio como variables importantes. Este fenómeno se pudo observar al calcular la importancia de las variables en la zona surponiente, donde la variable distancia a la plaza de Maipú alcanza los 30 % de porcentaje de participación. La representación espacial del índice de calidad en esta zona muestra que existe una segregación este-oeste en esta zona en término de calidad de la construcción, por lo que existe una correlación entre el índice de calidad de construcción de los edificios y la distancia a la plaza de Maipú. El porcentaje de participación obtenido por la variable distancia a la plaza de Maipú señala que el algoritmo separó la muestra en base a esta variable, aunque la diferencia de precio entre el este y el oeste de la zona es explicado por la calidad de la construcción de los edificios. En esta sección se presentan las fuentes de sesgo en el cálculo de la importancia de las variables y las soluciones para evitarlo a partir de los trabajos de Strobl et al. (2007) y Strobl, Boulesteix, Kneib, Augustin y Zeileis (2008).

En esta memoria se utiliza el algoritmo de bosque aleatorio tal como descrito en la sección 1.2.1, y se calcula la importancia de las variables a partir del índice de Gini de los nodos del árbol. Otra manera de calcular la importancia de una variable ($X_j$) es calcular la precisión del bosque aleatorio antes y después de retirarla del conjunto de las variables predictoras. Este método se llama importancia basada en permutaciones (*permutation importance*).

La base de datos utilizada contiene variables continuas (por ejemplo la distancia a los centros urbanos y el índice de áreas verdes) y variables discontinuas con distintas cantidades de categorías (la variable índice de calidad comporta cuatro categorías y la variable que indica el estado nuevo o usado del bien es una variable *dummy*). Cuando la base de datos contiene variables de distintos tipos, hay un posible sesgo cuando se calculan la importancia de las variables. Strobl et al. (2007) revela la existencia de este sesgo y muestra que al calcular las importancias de las variables, se favorecen las variables continuas y discontinuas con muchas categorías, y las variables correlacionadas entre ellas.

### 4.5.1. Sesgo por tipo de variable

Como el cálculo de las importancias de las variables depende directamente de la arquitectura de los árboles, el sesgo en la importancia también proviene de un sesgo en la construcción del árbol. Por una parte, cuando se computa la importancia de Gini de una variable $X_m$, se consideran todos los nodos donde se utiliza $X_m$ para dividirlos y se calcula la variación del índice de Gini entre los nodos hijos y el nodo padre. De esta manera, la importancia de Gini de $X_m$ depende del número de nodos divididos en base a $X_m$. Por otra parte, cuando se divide un nodo, se calcula la variación del índice de Gini entre los nodos hijos y su nodo padre por cada variable $X_m$ y cada umbral $S$. Entonces, las variables continuas y discontinuas con muchas categorías tienen más probabilidad de ser elegidas porque presentan un número alto de posibles umbrales $S$. Además, tienden a ubicarse más cerca del nodo raíz.

En el caso de la importancia en base a permutaciones, el sesgo también proviene de la arquitectura de los árboles. Por construcción, las variables continuas y con muchas categorías son utilizadas para dividir muchos nodos, con una preferencia para los nodos ubicados cerca

del nodo raíz. Entonces, la permutación de una variable $X_m$ que aparece muchas veces en el árbol afecta a muchos nodos, es decir a muchas observaciones. Por esta razón, la variación de la precisión del árbol antes y después de permutar una variable continua o con muchas categorías $X_m$ es más importante.

Los árboles de inferencia condicional (Hothorn et al., 2006) son construidos en base a una partición recursiva no sesgada (*unbiased recursive partitioning*) y permiten calcular la importancia de las variables en base a permutaciones sin sesgo cuando las muestras utilizadas para construir los árboles son submuestras aleatorias sin reemplazo y no de tipo *bootstrap*.

### 4.5.2. Sesgo por correlación entre las variables predictoras

Otra fuente de sesgo en el cálculo de la importancia en base a permutaciones es la existencia de variables correlacionadas entre ellas en el conjunto de la variables predictoras, como es el caso en la base de datos utilizada en este trabajo. Por ejemplo, el índice de calidad de los bienes calculado por el Servicio de Impuestos Interno toma en cuenta la fecha de construcción de los departamentos por lo que esta variable y la variable fecha de construcción están correlacionadas. Strobl et al. (2007) explica el origen del sesgo mostrando que la hipótesis nula considerada para detectar correlaciones entre una variable predictora $X_m$ et la variable a predecir $Y$ es:

$$H_o : X_m \perp Y, Z \tag{4.1}$$

donde Z es el conjunto de variables predictoras, menos $X_m$.

Entonces, si $X_m$ resulta importante, es decir que se violó $H_o$, puede significar que $X_m$ tiene correlación con $Y$ o que $X_m$ tiene correlación con Z, por lo que es necesario introducir un nuevo método de cálculo de la importancia de las variables que pueda detectar solamente las correlaciones entre $X_m$ e $Y$. Strobl et al. (2008) propone la importancia condicional (*conditional variable importance*), que consiste en retirar la variable $X_m$ del conjunto de variables predictoras solamente para un conjunto de observaciones tal que $Z = z$ para calcular la importancia basada en permutaciones. Para ello, utilizan la partición de las variables que resulta de la construcción de los árboles y muestran que este nuevo método, aunque no elimine totalmente la preferencia para las variables correlacionadas, permite reflejar mejor la importancia de cada variable.

Finalmente, cuando la base de datos contiene variables de distintos tipos y que algunas de estas variables están correlacionadas entre ellas, el algoritmo de bosque aleatorio adaptado a un problema de regresión es un bosque compuesto de árboles de inferencia condicional (Hothorn et al., 2006). Además, la importancia de las variables se debe calcular utilizando el método de importancia condicional propuesto por Strobl et al. (2008).

# Conclusión

En este trabajo se presentaron los resultados de la predicción de rentas de departamentos de Santiago de Chile utilizando el método de bosque aleatorio, después de haber comparado el poder predictivo de la red neuronal, del algoritmo SVR y del bosque aleatorio a partir de una muestra de transacciones de la comuna de La Florida. De manera similar a Antipov y Pokryshevskaya (2012), se mostró que baja el error absoluto porcentual medio al elegir como variable a predecir el precio del metro cuadrado de los departamentos y no su renta total. El bosque aleatorio genera el error porcentual más bajo (9,67 %) y más estable cuando cambia el año de venta de los bienes. La red neuronal y el algoritmo SVR (19,17 % y 14,69 % de error absoluto porcentual medio respectivamente) son más sensibles al número de departamentos vendidos por año, y ambos ven su poder predictivo afectado cuando disminuye el número de observaciones. La red neuronal es el algoritmo que requiere más tiempo de ejecución, seguido por el bosque aleatorio y el algoritmo SVR. La distribución espacial del error, obtenida mediante un software GIS, es similar para los tres algoritmos, lo que sugiere que la base de datos no contiene todas las variables explicativas necesarias.

Utilizando el algoritmo de bosque aleatorio, se obtuvo un error absoluto porcentual medio de 11 % para la predicción de rentas de departamentos en la ciudad de Santiago . Este valor, comparable con el valor encontrado en otras ciudades del mundo con el modelo hedónico y la red neuronal, permite concluir que el bosque aleatorio es una herramienta adaptada al problema de rentas. La separación de la base de datos en distintas zonas geográficas basadas en el trabajo de Fuentes et al. (2017) y en dos sectores de ingresos permitió confirmar la heterogeneidad del mercado inmobiliario en Santiago y entender cómo contribuyen los atributos de los departamentos en el valor de la renta, en los distintos casos. Sin embargo, la fuerte correlación entre los atributos hace difícil capturar los efectos aislados de estos.

Finalmente, el bosque aleatorio es un método fácil de implementar que puede funcionar con muestras de datos de varios tamaños. Además, el funcionamiento del bosque aleatorio es parecido al proceso de tasación real, lo que constituye una razón adicional para utilizarlo para la predicción de rentas. Sin embargo, existe un sesgo en el cálculo de las variables cuando las variables explicativas son correlacionadas entre ellas o cuando son de distintos tipos (continuas y discretas). En este caso, la importancia de las variables continuas o discretas es artificialmente alta porque la probabilidad de seleccionarlas para dividir los nodos del árbol es mayor (Strobl et al., 2007). Las variables correlacionadas entre ellas también cobran artificialmente más importancia porque violan la hipótesis nula de independencia entre los atributos de los bienes y su precio (Strobl et al., 2008). Los árboles de inferencia condicional (Hothorn et al., 2006) con muestras aleatorias permiten obtener importancias no sesgadas.

# Bibliografía


Allen, W. y Zumwalt, J. (1994). Neural Networks: a word of caution. *Unpublished Working Paper, Colorado State University*, 127-145.

Antipov, E. A. y Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications*, *39*(2), 1772-1778.

Barabasi, A.-L. y Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, *5*(2), 101.

Borst, R. A. (1991). Artificial neural networks: the next modelling/calibration technology for the assessment community. *Property Tax Journal*, *10*(1), 69-94.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Čeh, M., Kilibarda, M., Lisec, A. y Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS International Journal of Geo-Information*, *7*(5), 168.

Clark, D. E. y Herrin, W. E. (2000). The impact of public school attributes on home sale prices in California. *Growth and change*, *31*(3), 385-407.

Des Rosiers, F., Lagana, A., Thériault, M. y Beaudoin, M. (1996). Shopping centres and house values: an empirical investigation. *Journal of Property Valuation and Investment*, *14*(4), 41-62.

Do, A. Q. y Grudnitski, G. (1992). A neural network approach to residential property appraisal. *The Real Estate Appraiser*, *58*(3), 38-45.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J. y Vapnik, V. (1997). Support vector regression machines. En *Advances in neural information processing systems* (pp. 155-161).

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Siam.

Figueroa, E. (1992). Determinantes del precio de la vivienda en Santiago: Una estimación hedónica. *Estudios de Economıa*, *19*(1), 67-84.

Fuentes, L., Mac-Clure, O., Moya, C. y Olivos, C. (2017). Santiago de Chile: ?ciudad de ciudades? Desigualdades sociales en zonas de mercado laboral local. *Revista CEPAL*.

Garrod, G. D. y Willis, K. G. (1992). Valuing goods' characteristics: an application of the hedonic price method to environmental attributes. *Journal of Environmental management*, *34*(1), 59-76.

Gevrey, M., Dimopoulos, I. y Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecological modelling*, *160*(3), 249-264.

Glorot, X., Bordes, A. y Bengio, Y. (2011). Deep sparse rectifier neural networks. En *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 315-323).

Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J. y Seung, H. S. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*(6789), 947.

He, K., Zhang, X., Ren, S. y Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. En *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).

Hornik, K., Stinchcombe, M. y White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

Hothorn, T., Hornik, K. y Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651-674.

Lavın, F. V., Dresdner, J. y Aguilar, R. (2011). The value of air quality and crime in Chile: a hedonic wage approach. *Environment and Development Economics*, *16*(3), 329-355.

Lin, H. y Chen, K. (2011). Predicting price of Taiwan real estates by neural networks and support vector regression. En *Proceedings of the 15th WSEAS International Conference on Systems, Corfu Island, Greece,* (Vol. 7, *14*, p. 2011).

Louppe, G., Wehenkel, L., Sutera, A. y Geurts, P. (2013). Understanding variable importances in forests of randomized trees. En *Advances in neural information processing systems* (pp. 431-439).

McCulloch, W. S. y Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Quiroga, B. F. (2005). Precios hedónicos para valoración de atributos de viviendas sociales en la región metropolitana de Santiago.

Richardson, H. W., Vipond, J. y Furbey, R. A. (1974). Determinants of urban house prices. *Urban studies*, *11*(2), 189-199.

Ridker, R. G. y Henning, J. A. (1967). The determinants of residential property values with special reference to air pollution. *The Review of Economics and Statistics*, 246-257.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, *82*(1), 34-55.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Rowley, H. A., Baluja, S. y Kanade, T. (1998). Neural network-based face detection. *IEEE Transactions on pattern analysis and machine intelligence*, *20*(1), 23-38.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Rumelhart, D. E. y McClelland, J. L. (1986). Parallel distributed processing: explorations in the microstructure of cognition. volume 1. foundations.

So, H. M., Tse, R. Y. y Ganesan, S. (1997). Estimating the influence of transport on house prices: evidence from Hong Kong. *Journal of Property Valuation and Investment*, *15*(1), 40-47.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. y Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, *15*(1), 1929-1958.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T. y Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, *9*(1), 307.

Strobl, C., Boulesteix, A.-L., Zeileis, A. y Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, *8*(1), 25.

Tay, D. P. y Ho, D. K. (1992). Artificial intelligence and the mass appraisal of residential apartments. *Journal of Property Valuation and Investment*, *10*(2), 525-540.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tikhonov, A. N. (1963). On the solution of ill-posed problems and the method of regularization. En *Doklady Akademii Nauk* (Vol. 151, *3*, pp. 501-504). Russian Academy of Sciences.

Vargas, M. (2006). Causes of Residential Segregation: the Case of Santiago, Chile. *mimeo, Centre for Spatial and Real Estate Economics, Department of Economics, The University of Reading, United Kingdom.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... y Polosukhin, I. (2017). Attention is all you need. En *Advances in Neural Information Processing Systems* (pp. 5998-6008).

Vega Flores, R. A. (2017). Valoración de amenidades urbanas mediante precios hedónicos: el caso de Santiago de Chile.

Worzala, E., Lenk, M. y Silva, A. (1995). An exploration of neural networks and its application to real estate valuation. *Journal of Real Estate Research*, *10*(2), 185-201.

Yoo, S., Im, J. y Wagner, J. E. (2012). Variable selection for hedonic model using machine learning approaches: A case study in Onondaga County, NY. *Landscape and Urban Planning*, *107*(3), 293-306.

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.