Precios de venta del mercado inmobiliario de Melbourne

Jorge Luis Ramos Zavaleta
Centro de Investigación en Matemáticas. Unidad Monterrey.
Email: jorge.ramos@cimat.mx

***Resumen*—En este reporte se detalla un modelo de regresión multilineal que nos permitirá establecer variables de impacto que se puedan considerar como buenos predictores para establecer los precios de casas en la ciudad de Melbourne durante el año 2016. La información utilizada fue obtenida desde una página web *Domain.com.au* que contiene las ofertas inmobiliarias provistas por diversas agencias inmobiliarias del pais.**

I. Introducción

El problema de los precios óptimos de vivienda es un problema muy viejo y ha sido explorado desde diversos enfoques. Sin embargo es imposible considerar un modelo universal que gobierne por completo la relacióon entre las variables explicativas y el precio de las viviendas debido a que cada mercado a nivel desagregados tiende a tener sus particularidades por lo que las variables que son importantes en un mercado pueden no serlo en otro.

Aparte del punto anterior cabe mencionar que nuevos tipos de mecanismos de venta de viviendas se han estado presentando en paises economicamente desarrollados que permiten maximizar el precio final de venta, por lo que estos métodos no permiten que se refleje el precio real de la vivienda.

MelBourne es la capital de Victoria y es la ciudad más grande de dicho estado con una población de cerca de 4.5 millones dentro de su area urbana. Melbourne es reconocida como la capital deportiva de Australia ya que fue sede los juegos olimpicos en 1956 y los juegos de la mancomunidad en 2006, aparte de que es sede de otros varios eventos deportivos internacionales.

Los datos utilizados para este reporte fueron extraídos a lo largo del año 2016 y una pequeña parte de 2017 (aunque estos últimos se eliminaron en el proceso de limpieza de los datos) de la página web *Domain.com.au*, que contiene información de venta y renta de inmuebles provista por distintas agencias inmobiliarias del pais lo que nos permitiría creer que es una información más completa y real sobre los inmuebles vendidos. El conjunto de datos fue conseguido a través de la plataforma **Kaggle.** [1]

[1]https://www.kaggle.com/anthonypino/melbourne-housing-market

Figura 1. Mapa del estado de Victoria, con el valor de 1 se encuentra la ciudad de Melbourne

II. Objetivo

El objetivo de este trabajo es identificar el efecto de algunas variables explicativas en el precio de la vivienda en Melbourne en el año 2016, intentando con ello predecir los precios de las ventas de viviendas en dicho año. Es decir se busca establecer la importancia de dichas variables para que en un futuro se puedan emplear en un método más robusto que incluya el efecto del tiempo para intentar predecir precios de viviendas a futuro en dicho mercado.

III. Antecedentes

Dado que los datos se encontraron en la plataforma **Kaggle** se han establecido varios intentos para encontrar relaciones entre el precio de venta y las variables explicativas, así como métodos de imputación de datos para los datos faltantes. Sin embargo las metodologías empleadas carecen de rigor estadístico por lo que los resultados que se obtienen son artificiales y no permiten indicar relaciones entre las variables explicativas y el precio de venta.

Por ejemplo se usa imputación por la media tanto para datos continuos como categóricos sin considerar la naturaleza de la variable ni el factor espacial. Por ejemplo en el caso del tamaño de la cochera respecto al número de autos que entran en ella en los suburbios más caros las cocheras pueden almacenar hasta 9 automoviles mientras que en otros suburbios las casas no cuentan con cochera, por lo que ese tipo de imputación implicaría pensar que las viviendas siguen una distribución normal sin embargo dado que se trata de un

dato categórico debe seguir una distribución discreta.

Otros errores que se encontraron fue el hacer uso de regresión sin considerar la distribución de los precios de las viviendas para verificar la normalidad de la variable y con eso cumplir uno de los supuestos de la regresión, este error se hará mas evidente más adelante en este reporte. Además de esto se realiza regresión sin establecer una codificación apropiada para los datos categóricos.

IV. El modelo

Como se menciono antes los datos se obtuvieron a través del minado de una página web. El total de observaciones es de 34,857 y contiene 21 variables, sin embargo algunas de estas variables aunque importantes eran redundantes con respecto a otras y en algunas era imposible darles una imputación significativa por lo que se optó por eliminarlas del modelo. Además de esto algunas variables agregan una complejidad mayor al modelo y se decidió omitirlas debido a que aún cuando se usaron variables de complejidad menor en este caso regiones en lugar de usar suburbios se encontraron problemas en el calculo de la regresión.

Primero que nada se observo que algunas de las observaciones de precios de viviendas no tenian asignado un valor por lo que se eliminaron dichas observaciones debido ya que su imputación podría haber alterado demasiado los resultados. Para realizar la imputación de los datos faltantes se considero como supuesto que las casas entre suburbios eran muy parecidas en sus características dado que esta es la medida geográfica con que contabamos para agrupar las casas.

Siguiendo la idea expuesta anteriormente se imputaron las variables *tamaño del terreno, número de baños y número de coches* considerando la mediana por suburbio. En el caso de que no fuese posible imputar de esta manera las variables se eliminaban los registros con NA's, en este caso solo sucedio en 26 observaciones por parte de la variables *tamaño del terreno*. En el caso de la variable *tamaño del terreno* también se tenian observaciones inconsistentes como el hecho de que el terreno en donde estaba situada la casa era de 0 m^2, por lo que se opto por cambiar los registros con un número menor de 100 metros cuadrados por 100 metros cuadrados.

Por último se imputo la variable *tamaño de la edificación* haciendo uso de la mediana por suburbio, y cuando no era posible esto se usaba una proporción entre 40 % y 90 % del tamaño del terreno elegido de manera aleatoria (fijando una semilla de 100). Con este proceso armado terminamos con 27,169 observaciones. Se realizó una primera versión del modelo sin considerar imputaciones borrando todos los registros con información faltante y se encontró una diferencia significativa en la mediana de los errores de predicción de cerca de 5 mil dólares australianos de más, por lo que se considero que la imputación fue una buena medida para mejorar el modelo.

Se inició un proceso exploratorio para ver que variables elegir. Primero se observo la distribución de la variable *Precio de la vivienda* la cual puede observarse en la figura 2, y dado que no representa una distribución normal entonces se considero su logaritmo el cual tiene una forma más cercana a la distribución normal la cual puede observarse en la figura 3, y se realizó un Q-Q plot que puede observarse en la figura 4, y se observa que dicha variable cerca de las colas tiende a despegarse un poco de la recta normal, pero para fines prácticos ese pequeño desplazamiento puede considerarse como despreciable.

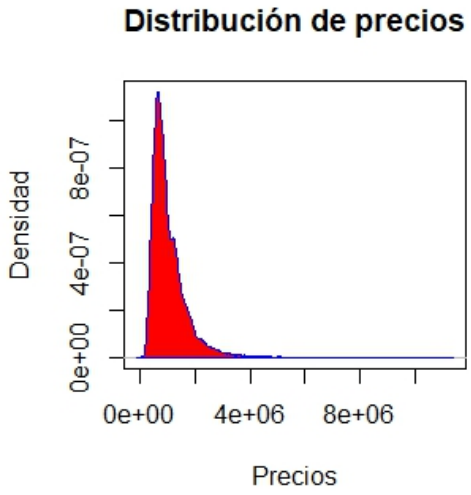

Figura 2. Distribución de los precios

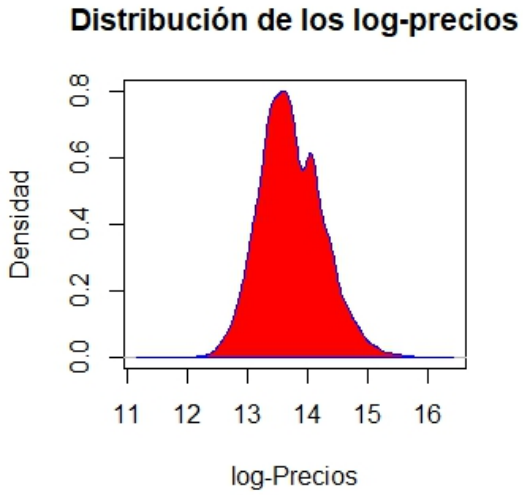

Figura 3. Distribución de los log-precios

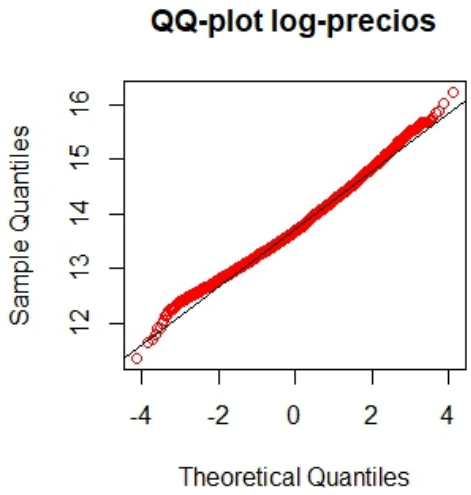

Figura 4. Q-Q plot de los log-precios

En el caso del tamaño del terreno se observaron la presencia de datos atípicos oara algunos suburbios, lo cual puede verificarse en la figura 5 donde se muestran círculos por tamaño correspondiente a la mediana del tamaño por suburbio, y se puede observar que algunos de los suburbios tienen terrenos muy amplios a la venta comparados con el resto de los terrenos. Además se considero el precio promedio por vivienda en cada suburbio observandose que los precios más altos se encuentran en la zonas más cercanas de la central corporativa de la ciudad como se observa en la figura 6 y se encuentran principalmente en la región sur metropolitana como. Las visualizaciones se hicieron en *Tableau Public* por lo que pueden ser consultadas de manera interactiva.[2]

Figura 5. Distribución de la mediana del tamaño de terreno por suburbio en la ciudad de Melbourne

[2]https://public.tableau.com/profile/jorge.ramos3092#!/vizhome/CasasMelbourne/Precio-Suburbio

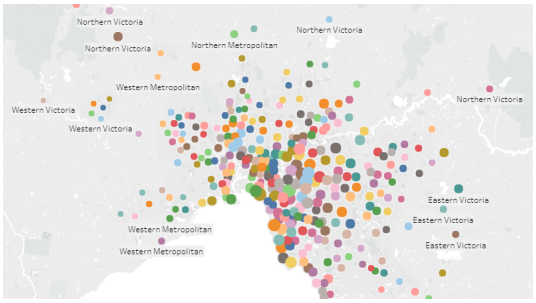

Figura 6. Distribución del precio promedio de las viviendas por suburbio en la ciudad de Melbourne

Se prosiguió el análisis descriptivo observando el total del monto de ventas con respecto al método usado para la venta lo cual se detalla en la figura 7, donde se deja ver que el método SA y probablemente el VB pueden no ser relevantes estadísticamente ya que su monto de ventas es muy bajo comparado con los otros métodos. El método SA corresponde a una propiedad vendida por subasta y el método VB a una venta por subasta con oferta del vendedor. El método PI corresponde a una propiedad pasada, es decir una que no alcanzó el precio de reserva en una subasta y se vende al mejor postor como una negoción extraoficial. El método S corresponde a una venta tradicional y el método Sp corresponde a una venta que se realiza antes pero se pacta una temporada para desalojar el inmueble.

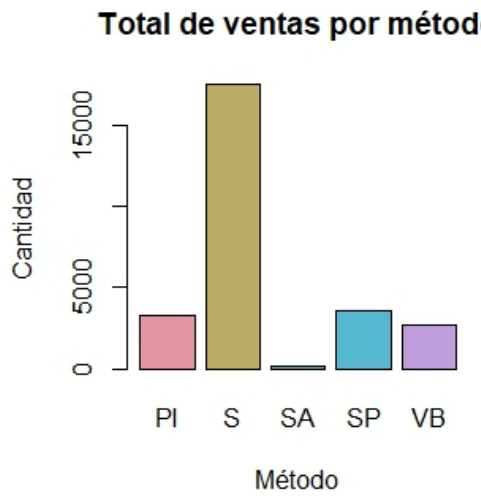

Figura 7. Ventas por método empleado

Por otro lado se considero el total de viviendas vendidas con respecto al tipo de vivienda. En la figura 8 se puede apreciar que las ventas de casas unitarias y duplex etiquetadas como unit pueden tener poca relevancia estadística debido al número tan bajo que representan. Por último se consideraron la cantidad de casas vendidas por tamaño de terreno, debido

al número de observaciones distintas de esta variable se optó por usar la división de esta variable en sus cuartiles y con ello expresar la cantidad de casas vendidas por cada cuartil de la variable, esto se observa en la figura 9 donde puede apreciarse que dicha variable si tendrá un peso significativo en el modelo.

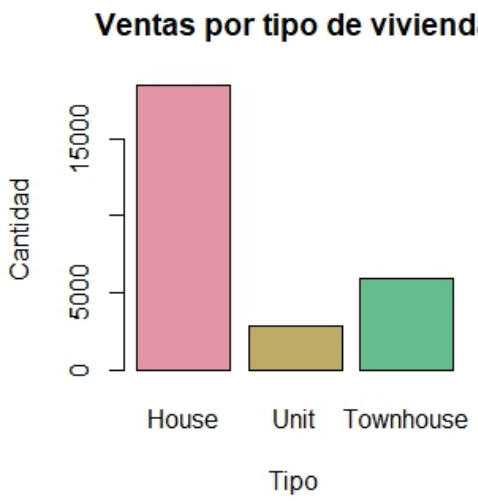

Figura 8. Casas vendidas por tipo de vivienda

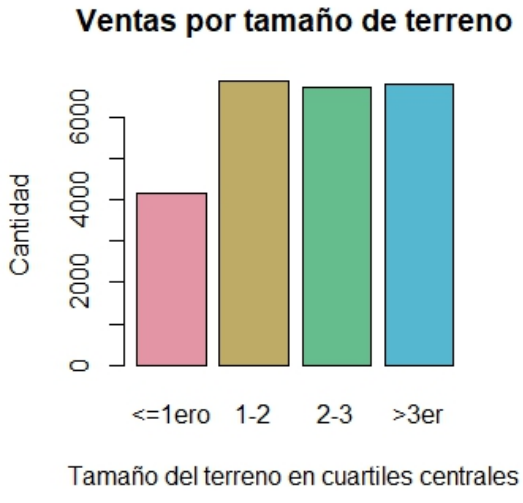

Figura 9. Ventas por cuartiles centrales de tamaño del terreno

Para el modelo se uso una codificación simple para las variables categóricas que en este caso fueron *Mehod, type y Regionname* por lo que cada etiqueta dentro de la categoria se consideraba una variable y se colocaba un uno en caso de aparición y un cero en caso contrario. Las variables elegidas para el primer modelo fueron:

- Distance
- Bathroom
- Cars
- Landsize
- Rooms
- Type
- Method
- Regionname

Estas ultimas 3 con su respectiva codificación. Un vistazo a todas las variables originales con su descripción se detallará en el anexo a este reporte además del porqué no se eligieron ciertas variables.

Además de lo anterior en el anexo se agregan algunos gráficos de correlación entre las variables continuas y otro con respecto a las discretas para entrever relaciones apriori entre ellas. Algunas relaciones importantes que son conocidas en la literatura especializada en el modelado de bienes raíces es el número de baños y el número de cuartos con respecto al precio de la vivieda y en los gráficos se aprecia esta relación. Otra relación importante es la correlación negativa entre la distancia entre los suburbios y el centro de negocios y el precio, lo cual es claro indicador de que el centro de negocios juega un papel muy importante en la ciudad. Otra relación importante es el número de coches, el número de cuartos y el número de baños por vivienda con la distancia al centro de negocios indicandonos que las casas más grandes estan generalmente más alejadas del centro de negocios.

Por parte de las correlaciones entre las variables discretas encontramos correlaciones negativas entre los diferentes tipos de vivienda, así también entre los distintos métodos de venta lo que es indicativo de la preferencia de elección de un tipo de vivienda por otra y de un método por otro. En el caso de las regiones se encontraron correlaciones muy pequeñas lo que es indicativo de que no hay presencia de autocorrelación espacial al generar la separación por región, debido a que los precios entre regiones tiende a ser muy variable aún cuando una este cerca de otra.

Sin embargo debido a que como se detallo anteriormente algunas de las variables categóricas tienen muy pocos datos por lo que entre ellas pueden generarse combinaciones lineales lo que resultó en problemas de colinealidad. Y en este caso primer modelo con todas las variables y sus codificaciones termino resultando un modelo con multicolinealidad. Por lo que se decidió eliminar las subvariables generadas a partir de la codificación con un número de observaciones muy pequeño, con lo que se eligió eliminar cuatro variables dummy en este caso fueron: Viviendas de tipo unit, ventas por método SA, ventas por método VB y pertenencia a la región de Western Virginia. Los resultados obtenidos en cada modelo se detallarán en el anexo.

Cabe mencionar que para este segundo modelo casi todas las variables resultaron estadísticamente significativas con excepción de la pertenencia a la región del este metropolitano.

Con este modelo se obtuvo una R^2 ajustada de 0.6821 y un error cuadrado medio de 0.08508 por lo que puede considerarse un modelo suficientemente explicativo del fenómeno de los log-precios de las viviendas en Melbourne.

El modelo final quedo de la siguiente manera

$$\begin{aligned} logPrice_{2016} = {} & 12{,}75 - 0{,}03994 Distance + \\ & 0{,}09557 Bathrooms + 0{,}02644 Cars + \\ & 0{,}0000033 Landsize + 0{,}000025 BuildingArea + \\ & 0{,}1845 Rooms + 0{,}529 TypeH + 0{,}3221 TypeT - \\ & 0{,}017 MethodPI + 0{,}0593 MethodS + 0{,}0157 MethodS \\ & 0{,}2042 RegionEM + 0{,}3551 RegionEV + \\ & 0{,}6057 RegionNM + 0{,}3817 RegionSEM + \\ & 0{,}5411 RegionSM + 0{,}4772 RegionWM \end{aligned}$$

Y las gráficas de residuos vs ajustados se encuentra en la figura 10, aquí no se puede observar ninguna tendencia en la gráfica por lo que se procedió a hacer el gráfico de densidad de los residuos que se encuentra en la figura 11 y puede observarse que es bastante cercana a la distribución normal por lo que podemos presuponer que el modelo cumple con los supuestos de regresión.

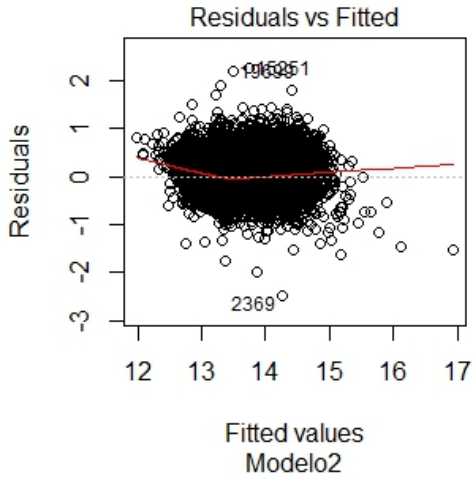

Figura 10. Gráfica de Residuos vs Ajustados del modelo 2

V. Resultados

Primer hay que considerar que para este modelo no se estan considerando algunas variables que pueden influir directamente en el precio de las viviendas como puede ser alguna medición de criminalidad por región dado que nuestro modelo se esta pensado regionalmente, o la cercanía a servicios de emergencia, educación y transporte, además de que no se consideró una variable muy importante que es la antigüedad de la edificación por lo que no esperamos una predicción de los precios exacta, pero dado que ya encontramos variables importantes que ayudan a predecir los log-precios el objetivo del trabajo se logro.

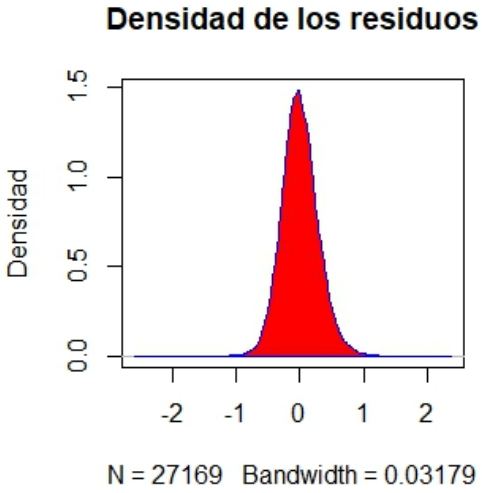

Figura 11. Distribución de los residuos del modelo 2

Comparando los valores predichos con los precios-logarítmo reales de venta de viviendas se obtuvo un error mínimo de $3{,}48x10^{-6}$, una mediana del error de $0{,}1823$ y un error máximo de $2{,}4913$. Sin embargo considerando los suburbios los resultados pueden variar mucho entre cada uno de ellos.

En el anexo se agregaron un par de gráficos donde se pueden observar las diferencias entre los log-precios para las primeras 100 observaciones y para el suburbio Meadow Heights.

VI. Referencias

Kleiber, C., & Zeileis, A. (2008). Applied econometrics with R. Springer Science & Business Media.

Linneman, B. P. (2011). Real Estate Finance and Investments. Philadelphia: Linneman Associates.

Richard, A. J., & Dean, W. W. (2002). Applied multivariate statistical analysis. London: Prenticee Hall, 265.

Variables originales

- **Suburb:** Distintos suburbios de Melbourne. Esta variable se uso para imputar por lo que se espera que genere efectos en el modelo sin incluirla explicitamente.
- **Address:** El domicilio de la vivienda. No se uso esta variable porque no tiene un poder explicativo directo como su cercanía al centro de Melbourne.
- **Rooms:** Numero de cuartos de la vivienda.
- **Price:** Precio de venta de la vivienda en dólares australianos
- **Method:** S - Propiedad vendida tradicionalmente; SP - Propiedad vendidad con previo aviso; PI - Propiedad pasada; VB - Subasta con posibilidad de recompra por parte del vendedor; SA - Venta por subasta normal.
- **Type:** br - bedroom(s); h - house,cottage,villa, semi,terrace; u - unit, duplex; t - townhouse; dev site - development site; o res - other residential. Existían otros métodos pero desaparecieron debido a que no reportaron el precio de venta de la vivienda.
- **SellerG:** Agente inmobiliario.
- **Date:** Fecha de venta. No se uso debido a que solo eran unas cuantas observaciones de 2017 y se eliminaron prácticamente al limpiar los datos, por lo que se consideraron dichos precios como si fueran del 2016.
- **Distance:** Distancia hacia el centro de negocios de Melbourne en km.
- **Regionname:** Regiones de la ciudad.
- **Propertycount:** Número de propiedades en el suburbio. No se uso debido a que tenía una presencia muy grande de datos atípicos.
- **Bedroom2:** Número de cuartos por vivienda obtenidos de otro sitio web. No se incluyo por que era redundante con la variable rooms.
- **Bathroom:** Número de baños en la vivienda.
- **Car:** Número de lugares disponibles para automóvil en las viviendas.
- **Landsize:** Tamaño de del terreno.
- **BuildingArea:** Tamaño de la edificación.
- **YearBuilt:** Año de la construcción de la edificación. No se uso porque tenía muchos datos faltantes y no había posibilidad de imputar usando los suburbios.
- **CouncilArea:** Area de gobierno a la que pertenece la vivienda. No se utilizo porque presentaría redundancias con respecto a la variable regionname.
- **Lattitude:**
- **Longtitude:** Estas últimas variables no se emplearon directamente pero se utilizaron para el análisis exploratorio.

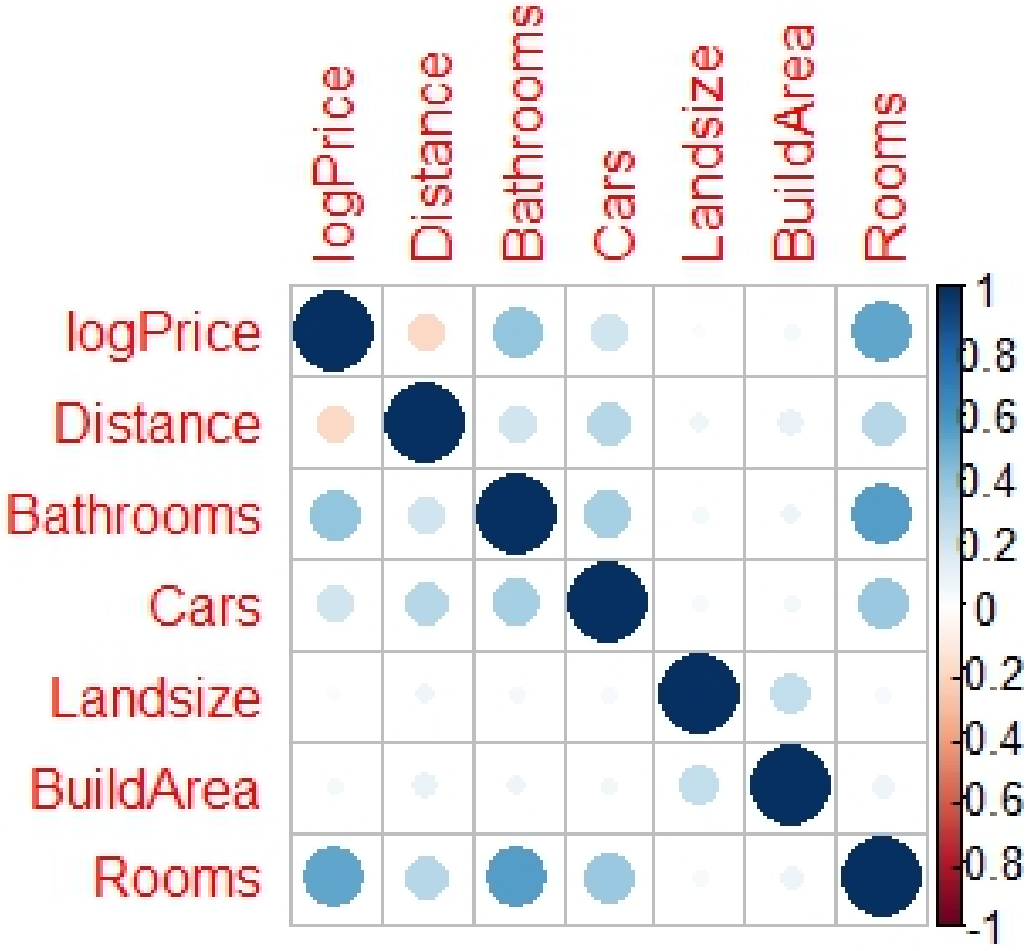

Figura 12. Correlaciones entre variables continuas.

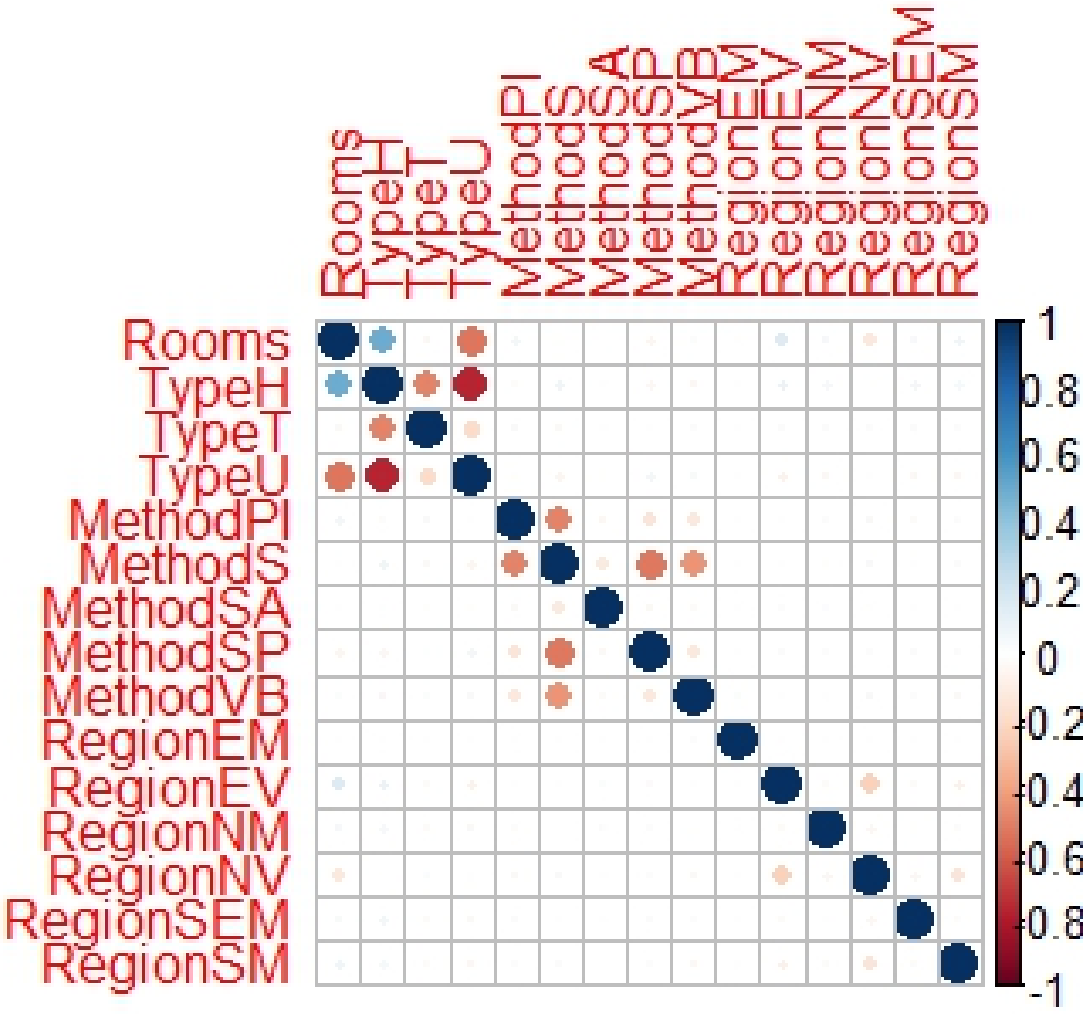

Figura 13. Correlaciones entre variables discretas codificadas.

Resultados del primer modelo

```
Call:
lm(formula = logPrice ~ ., data = new.df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49139 -0.18900 -0.01179  0.17533  2.28073

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.280e+01  3.253e-02  393.551  < 2e-16 ***
Distance    -4.012e-02  3.694e-04 -108.595  < 2e-16 ***
Bathrooms    9.580e-02  3.370e-03   28.423  < 2e-16 ***
Cars         2.653e-02  2.216e-03   11.973  < 2e-16 ***
Landsize     3.300e-06  5.937e-07    5.558 2.76e-08 ***
BuildArea    2.520e-05  3.476e-06    7.249 4.31e-13 ***
Rooms        1.847e-01  2.687e-03   68.718  < 2e-16 ***
TypeH        5.289e-01  5.408e-03   97.803  < 2e-16 ***
TypeT        3.222e-01  6.979e-03   46.173  < 2e-16 ***
TypeU               NA         NA       NA       NA
MethodPI    -1.445e-02  7.627e-03   -1.894   0.0582 .
MethodS      6.193e-02  6.134e-03   10.096  < 2e-16 ***
MethodSA     3.678e-02  2.206e-02    1.667   0.0955 .
MethodSP     1.832e-02  7.570e-03    2.420   0.0155 *
MethodVB            NA         NA       NA       NA
RegionEM     1.503e-01  2.087e-01    0.720   0.4714
RegionEV     3.010e-01  3.112e-02    9.672  < 2e-16 ***
RegionNM     5.544e-01  3.796e-02   14.606  < 2e-16 ***
RegionNV     4.980e-03  3.122e-02    0.160   0.8733
RegionSEM    3.305e-01  3.780e-02    8.744  < 2e-16 ***
RegionSM     4.886e-01  3.125e-02   15.637  < 2e-16 ***
RegionWM     4.224e-01  3.128e-02   13.505  < 2e-16 ***
RegionWV    -5.543e-02  3.117e-02   -1.778   0.0754 .
---

Residual standard error: 0.2918 on 27148 degrees of freedom
Multiple R-squared:  0.6823,    Adjusted R-squared:  0.6821
F-statistic:  2916 on 20 and 27148 DF,  p-value: < 2.2e-16
```

Resultados del segundo modelo

```
Call:
lm(formula = logPrice ~ . - TypeU - MethodSA - RegionWV - MethodVB,
    data = new.df)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49134 -0.18943 -0.01196  0.17525  2.28108

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)
(Intercept)  1.275e+01  9.235e-03 1380.495  < 2e-16 ***
Distance    -3.994e-02  3.595e-04 -111.108  < 2e-16 ***
Bathrooms    9.557e-02  3.369e-03   28.366  < 2e-16 ***
Cars         2.644e-02  2.216e-03   11.935  < 2e-16 ***
Landsize     3.301e-06  5.937e-07    5.559 2.73e-08 ***
BuildArea    2.510e-05  3.476e-06    7.221 5.31e-13 ***
Rooms        1.845e-01  2.687e-03   68.679  < 2e-16 ***
TypeH        5.290e-01  5.408e-03   97.817  < 2e-16 ***
TypeT        3.221e-01  6.978e-03   46.161  < 2e-16 ***
MethodPI    -1.700e-02  7.487e-03   -2.270   0.0232 *
MethodS      5.930e-02  5.949e-03    9.967  < 2e-16 ***
MethodSP     1.568e-02  7.419e-03    2.113   0.0346 *
RegionEM     2.042e-01  2.064e-01    0.990   0.3224
RegionEV     3.551e-01  6.518e-03   54.481  < 2e-16 ***
RegionNM     6.057e-01  2.470e-02   24.526  < 2e-16 ***
RegionNV     5.977e-02  5.104e-03   11.709  < 2e-16 ***
RegionSEM    3.817e-01  2.432e-02   15.694  < 2e-16 ***
RegionSM     5.411e-01  9.942e-03   54.431  < 2e-16 ***
RegionWM     4.772e-01  5.148e-03   92.703  < 2e-16 ***
---

Residual standard error: 0.2918 on 27150 degrees of freedom
Multiple R-squared:  0.6823,    Adjusted R-squared:  0.6821
F-statistic:  3239 on 18 and 27150 DF,  p-value: < 2.2e-16
```

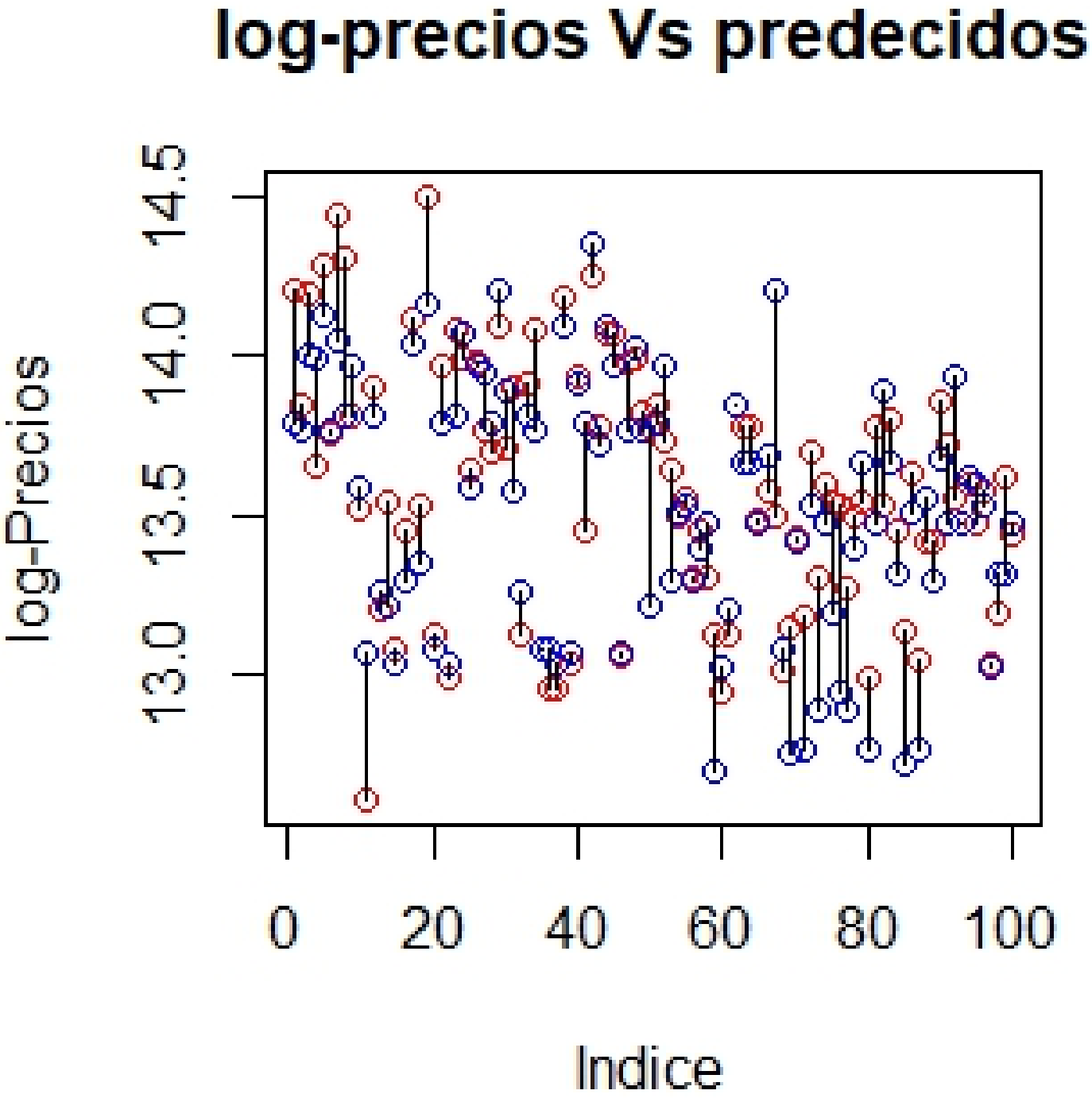

Figura 14. Distancias entre valores predecidos y los reales para las primeras 100 observaciones. En rojo se encuentran los valores reales y en azul los predecidos.

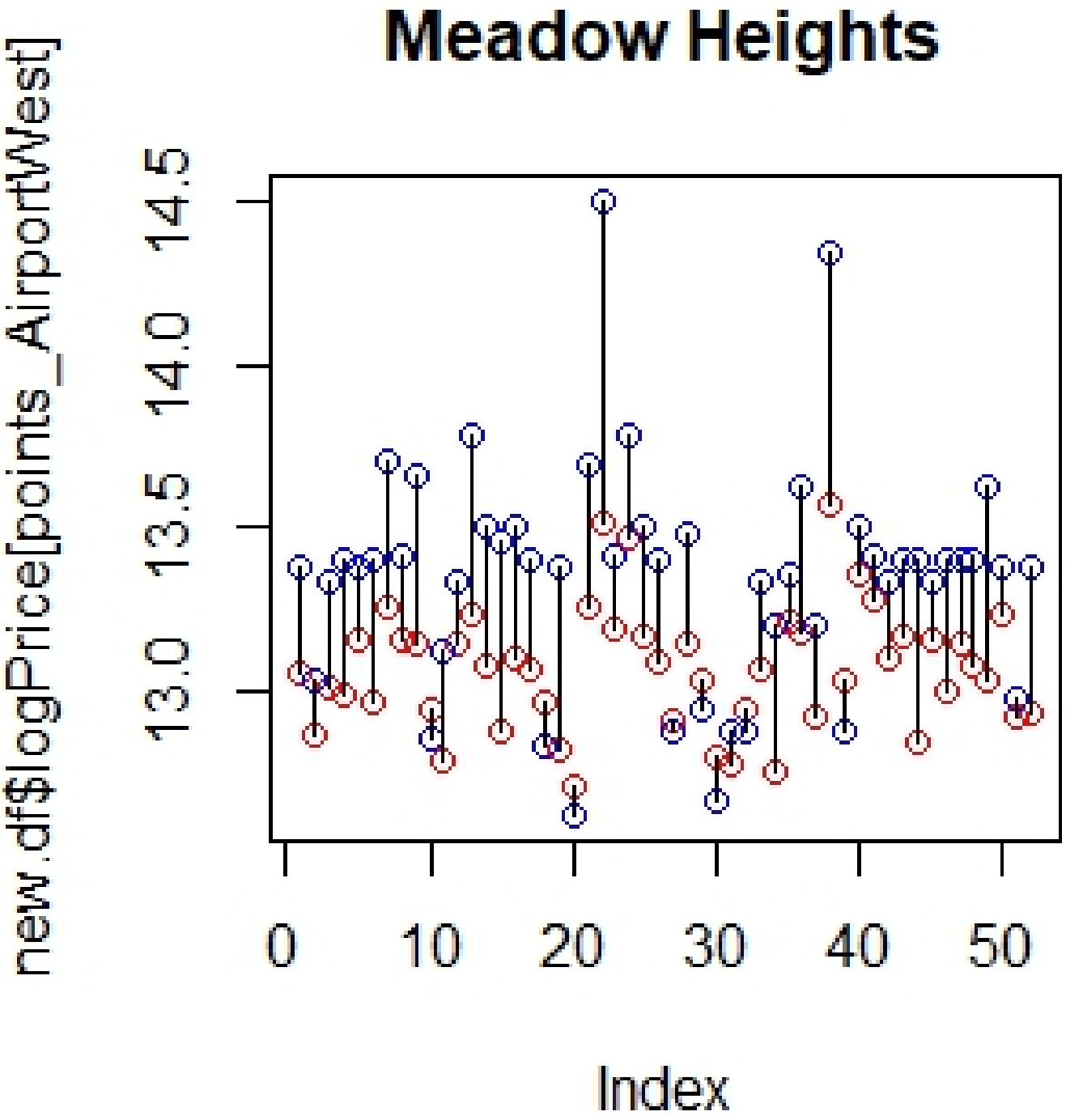

Figura 15. Distancias entre valores predecidos y los reales para el suburbio Meadow Heights. En rojo se encuentran los valores reales y en azul los predecidos.