DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Metros Cuadrados Mínimos Lineales

Métodos Numéricos

Integrante	LU	Correo electrónico
Gómez, Bruno Agustín	428/18	bgomez@dc.uba.ar

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

1. Resumen

La predicción de precios de propiedades a partir de sus principales características es de interés para realizar estimaciones de valoración sobre el precio relativo de propiedades, así como para determinar valores futuros de propiedades que aún no finalizaron su construcción. En el presente informe se propone realizar un modelado a partir de un conjunto de características o *features* que permita predecir el precio y la cantidad de metros cubiertos de un conjunto de propiedades pertenecientes a los Estados Unidos Mexicanos. Para ello se utilizó el método de regresión lineal múltiple. Previamente se agregaron posibles variables predictoras mediante mecanismos de *feature engineering* tanto de información de las descripciones como de fuentes externas. Además, se removieron o completaron los datos faltantes y se realizó un barrido de los *outliers*. Para determinar la precisión de las predicciones de los modelos se utilizaron métricas de error (MAE, MSE, RMSE y RMSLE) así como estimadores del modelo (R2-score y AIC). Se comparó su comportamiento con el del método de principal component regression (PCR), donde se observó que si bien PCR es más veloz, presenta peores resultados para un bajo número de variables. Se realizaron cuatro tipo de segmentación por tipo de propiedad, regiones geográficas y categorías de precios que permitieron observar mejores resultados que al analizar el modelo general. Se observó en líneas generales que los dos factores más importantes para obtener buenas estimaciones fueron el número de entradas presentes en los datos utilizados y la homogeneidad en el tipo de propiedad. Adem ás, se pudo observar la importancia del manejo previo de los datos, sobre todo con la remoción de *outliers*.

Palabras claves: *modelos predictivos, PCR, RMSLE, AIC, R2-score, experimentación, métodos numéricos, K-Fold, outliers*

2. Introducción

El precio de la vivienda varía según sus características. Tamaño, calidad y ubicación (rural, urbana, centro, periferia), así como factores y circunstancias socio económicas de los distintos países como demografía (crecimiento de la población, equilibrio o descenso, flujos migratorios), crecimiento económico, desempleo, precio del suelo, existencia o no de burbuja inmobiliaria, interés bancario, impuestos, desgravaciones, capacidad de endeudamiento personal y familiar, recesión y depresión económica entre otras. El precio de la vivienda condiciona la vida económica de los ciudadanos y familias que, en el caso de la compra de vivienda, habitualmente deben suscribir a créditos hipotecarios de larga duración con entidades bancarias para poder acometer la compra más importante a lo largo de la vida personal y familiar.

Existen indicadores estructurales que permiten valorar si el precio de una vivienda es adecuado y económicamente viable, ya sean indicadores sobre los ingresos o salarios (relación entre el precio de la vivienda y la renta bruta disponible), en relación con el precio del alquiler de la vivienda que quiere comprar (rentabilidad del alquiler) o en relación con el monto y duración de la hipoteca a solicitar. En el precio de la vivienda en relación con la renta anual disponible se considera adecuada una relación de 4, es decir que el precio de compra de la vivienda es económicamente viable si se aproxima a cuatro años de la renta familiar bruta disponible[1].

Además, el precio del metro cuadrado de la vivienda suele ser uno de los indicadores usuales para analizar la evolución de la riqueza real del sector privado en la región geográfica analizada. Su evolución no solamente interesa a intermediarios y empresas de la construcción, sino también a entidades financieras y extrabancarias que otorgan préstamos hipotecarios. Además, con el florecimiento de las inversiones en pozos para invertir o adquirir propiedades, resulta de interés tener algún tipo de modelo que permita estimar los precios finales a partir de las características que tendrá la propiedad al finalizar su construcción.

Es posible realizar predicciones confeccionando un modelo a partir de una regresión lineal simple o múltiple, dependiendo la cantidad de *features* o características que se incluyan. Observando la cantidad de factores que pueden afectar el precio de una propiedad, es comprensible que diseñar un modelo que

permita predecir precios a partir de un conjunto de características no sea tarea sencilla.

2.1. Método de cuadrados mínimos lineales

En los experimentos se realizan mediciones a partir de los fenómenos bajo estudio. Las mediciones realizadas normalmente no son precisas, pues existen errores asociados al instrumento de medición, la técnica empleada o al individuo. De hecho, si se realiza la misma medición varias veces, los resultados variarán. Entonces, ¿cuál es la mejor estimación para la medición real?.

El método de mínimos cuadrados proporciona una forma de encontrar la mejor estimación, suponiendo que los errores, es decir las diferencias con el valor verdadero, son aleatorios e imparciales. Es un procedimiento de análisis numérico en el que, dado un conjunto de datos (pares ordenados y familia de funciones), se intenta determinar la función continua que mejor se aproxima a los datos (línea de regresión o la línea de mejor ajuste), proporcionando una imagen que demuestra la relación entre los puntos de la misma. En su forma más simple, busca minimizar la suma de cuadrados de las diferencias ordenadas entre los valores predichos por el modelo y los observados.

Este método se usa comúnmente para analizar una serie de datos obtenidos de un estudio, con el fin de expresar su comportamiento de forma lineal y así minimizar los errores de los datos tomados.

> *Dado un conjunto de pares ordenados de valores* (x_i, y_i) *para* $i = 1, ..., m$ *se busca una función* $f(x)$ *perteneciente a una familia* $\mathcal{F}$ *tal que "mejor aproxime" a los datos.*

$$min_{f \in \mathcal{F}} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

Por otro lado, el método de cuadrados mínimos lineales es una reformulación de este problema.

> *Dado un conjunto de funciones* $\phi_1, ..., \phi_n$ *linealmente independientes, se define* $\mathcal{F} = f(x) = \sum_{j=1}^{n} c_j \phi_j$

$$min_{c_1,...,c_n} \sum_{i=1}^{m} (\sum_{j=1}^{n} c_j \phi_j(x_i) - y_i)^2$$

Sean $A \in \mathcal{R}^{m \times n}$, $b \in \mathcal{R}^m$ y $x \in \mathcal{R}^n$ tales que:

$$A = \begin{pmatrix} \phi_1(x_1) & \phi_2(x_1) & \cdots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \cdots & \phi_n(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(x_m) & \phi_2(x_m) & \cdots & \phi_n(x_m) \end{pmatrix} b = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} x = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

Se puede desarrollar el problema de cuadrados mínimos lineales como se encuentra definido en la siguiente ecuación.

$$min_{x \in \mathcal{R}^n} \|Ax - b\|_2^2$$

2.2. Análisis de componentes principales (PCA)

El análisis de componentes principales es utilizado para transformar el espacio de variables que presentan cierto nivel de correlación entre ellas a un espacio donde se definen la misma cantidad de variables pero completamente independientes entre sí. A estas variables se las conoce como "componentes". Estas variables generadas se ordenan según el grado de varianza del sistema que explican. La traza de la matriz generada, que sería la sumatoria de las varianzas de cada componente, representa la varianza total del sistema, debido a que se tratan de variables independientes.

Teniendo *n* muestras de *m* variables, se construye la matriz $X \in \mathbb{R}^{n \times m}$, donde cada muestra corresponde a una fila y tiene media cero $x_i = \frac{(x_i - \mu)}{\sqrt{n-1}}$, donde μ es el vector que contiene la media de cada una de las variables (columnas).

A partir de esta matriz, se construye la matriz de covarianzas como $\frac{X^t X}{n-1}$ y, además, la matriz *V* que tiene como columnas los autovectores de la matriz de covarianzas. Se realiza un cambio de base definido como $V^t X^t$, que reduce la redundancia entre las variables y asigna a cada muestra un nuevo nombre mediante un cambio de coordenadas. Luego del cambio de base, es posible seleccionar un subconjunto de estos componentes permitiendo simplificar los cálculos utilizando un menor número de variables que expliquen gran parte de la variabilidad del sistema. Esto permite, sobre todo para casos con *m* grandes, reducir el *overhead* de los cálculos de predicción.

3. Desarrollo

Para lograr definir un modelo que permita realizar las predicciones propuestas, es necesario indagar sobre varios métodos y procedimientos en profundidad. Se debe realizar un análisis exhaustivo sobre los datos a utilizar para determinar cómo su composición puede afectar a los métodos utilizados. Por lo que es necesario estudiar más en detalle la relación entre los datos y las variables utilizadas para lograr un rendimiento más óptimo de los métodos. A continuación se introducen los métodos utilizados y la investigación previa realizada.

3.1. Métodos de resolución de cuadrados mínimos lineales

Una característica distintiva del método de cuadrados mínimos es que siempre tiene solución. Sin embargo, puede haber una única o infinitas soluciones. Esto depende de la independencia lineal entre las columnas de A, siendo esta la matriz que se utiliza en la formulación a la cual puede ser reducido del problema de cuadrados mínimos.

$$\min_{x \in \mathcal{R}^n} \|Ax - b\|_2^2$$

3.1.1. Resolución por Ecuaciones normales

Este método tiene una formulación sencilla y en la práctica en general es muy utilizado.

> *Sea* $\overline{x} \in \mathcal{R}^n$ *, si* $\overline{x}$ *es solución del problema de cuadrados mínimos entonces* $A^tA\overline{x} = A^tb$. *Si* (A^tA) *es inversible, entonces* $\overline{x} = (A^tA)^{-1}A^tb$.

Para que A^tA sea inversible, las columnas de A deben ser linealmente independientes. Sin embargo, para poder determinar esto se debería realizar un análisis sobre la matriz A, lo cual dada la magnitud del conjunto de datos de este trabajo resultaría demasiado costoso. A pesar de esto, analizando la base de datos se puede ver que tiene 240.000 entradas de datos para 22 variables inicialmente. Luego de remover los valores faltantes y agregar *features* nuevas, la matriz cuenta con aproximadamente 58.000 entradas. Esto hace que la probabilidad de que las columnas de la matriz que se seleccionen sean linealmente dependientes con estos datos sea muy baja, lo que implicaría que para dos variables distintas, deberían coincidir exactamente los 240.000 o 58.000 valores. Dado este análisis, se asume que la matriz de datos cumplen lo anterior, para poder aplicar este método de resolución.

Para implementar la resolución por ecuaciones normales, se utiliza el método **FIT** de la clase **LINEARREGRESSION**. Se plantea la ecuación y se la resuelve para los parámetros recibidos.

Listing 1: Resolución por ecuaciones normales

```
1  EcuacionesNormales(A, b) {
2      Matrix AtA = A.transpose() * A
3      Matrix Atb = A.transpose() * b
4
5      _x_solucion = AtA.inverse() * Atb
6  }
```

3.1.2. Resolución por descomposición QR

En general hay distintos métodos de factorización o descomposición de matrices que se utilizan de manera conveniente en los cálculos. Entre ellos, un método característico es el de factorización QR que consiste en descomponer una matriz en el producto de una matriz ortogonal y otra matriz triangular superior.

$A = QR$, donde Q es una matriz ortogonal de m por m, y R es una matriz triangular superior m por n.

La ventaja de este tipo de factorización, es que siempre puede realizarse para cualquier matriz lo que puede aprovecharse para reformular el problema de cuadrados mínimos en terminos de Q y de R de la siguiente manera[1]:

Sea $\overline{x} \in \mathcal{R}^n$, si $\overline{x}$ es solución del problema de cuadrados mínimos entonces $\overline{x} = (R)^{-1}b$.

Se utilizó la implementación de la biblioteca *Eigen* para poder realizar comparaciones entre los distintos métodos planteados en terminos de precisión y tiempos de ejecución.

3.1.3. Resolución por descomposición en valores singulares (SVD)

Otro tipo de factorización muy utilizada, es la de valores singulares o SVD. Si bien presenta un mayor número de restricciones para poder aplicarse, es de gran utilidad en distintos métodos estadísticos (PCA), teoría de control, compresión de imagenes, etc. La descomposición consiste en expresar una matriz como producto de dos matrices ortogonales y una matriz diagonal, la cual contiene en su diagonal los valores singulares de la matriz que se busca descomponer. Esta descomposición se define como:

Una SVD de A es una factorización del tipo $A = U\Sigma V^t$ con $U \in \mathcal{R}^{m\times m}$, $V \in \mathcal{R}^{n\times n}$ ortogonales y $\Sigma \in \mathcal{R}^{m\times n}$ una matriz formada con los valores singulares de A en su diagonal principal ordenados de mayor a menor.

Al igual que la descomposición QR, esta factorización se puede utilizar para implementar un método alternativo de resolución al problema de cuadrados mínimos, pudiendo reformularse de la siguiente manera:

Sea $\overline{x} \in \mathcal{R}^n$, si $\overline{x}$ es solución del problema de cuadrados mínimos, entonces $\overline{x} = V * \overline{y}$, donde $\overline{y}_i = \frac{(U^t b)_i}{\sigma_i}$ para todo $i = 1, ..., r$ y cualquier valor para $i = r + 1, ..., n$

Al igual que el método de resolución por descomposición QR, se utilizó la implementación de la biblioteca *Eigen* para poder realizar comparaciones.

3.2. Funciones de pérdida

El problema que se presenta en este trabajo requiere utilizar un modelo de regresión para poder predecir ciertas variables sobre un conjunto de datos dado. Pero, ¿cómo podemos estar seguros de que este modelo dará el mejor resultado posible? ¿Hay algún tipo de técnica que podamos utilizar para evaluar el desempeño del modelo utilizado? Para dar respuesta a estas preguntas se utilizan las funciones de pérdida que además de proporcionar una representación estática de cómo se está desempeñando su modelo, también nos indican cuánto se ajusta a los datos.

Las funciones de pérdida pueden clasificarse ampliamente en 2 tipos: *Clasification* & *Regression loss*. En este trabajo se utilizan las funciones de *Regression loss*.

3.2.1. Mean Absolute Error (MAE)

Es una función de pérdida comúnmente utilizada en modelos de regresión. MAE es la suma de las diferencias absolutas entre los valores observados y los predichos. Por lo tanto, mide la magnitud promedio de los errores en un conjunto de predicciones.

[1] Suponiendo que las columnas de A son linealmente independientes.

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

Suele ser útil si los datos de entrenamiento se corrompen con valores atípicos, es decir si por error se reciben valores negativos o positivos demasiado grandes en el entorno de entrenamiento, pero no en el de prueba.

Si se trata de minimizar el MAE, esa predicción sería la mediana de todas las observaciones. Esto se debe a que la mediana es más robusta para los valores atípicos que la media, lo que en consecuencia hace que el MAE sea más robusto para los valores atípicos que *Mean Square Error*.

3.2.2. Mean Square Error (MSE)

Mean Square Error es la función de *Regression loss* más utilizada. Se define como la suma de las distancias al cuadrado entre los valores observados y los predichos.

$$MSE = \sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}$$

Cabe destacar ciertas propiedades que cumple esta función de pérdida.

- Penaliza al modelo por cometer grandes errores al elevarlos al cuadrado. Esto podría ser beneficioso cuando se desea entrenar al modelo donde no hay predicciones de valores atípicos con errores muy grandes.
- Si hay *outliers* muy grandes en un conjunto de datos, pueden ser afectado drásticamente. En ese sentido, el MSE no es "robusto" para *outliers*.
- La pérdida de MSE es estable. En el caso de la función de pérdida de MSE, si se introduce una perturbación de $\delta << 1$, entonces la salida será perturbada por un orden de $\delta^2 <<< 1$.

3.2.3. Root Mean Square Error (RMSE)

Root Mean Square Error es una de las métricas de evaluación más popular utilizada en problemas de regresión. Sigue el supuesto de que el error es imparcial y sigue una distribución normal. Esta función de pérdida se define como:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}}$$

Estos son los puntos clave a considerar en RMSE:

- La "raíz cuadrada" permite que esta métrica muestre grandes desviaciones numéricas.
- Elevar el error al cuadrado ayuda a ofrecer resultados más sólidos, lo que evita la cancelación de los valores de error positivo y negativo. En otras palabras, esta métrica muestra acertadamente la magnitud plausible del término de error.
- Cuando cuenta con más muestras, se considera que la reconstrucción de la distribución de errores usando RMSE es más confiable.
- RMSE, al igual que MSE, se ve muy afectado por valores atípicos.
- En comparación con MAE, RMSE proporciona un mayor peso y castiga los errores grandes.

3.2.4. Root Mean Square Log Error (RMSLE)

Para esta función de pérdida, al igual que en las anteriores, se utiliza el registro de las predicciones y los valores reales. RMSLE se usa generalmente cuando no se quiere penalizar grandes diferencias entre los valores predichos y reales. Se define de la siguiente manera:

$$\text{RMSLE} = \sqrt{\sum_{i=1}^{n} \frac{(\log(y_i + 1)) - (\log(\hat{y_i} + 1))^2}{n}}$$

Para esta función de perdida, hay que destacar ciertas propiedades:

- Tiene en cuenta los valores positivos y negativos.
- RMSLE, al contrario que RMSE, penaliza más los errores menores. Por lo general se utiliza cuando no se desea que influya en los resultados si hay grandes errores.
- Cuando los valores reales y predichos son bajos, RMSLE y RMSE son iguales.
- Por otro lado, cuando algún valor predicho o real es alto. RMSE >RMSLE

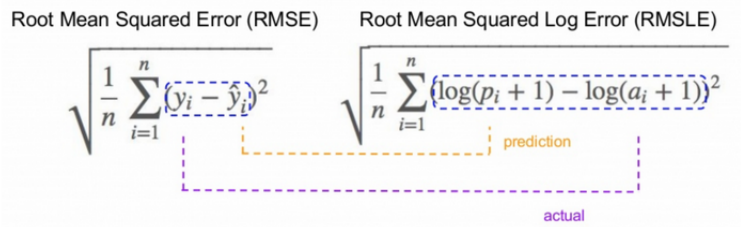

Figura 1: Comparación gráfica entre las formilas de RSME y RSMLE.

3.2.5. R2 - Score

R2-Score es el valor que muestra cuán bien se ajusta un modelo a sus datos de entrenamiento. Sin embargo, hay una diferencia entre ajuste y ajuste óptimo. Cuando se trata de la eficiencia de predictibilidad de un modelo, **R2-Score** se torna inválido debido a que es una medida de cuán bien se ajustan los datos de entrenamiento al modelo.

Primero es necesario definir la varianza en términos de la regresión lineal. Es una medida de hasta qué punto los valores observados difieren del promedio de los predichos, es decir, su diferencia de la media del predicho.

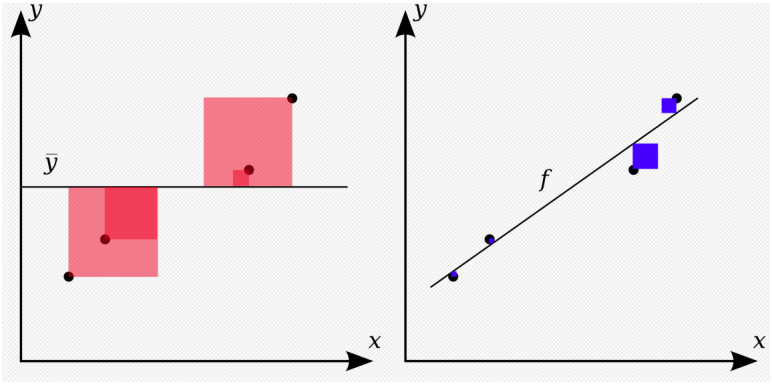

Figura 2: Cuanto mejor se ajuste la regresión lineal (derecha) a los datos en comparación con el promedio simple (izquierda), más cerca estará el valor de R^2 de la unidad. Las áreas de los cuadrados azules representan los residuos cuadrados con respecto a la regresión lineal. Las áreas de los cuadrados rojos representan los residuos al cuadrado con respecto al valor medio.

Sea $\bar{y}$ la media de un conjunto de datos observados, se puede definir ***R2-Score****, usando tres sumatorias de fórmulas cuadradas, de la siguiente manera.*

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2,$$

$$SS_{res} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2,$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R2-Score está estrechamente relacionado con el MSE. El R2 representa la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes. También puede verse como el cociente entre la varianza total explicada por el modelo y la varianza total.

Del mismo modo, tampoco hay una respuesta correcta sobre lo que R2 debería ser. 1 significa correlación perfecta. Sin embargo, hay modelos con un R2 bajo que siguen siendo buenos modelos.

Esta métrica fue utilizada para seleccionar cuál era el conjunto de variables que brindaba la mejor precisión en la predicción. Esta selección fue comparada también con la obtenida por el AIC (Akaike's Information Criterion)[2], que brindó resultados similares.

3.3. Método de K-Fold cross validation

Cross-validation es un procedimiento de remuestreo utilizado para evaluar modelos de aprendizaje automático en una muestra de datos limitada.

El procedimiento tiene un único parámetro llamado *K* que refiere al número de grupos en los que se dividirá la muestra de datos. El procedimiento a menudo se llama *K-Fold cross-validation*. Cuando se fija el valor de *K* en 10, por ejemplo, el método se determina como *10-folds cross-validation*.

Cross-validation se utiliza principalmente en aprendizaje automático aplicado para entrenar un modelo en datos no vistos. Es decir, usar una muestra limitada para estimar cómo se espera que el modelo

funcione en general, realizando las predicciones sobre datos que no se usaron durante el entrenamiento del modelo.

Es un método popular debido a su simpleza y porque generalmente da como resultado una estimación menos sesgada de la calidad del modelo que otros métodos que podrían llegar a generar *overfitting*.

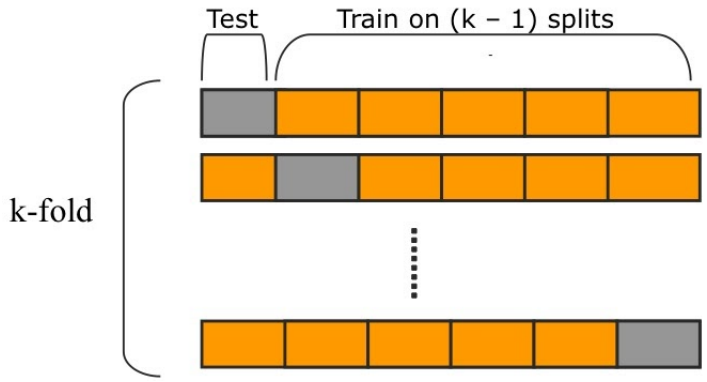

Figura 3: Esquema explicativo del método de *K-Fold cross-validation*.

El procedimiento general es el siguiente:

- Se divide el conjunto de datos en k grupos
- Se elige el grupo para el conjunto de datos de prueba
- A los grupos restantes se los utiliza como un conjunto de datos de entrenamiento.
- Se ajusta el modelo al conjunto de entrenamiento y se lo evalúa en el conjunto de prueba
- Para cada grupo se guarda el puntaje de evaluación y luego se descarta el modelo
- Se obtiene una medida de calidad del modelo utilizando la muestra de puntajes de evaluación.

Es importante destacar que cada observación en la muestra de datos se asigna a un grupo individual y permanece en ese grupo durante la ejecución del procedimiento. Esto significa que cada muestra va a ser utilizada en el conjunto de prueba una vez y se usa para entrenar el modelo k-1 veces.

3.4. Manipulación de datos

3.4.1. Remoción de *outliers*

Se denomina valor atípico o *outlier* a una observación que es numéricamente distante del resto de los datos. Dichos puntos pueden representar datos erróneos o pueden indicar una línea de regresión mal ajustada. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos pueden resultar engañosas.

Si un punto se encuentra lejos de los otros datos en la dirección horizontal, se conoce como una observación influyente. La razón de esta distinción es que estos puntos pueden tener un impacto significativo en la pendiente de la línea de regresión.

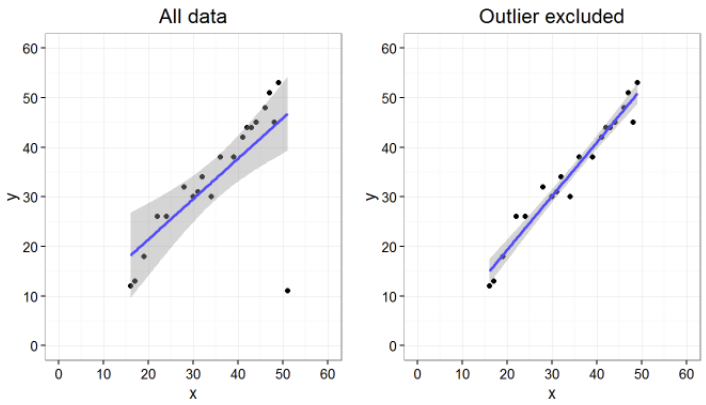

Figura 4: El comportamiento de la linea de regresión frente a la remoción de outliers.

El método de cuadrados mínimos solamente trabaja con los puntos de datos que componen una muestra en términos de su media y sus diferencias. Entonces el caso ideal sería una distribución simétrica, preferiblemente normal. Los *outliers* lográn distorsionar las predicciones realizadas por este modelo. Este tipo de métodos no funcionan óptimamente con patrones de datos demasiados complejos, sólo pueden proporcionar la mejor solución al conjunto de diferencias cuadráticas bajo una premisa de distribución normal.

Para encontrar y remover los *outliers* se utiliza la función **Z-score**, definida como el número de desviaciones típicas que un valor dado toma con respecto a la media de su muestra o población.

Sea x un valor observado, con media μ y desviación típica σ. Z-Score se define como:

$$Z = \frac{x - \mu}{\sigma}$$

Z-score permite describir cualquier punto de los datos al encontrar su relación con la desviación estándar y la media. Al calcular el **Z-score**, se centran los datos y se buscan puntos que están demasiado lejos de la media. Estos puntos serán tratados como *outliers*. En la mayoría de los casos, se usa un umbral de 3 o -3, es decir, si el valor de la puntuación Z es mayor o menor que 3 o -3 respectivamente, ese punto se identificará como *outlier*.

En este trabajo, se utilizó la biblioteca *scipy* con su módulo llamado *stats* que provee la implementación de la función matemática **Z-Score**.

3.4.2. Manejo de datos faltantes

La matrices obtenidas a partir de bases de datos de experimentos o censos, entre otros, suelen contener celdas o espacios con información faltante. En algunos casos simplemente aparece la celda vacía, en otro, completada con el valor *NaN*. A veces, según la metodología empleada para recolectar los datos, puede completarse incluso con otros valores que puedan alterar el cálculo de estimadores estadísticos. Un caso muy común es la introducción de valores como -1 o -9, que al calcular la media o mediana de una variable numérica, puede llevar a una estimación inexacta. Es por ello que el manejo de datos faltantes es un procedimiento importante y necesario previo al trabajo con datos.

En este trabajo se optó por utilizar dos técnicas de manejo de datos faltantes bastante utilizadas en la práctica. Por un lado, el proceso más simple es identificar los datos faltantes y eliminar todas las entradas que presenten al menos un dato faltante para alguna de las variables analizadas. Este procedimiento, si bien es sencillo, puede llevar a una pérdida importante de información, debido a que podemos llegar a descartar más de la mitad de las entradas originales.

Por otro lado, se puede optar por completar cada celda faltante por alguna aproximación estadística relativa a esa variable. Los ejemplos más comunes incluyen reemplazar los datos faltantes por la media o

la mediana. En este trabajo se decidió utilizar la mediana, dado que la consideramos más representativa al verse menos afectada por los valores extremos.

3.4.3. Análisis de componentes principales (PCA)

La implementación de PCA cuenta con un valor α que representa el número de componentes al que se reduce la matriz original. El método *fit* recibe la matriz de entrenamiento con la que realiza la clasificación y calcula los autovectores de la matriz para luego utilizarlos para transformar las matrices. El método *transform* es el que recibe una matriz y la transforma al nuevo espacio de variables (componentes principales). A partir del conjunto de entrenamiento, se genera la matriz de covarianzas y se utiliza el método de la potencia junto con el de la deflación para obtener los α autovectores asociados a la matriz de covarianzas en orden decreciente a su varianza. A continuación se describien los pasos del algoritmo:

- A las muestras del conjunto de datos recibido se le resta la media de cada una de las variables (columnas)
- Luego, a partir de la matriz obtenida se genera la matriz de covarianzas como $M_x = \frac{X^t X}{n-1}$
- Se aplica el método de la potencia junto con el de deflación para obtener los primeros α autovectores de M_x, es decir los componentes principales.
- Se devuelve el producto de X.V, siendo V la matriz que tiene por columnas los α autovectores de M_x.

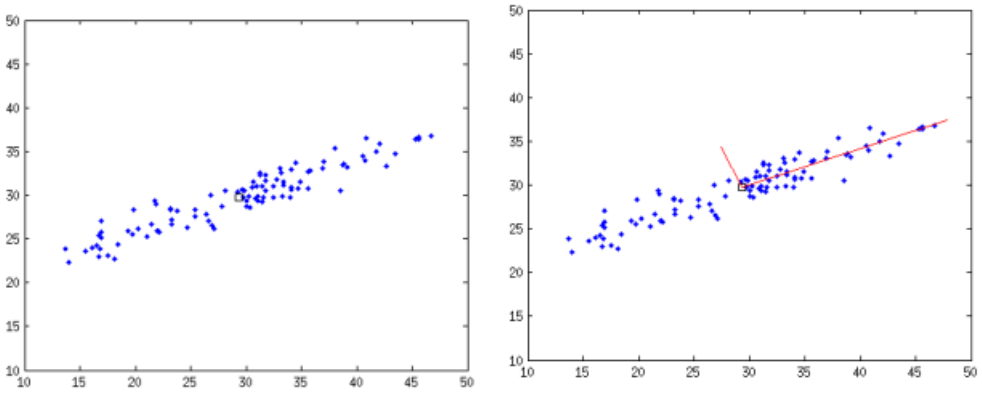

Figura 5: Figuras que muestran cómo actúa PCA en dispersión de datos en R^2. A la izquierda los datos antes de aplicar PCA y a la derecha los componentes que son tomados luego de la aplicación del método.

3.5. Exploración de datos y análisis preliminar

La exploración de datos inicial permite poder obtener información general sobre nuestra base de datos. Esto incluye el tipo de distribución aproximada, cantidad de datos faltantes, presencia de *outliers*, grado y tipo de correlación entre pares de variables, entre otros. En la mayor parte de los casos se recurren a métodos o funciones provistas por la biblioteca *numpy* dentro de *Python*.

En primer lugar, se procede a obtener información detallada de los tipos de datos contenidos en cada variable, así como el número de datos faltantes. Esto permite decidir qué rol cumplirá cada variable en el modelado, si se utilizará como variable predictora, como variable de segmentación y, además, determinar si es necesario realizar manejo de datos faltantes.

Luego, puede realizarse un análisis de los estadísticos principales de la variable a predecir, como pueden ser el rango de valores presentes, la media, mediana, desvío estándar.

Además, puede realizarse un gráfico de distribuciones y gráficos de a pares (*pairplot*). Estos gráficos consisten en un conjunto de $n \times n$ subgráficos. En la diagonal, se disponen los gráficos de distribución de cada uno de las variables representados mediante histogramas. En el resto de las posiciones, se cuenta con los gráficos de dispersión de cada par de variables, permitiendo observar el tipo de relación y el grado de correlación estimado que pueden presentar.

Luego, se realiza el cálculo de las matrices de correlación entre las variables y se grafican los *heatmaps* asociados para facilitar su interpretación gráfica.

Por último, se realiza un gráfico de *boxplot* utilizando (o no) alguna variable categórica (como puede ser *provincias* en este trabajo) para subdividir los datos. De esta manera, se puede observar la presencia de valores extremos (*outliers*), así como las diferencias entre los intervalos de confianza y las medias para cada división.

3.5.1. Segmentación de datos

A partir de la exploración inicial de los datos es posible formular algunas segmentaciones que puedan llegar a presentar resultados interesantes.

3.5.1.A. Segmentación por tipo de propiedad

A partir de los valores adoptados por el *feature* tipo de propiedad, se observó que se podía realizar una segmentación en dos grupos denominados *Casas* y *Departamento*. El primero contiene todos los tipos de propiedad que contuvieran la palabra "casa" o que hicieran referencia a viviendas de ese tipo (como puede ser el caso de "duplex"). El segundo grupo sólo contiene entradas con tipos de propiedad "apartamento".

3.5.1.B. Segmentación por tipo de explotación de la propiedad

En este caso se decidió realizar un segmentación dividiendo según la explotación de la propiedad, es decir, entre propiedades de uso "Comercial" y "Urbano". En el primer grupo se incluyen categorías como "oficinas comerciales", "terreno comercial", entre otras. En el segundo, "casa", "apartamento", "duplex", "terreno", entre otras.

3.5.1.C. Segmentación por categorías de precios

Se decidió separar en subgrupos de propiedades según rangos definidos a partir de los precios observados. Primero, se procedió a encontrar el valor máximo y mínimo de precios, definiéndose así las cotas del intervalo. Luego, a cada propiedad se le restó el valor mínimo y se procedió a dividir el intervalo $[0, max - min]$ en tres intervalos de menor tamaño. Se asignaron las etiquetas $0, 1$ y 2 según en qué intervalo cayera el precio de la propiedad.

Es interesante resaltar que se observaron muchas propiedades en la categoría de menores precios ("bajos"), mientras que la categoría más alta ("altos") presentó pocas entradas.

3.5.1.D. Segmentación por regiones geográficas

Por último, se decidió utilizar los *features* "provincia" y "ciudad" para segmentar la matriz utilizando solamente aquellas provincias que presentaran más de 5.000 entradas y aquellas ciudades que presentaran más de 3.200 entradas. De esta manera, se lograron generar 9 segmentos de provincias y 21 de ciudades (contenidas en las mismas 9 provincias).

3.5.2. *Feature engineering*

Feature engineering se refiere a un proceso de selección y transformación de variables pertenecientes a los datos. El proceso implica una combinación de análisis de datos, aplicación de reglas generales y juicio.

Los datos utilizados para crear un modelo predictivo consisten en una variable de resultado, que contiene datos que deben predecirse, y una serie de variables predictoras que contienen datos que se

consideran útiles para poder realizar una predicción. En este trabajo, las variables a predecir fueron el precio y los metros cubiertos, mientras que algunas variables utilizadas para predecir fueron metros totales, número de habitaciones y antigüedad, entre otras.

Una "característica" es solo otro nombre para una variable predictiva. *Feature engineering* es el término general para crear y manipular predictores que permita generar nuevas características con el fin de mejorar el nivel de predictibilidad del modelo.

Hay distintas maneras posibles de hacer *Feature engineering*. En el presente trabajo se recurrió a la extracción de información de los campos de texto, como la *descripción* de la publicación de la propiedad, y incorporación de características de fuentes externas.

Exploración de las descripciones - Se realizó una observación de un conjunto aleatorio de *descripciones* de algunas entradas, lo que permitió evidenciar que ciertas palabras que se consideraron potencialmente relevantes se repetían con mucha frecuencia indicando otro tipo de característica sobre la vivienda. A modo de ejemplo, se muestra una descripción perteneciente a la base de datos.

> *casa nueva, en fraccionamiento con **vigilancia, jardines** comunales con excelente mantenimiento, salón para eventos con capacidad de 50 personas con baño de hombres y mujeres.*

Como se puede ver, las palabras "jardines" y "vigilancia" resaltadas pueden denotar dos características como **seguridad** y **jardín** que no están dentro de las variables predictoras iniciales. Se utilizaron sinónimos de conceptos a englobar, por ejemplo para el caso de la característica **jardín** se utilizó "garden" y "parque" para lograr captar todas las potenciales referencias. De esta manera, se agregaron características nuevas como **seguridad**, **luminoso** y **terraza** representadas como valores booleanos (presente o ausente). A pesar de lo rústico de este proceso, para la creación de nuevas características, resultó que no solo las apariciones eran significativas sino que la correlación con las otras variables era en algunos casos mayor que características preexistentes.

Datos de inseguridad - Por otro lado se decidió agregar información estadística relativa a la inseguridad, para determinar si podía presentar algún efecto sobre las predicciones de precios de las propiedades. Se logró encontrar dos tipos de características sobre la inseguridad referidas a cada provincia, la **percepción de inseguridad** y las **infracciones totales**[2].

En el caso de la característica de infracciones totales se realizó una normalización previa de los datos.

Sea I_p el número de infracciones cometidas en una provincia p, Tm_p la tasa media de crecimiento poblacional para cada provincia tomada en el periodo 2010 - 2015, $P_{tot2010}$ población total de México en el 2010 para cada provincia y $y \in \{2012, 2013, 2014, 2015, 2016\}$ el año donde se tomó el muestreo de las infracciones. El índice está definido como:

$$infracciones\_totales = \frac{I_p * 1000}{P_{tot2010} * Tm_p * (y - 2010)}$$

En resumen nos estaría indicando la cantidad de infracciones cometidas en una provincia cada 1000 habitantes.

En el caso de la **percepción de inseguridad** no fue necesario ningún procesamiento previo de los datos.

3.5.3. Forward selection

Al ver la correlación entre los datos pertenecientes a la base de datos surge el interrogante de cuál sería la mejor manera de seleccionar las variables que permita obtener un modelo más óptimo para predecir. Utilizando **R2-Score** para evaluar cuánto se ajusta el modelo a los datos, se realiza el siguiente procedimiento.

[2]Los datos fueron extraídos de [3]

Paso N° 1 : Obtener el *feature* más correlacionado con la variable que se intenta predecir. Se define un puntaje de R2 inicial igual a 0, para poder representar el mínimo.

Paso N° 2 : Luego se itera sobre las variables restantes y se busca seleccionar las variables que presenten el menor valor de R2. Para ello se realiza *cross validation* y se obtiene la variable que haya sido seleccionada como *la mejor* una mayor cantidad de veces y se obtiene un promedio de los valores de R2 que presentó en las sucesivas iteraciones de esa variable.

Paso N° 3 : Luego se observa cómo modifica el R2 agregar esa variable. Se compara con el valor previo a agregarla y se utiliza cierta tolerancia que denominaremos $\mathcal{E}$ para decidir si la diferencia presentada amerita adicionar la variable al modelo.

Paso N° 4 : En caso de encontrar que agregar la variable no modifica sustancialmente la métrica del modelo, se detiene el proceso y se devuelve el modelo formado hasta el momento con su respectiva métrica.

3.5.4. Principal component regression (PCR)

PCR se utiliza para estimar los coeficientes de regresión desconocidos en un modelo de regresión lineal estándar. Los componentes principales de las variables explicativas se usan como *regressors*[3]. Por lo general, se usa solo un subconjunto de todos los componentes principales para la regresión, lo que hace que PCR sea un procedimiento regularizado.

Al realizar la regresión solo en un subconjunto de todos los componentes principales, PCR puede dar como resultado una reducción de la dimensión al reducir sustancialmente el número efectivo de parámetros que caracterizan el modelo subyacente. Se utilizará este método para lograr una predicción eficiente del resultado según el modelo asumido.

El método de PCR puede separarse en tres pasos:

Paso N° 1 : Realizar PCA en la matriz de datos observada para las variables explicativas. Se obtienen las componentes principales y luego se selecciona un subconjunto, basado en algunos criterios apropiados.

Paso N° 2 : Utilizando el método de cuadrados mínimos se obtiene un vector de coeficientes de regresión estimados, los cuales tienen una dimensión igual al número de componentes principales seleccionados.

Paso N° 3 : Se transforma el vector de nuevo a la escala de las covariables reales, usando los vectores propios correspondientes a los componentes principales seleccionados para obtener el estimador de PCR final que se utiliza para estimar los coeficientes de regresión caracterizando el modelo original.

3.5.5. Comparación de distintos métodos de resolución

Se realizó un análisis cuantitativo del desempeño de los distintos métodos propuestos para resolver el problema de cuadrados mínimos lineales, con la intención de poder obtener algún tipo de indicio a la hora de elegir un método sobre otro.

Se aprovecharon las implementaciones de la biblioteca EIGEN sobre los métodos de resolución por **descomposición por valores singulares (SVD)**, **descomposición QR** y **ecuaciones normales (EN)**. Para evaluar cuál sería más conveniente se analizaron los tiempos de ejecución y las mediciones para las métricas MAE, MSE, RMSE y RMSLE de cada método aplicando regresión sobre una segmentación en particular.

Se observaron los siguientes resultados:

[3]La variable dependiente cuya variación se está estudiando, alterando las entradas.

- Para el caso de los tiempos de ejecución, el método de resolución por **ecuaciones normales** fue sin duda el mejor ya que mostró un tiempo de por lo menos 5 veces menor que el resto.
- Para el caso de las medidas tomadas para cada métrica de error, ninguno de los métodos de resolución mostró una diferencia significativa respecto al resto.

Esto permitió observar que el método de **ecuaciones normales** resulta ser el más óptimo en términos generales.

3.6. Hipótesis

3.6.1. Manejo de *outliers* y datos faltantes

> ***Hipótesis 1.*** *Remover los outliers y los valores faltantes o NaN proporcionará una mejor precisión en las predicciones.*

Como fue visto en las secciones anteriores, se plantearon distintas soluciones para intentar dilucidar de qué manera se podrían aplicar para lograr un modelo más preciso. Se consideró que como al remover los datos que contuvieran celdas vacías aún se contaba con una gran cantidad de datos para realizar las predicciones, esto no afectaría en gran medida en la predicción. Es más, se cree que removerlos y evitar agregar datos falsos podría generar ruido.

Para el caso de la remoción de *outliers*, dado que la regresión lineal es muy susceptible a estos datos por lo visto anteriormente, es razonable asumir que remover estos datos patológicos del conjunto de datos incrementará la precisión en la predicción.

Para poder ver el incremento en el rendimiento del modelo se decidió evaluarlos con distintas funciones de pérdida. Si bien es claro que cada función tiene sus ventajas y desventajas, resultó interesante observar cómo afectan en cada caso.

Para cada función de perdida consideremos lo siguiente:

* **MAE** : Es más robusta para los *outliers*, por lo tanto no presentará grandes alteraciones al quitar o preservarlos. Sí lo será, en cambio, ante la falta de información por quitar los *NaN*.
* **MSE** : en contraposición con el MAE, esta métrica es muy susceptible a los *outliers* presentes en los datos. Por lo tanto, se espera que demuestre una modificación con respecto a la presencia de valores atípicos.
* **RMSE** : Al igual que MSE, esta métrica suele verse afectada por la presencia de *outliers*. Se espera que con el agregado de la mediana para los datos faltantes resulte más precisa.
* **RMSLE** : Esta métrica resulta más robusta ante la presencia de grandes errores, como lo son los *outliers*.

3.6.2. Selección del mejor modelo

> ***Hipótesis 2.*** *Una selección más precisa como la utilizada en* ***Forward selection****, permitirá obtener un mejor modelo para la predicción. Sin embargo, si no se pudieran obtener los datos que mejor correlacionan las variables del conjunto de datos con la dependiente (quizás por presentar un alto costo computacional),* ***PCR*** *sería el método más óptimo.*

La selección del conjunto de variables a utilizar para el modelo no es una cuestión "trivial". Anteriormente se propusieron dos métodos distintos para mitigar esta problemática. Para uno, si bien lleva a cabo una selección más precisa, requiere tener el grado de correlación entre cada variable, lo que en muchos casos puede no ser tan fácil de obtener. Por otro lado, si se utiliza el método que combina el

análisis de componentes principales (PCA) con el método de cuadrados mínimos, llamado **principal component regression (PCR)**, no se requerirá información adicional del conjunto de datos con el que se esta trabajando, por lo que sería una interesante posibilidad ante el desconocimiento de los mismos.

Para encontrar la mejor estrategia se compararán estos métodos mencionados anteriormente junto con un método *shuffle*, que selecciona las variables a utilizar de una manera aleatoria. Para analizar la comparación se estudiarán los resultados obtenidos con las funciones de pérdida MAE, MSE, RMSE y RMSLE.

3.6.3. Predicción de precios y metros cubiertos

Se realizará un análisis de precisión en la predicción de los **precios** utilizando las métricas, features, métodos y segmentaciones antes explicadas.

Se realizará un análisis de precisión en la predicción de los **metros cubiertos** utilizando las métricas, features, métodos y segmentaciones antes explicadas.

Considerando lo explicado respecto a los diferentes elementos a considerar a la hora de realizar predicciones con el método de regresión lineal, se desea analizar los resultados para predicciones concretas. Se evaluarán las diferencias, los casos más problemáticos y los más precisos en función de buscar los mejores métodos de predicción.

Se analizarán todos los *features* del modelo óptimo obtenido a partir del método **Forward Selection** para cada segmentación, utilizando la técnica de *cross-validation* para evitar *overfitting* al aplicar la regresión lineal.

4. Resultados y discusión

En la sección anterior se analizaron algunas preguntas que podrían surgir en el ambito teorico. Con los métodos mencionados anteriormente para afrontar la problemática, se buscará la manera más optima de combinarlos para lograr un mejor modelo de predicción. En lo que resta de este trabajo se presentarán algunas simulaciones para los escenarios considerados, distintas soluciones posibles que se tuvieron en cuenta y experimentos concretos analizando sus resultados, relacionándolos con las hipótesis ya enunciadas.

4.1. Especificaciones generales

Para contextualizar las métricas, se adjuntan a continuación las especificaciones del equipo en donde se corrieron los experimentos.

- **CPU:** IntelCore i5-6300U CPU @ 2.40GHz
- **Memoria Cache:**
  - L1I 64 KiB 2x32 KiB 8-way set associative
  - L1D 64 KiB 2x32 KiB 8-way set associative write-back
  - L2 512 KiB 2x256 KiB 4-way set associative write-back
  - L3 3 MiB 2x1.5 MiB write-back
- **Memoria Principal:** 8GB RAM DIMM DDR4 Synchronous @ 2400 MHz
- **Memoria Secundaria:** 256GB SATA SC308 Sk Hynix @

Todos los experimentos fueron corridos en *Jupyter Notebook*. Se adjunta el *script* correspondiente para que puedan replicarse fácilmente los análisis realizados, donde a su vez puede variarse los parámetros y datos de entrada fácilmente.

4.2. Análisis exploratorios

Según las segmentaciones explicadas se decidió hacer una análisis de las correlaciones que pudieran existir entre las variables.

En primer lugar se realizaron *pairplot* para poder analizar las distribuciones de los features y las relaciones que pudieran existir entre ellos. Esto permitió poder observar cuales serían más propensos a ser predichos mediante le método de regresión lineal y que otros features podrían ser utilizados para el mismo.

Figura 6: Matriz de correlación de toda la base de datos con los *features* correspondientes agregados.

Luego se observaron las **matrices de correlación** de diferentes segmentaciones, lo que permitió observar cuales features podían ser relevantes y ayudó a buscar nuevas variables para realizar *feature engineering* y así enriquecer la base de datos. Este paso fue crucial para poder concluir en la necesidad del método **Forward Selection** a la hora de elegir los features a utilizar en la predicción, puesto que fue clara la incidencia de ciertos features y la poca relación (o incluso la relación contraria) de otros.

Nota: En el apendice se adjuntan matrices de correlación según ciudades, provincias y tipos de propiedades en particular.

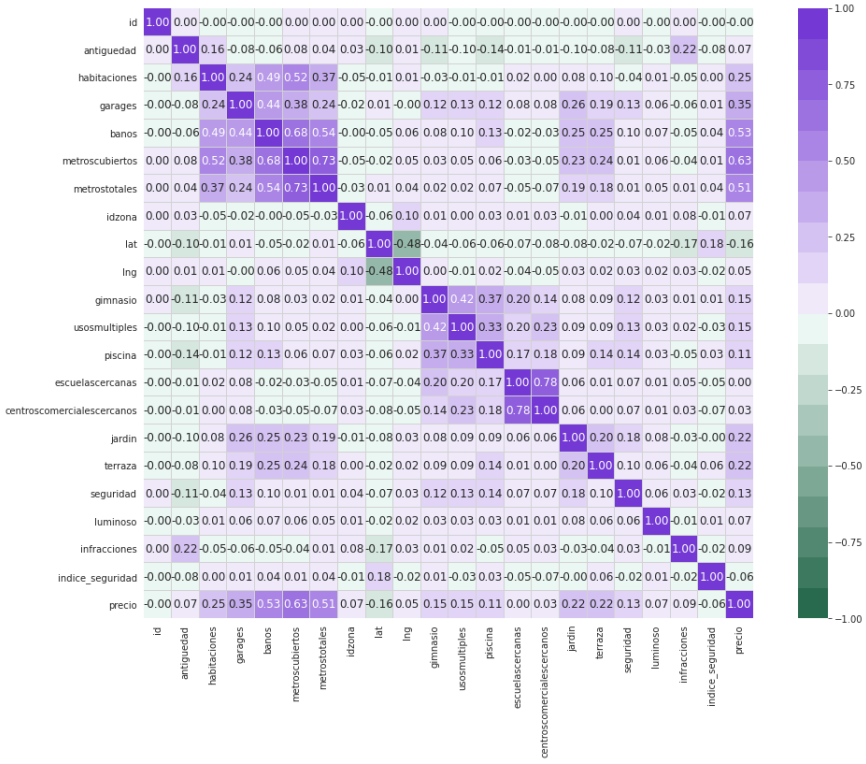

Figura 7: Matriz de correlación de toda la base de datos con los *features* correspondientes agregados.

4.3. Experimentación

4.3.1. Manejo de outliers y datos faltantes

Para realizar la experimentación se realizó una regresión lineal sobre una segmentación particular. Se seleccionaron las publicaciones de propiedades que correspondían a casas, intentando predecir el precio de los mismos. Se ejecutó el modelo de predicción con distintas modificaciones en los datos para analizar la problemática y mitigar los posibles conflictos.

- Rellenando los datos faltantes con la mediana y sin sacar outliers.
- Rellenando los datos faltantes con la mediana y sacando outliers.
- Descartando las filas incompletas y sin sacar outliers.
- Descartando las filas incompletas y sacando outliers.

Método	MSE	MAE	RMSE	RMSLE
Con outliers y rellenando NaN	1.914e+12	876277.401	1.383e+06	0.586
Con outliers y sacando NaN	9.757e+11	617325.983	9.877e+05	0.497
Sin outliers y rellenando NaN	1.135e+12	720090.823	1.065e+06	0.517
Sin outliers y sacando NaN	5.605e+11	512343.869	7.486e+05	0.426

Como se puede observar en la tabla de valores para cada ejecución, para los casos en los que se rellena o descarta datos faltantes cuando se compara los resultados obtenidos con tener y no outliers. Se puede ver para las métricas MSE y RMSE su resultado se ve reducido cuando se descartan los outliers. Por otro lado es interesante observar que RMSE para el caso en el que se remueven los datos faltantes y no se descartan los outliers, tiene un valor menor en comparación con el que no descarta los outliers pero rellena los datos faltantes con la mediana.

Sin embargo, el caso quitar los valores faltantes y los outliers, como se puede ver, representó una mejora en todas las métricas utilizadas. Esto a pesar de significar una reducción de la cantidad de información utilizada en la predicción. Puede significar que el llenado de valores faltantes por la mediana, para un caso como el nuestro, donde son demasiadas las entradas que tienen esta patología. Y el llenado si bien se utiliza para mejorar las estadísticas de los datos, en este caso podría estar sesgando los datos haciendo que una gran parte de ellos sea un valor representativo del conjunto como la mediana.

4.3.2. Selección del mejor modelo

Para analizar cual sería la mejor alternativa para seleccionar un modelo según los métodos propuestos anteriormente se implementó cada método y se ejecutó para distintas segmentaciones, donde para cada ejecución fueran evaluadas las métricas MAE, MSE, RMSE y RMSLE.

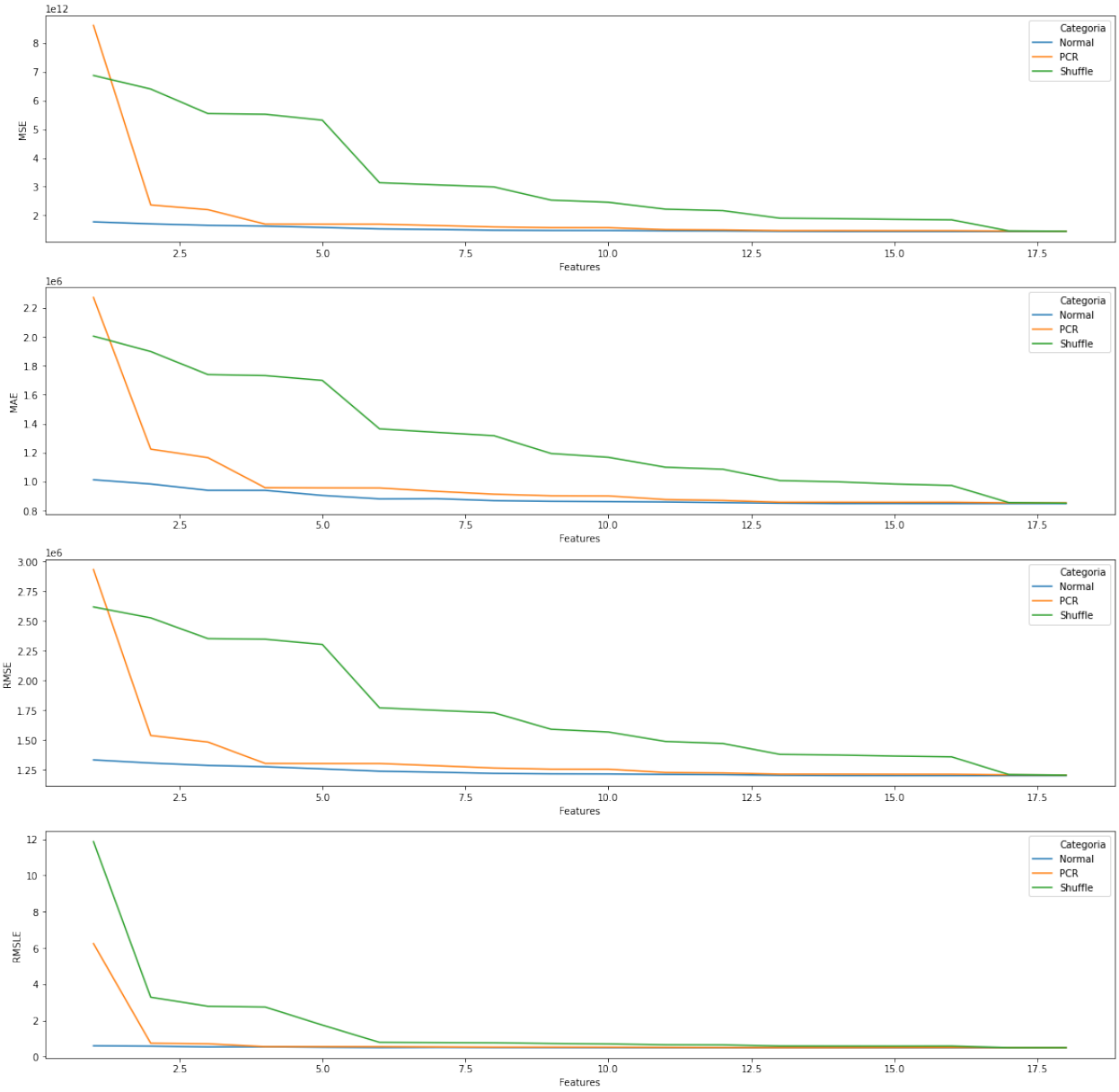

Figura 8: Análisis de las métricas para cada implementación de selección de Features en una predicción de precio para departamentos de todo México.

Como se puede ver en la figura, los resultados para el método **Forward Selection** fueron ampliamente mejores en todo momento (en la Figura 11 del apéndice se puede observar con mayor detalle lo que

sucede para valores grandes de feature y corroborar esto). A medida que se fueron agregando variables explicativas al modelo su error se redujo cada vez más (sobre el final algunos que estaban inversamente relacionados empeoraron levemente todas las métricas). Lo interesante de este método es que su error inicial, gracias a la selección de la variable más correlacionada con respecto a la que se intenta predecir, es considerablemente menor que el de los otros métodos. Esto demuestra que si bien aplicar este método requiere una mayor cantidad de información sobre el conjunto de datos con el que se está trabajando, se puede obtener resultados realmente buenos.

Por otro lado, respecto al método **PCR** se puede observar claramente como cada vez que se agrega una variable explicativa se corrige su error de predicción drásticamente, aunque el error inicial del modelo no sea bueno. De esta manera se logra un resultado excepcionalmente bueno considerando que el método no requiere ningún tipo de información previa o asunción sobre el conjunto de datos con el que se está trabajando. Sin embargo, en ningún momento el error de este método se vuelve menor que el de **Forward selection**, por lo que a lo largo del trabajo se utilizó el método de regresión lineal con **Forward Selection**.

El método de **Shuffle** fue el peor en todo momento. Sin embargo, se puede observar como a medida que se agregan características al modelo se genera una reducción del error, lo que tiene sentido considerando que se fueron agregando variables que pudieran estar más correlacionadas con el precio que las iniciales. Como era de suponerse, este método no suele ser una buena estrategia para seleccionar el mejor modelo de predicción.

Se pudo entonces concluir que la elección del método depende significativamente del contexto de trabajo. Si la extracción de los datos de correlación entre las variables explicativas que se van a utilizar en el modelo de predicción son faciles de obtener, la estrategia ganadora sería **Forward Selection**. Mientras que en un contexto donde no se puede conseguir fácilmente esta información extra sobre el conjunto de datos, se puede aplicar el método de PCR sin una perdida demasiado significativa en el desempeño de la predicción.

4.3.3. Predicción de Precios

Segmentación por tipos de propiedades:

Considerando todo lo dicho anteriormente se procedió a realizar la predicción de los precios utilizando los diferentes *features*, segmentaciones y métodos.

En primer lugar se hizo un análisis utilizando la segmentación por **categorías** y por **tipo de propiedad**. Aquí nos encontramos con los siguientes resultados:

Tipo de propiedad	MSE	MAE	RMSE	RMSLE
Urbano	8.084e+11	6.157e+05	8.989e+05	0.502
Comercial	9.676e+11	7.440e+05	9.755e+05	0.557
Casa	6.356e+11	5.353e+05	7.971e+05	0.459
Depto	1.164e+12	7.669e+05	1.079e+06	0.513
Bajos	3.979e+11	4.631e+05	6.307e+05	0.426
Medios	1.113e+12	8.777e+05	1.055e+06	0.179
Altos	7.650e+13	7.847e+06	8.738e+06	14.108

Tabla que representa con mayor precisión la evaluación de la predicción realizada sobre precio según una segmentación por tipos de propiedad, para cada métrica.

Como se puede observar, las distintas métricas fueron variando no sólo según que tan buena fue la segmentación sino que también se vieron afectadas por la cantidad de entradas que cada una tenía. En el caso de *urbano, depto, casa, bajos y medio* se tenía una buena y estable cantidad de elementos, mientras que se contaba con pocas entradas en las categorías *comercial, alta (que incluso no fue agregada por su gran distancia respecto al resto)*.

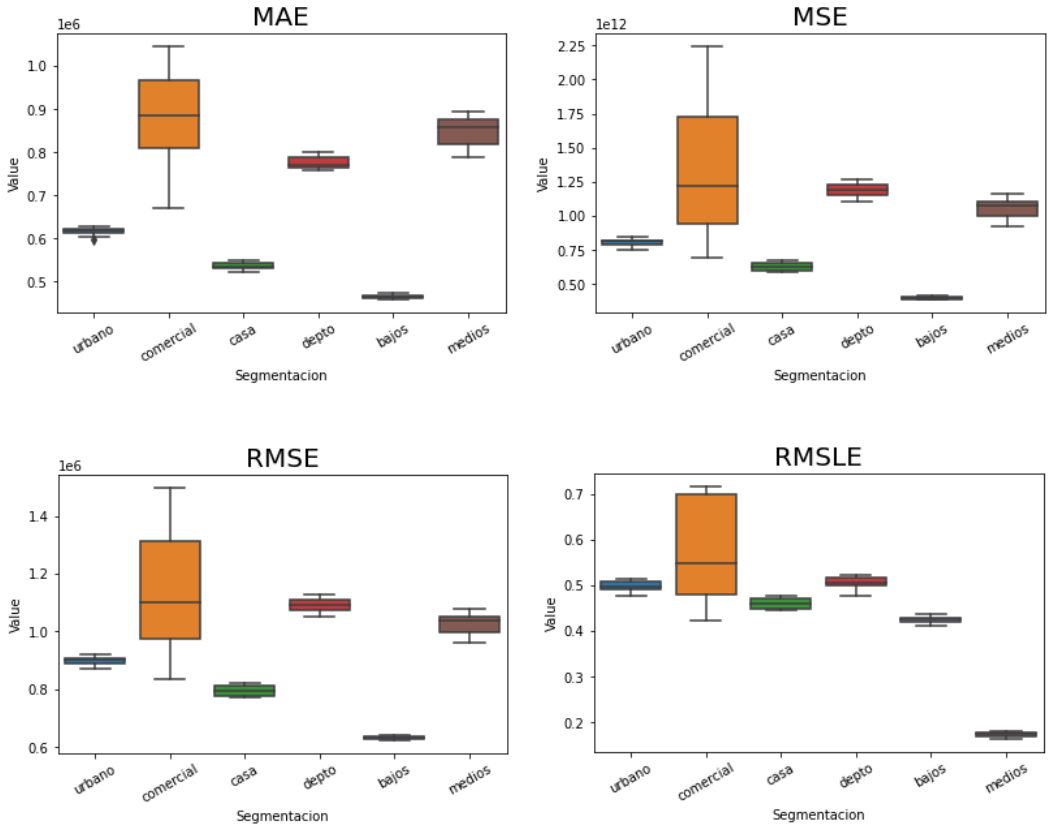

Figura 9: Boxplots que expresan, para cada métrica, las evaluaciones realizadas sobre la predicción de precio según una segmentación por distintos tipos de propiedad.

Segmentación por provincias y ciudades:

Luego se procedió con la segmentación por **provincias** y **ciudades** *los boxplots correspondientes a se encuentran en la Figura 17 del apéndice*. Aquí se pudo observar una considerable mejoría en las métricas en general. Analizando más en detalle se pudo observar que para ciertas ciudades y/o provincias que presentaban mejores métricas, los tipos de propiedades se reducían, por lo general, a un único tipo, lo que mejoró aún más la segmentación y, como consecuencia, sus métricas.

Provincia	MSE	MAE	RMSE	RMSLE
Distrito Federal	1.879e+12	956574.171	1.370e+06	0.441
Jalisco	4.619e+11	485031.164	6.794e+05	0.406
San luis Potosí	2.257e+11	331088.748	4.743e+05	0.486
Querétaro	1.275e+11	258819.416	3.568e+05	0.215
Edo. de México	9.032e+11	654438.989	9.496e+05	0.384
Nuevo León	4.078e+11	500102.969	6.364e+05	0.430
Puebla	2.336e+11	338385.640	4.824e+05	0.359
Yucatán	1.768e+11	311847.292	4.175e+05	0.323
Morelos	4.286e+11	467874.106	6.525e+05	0.295

Tabla que representa con mayor precisión la evaluación de la predicción realizada sobre precio según una segmentación por provincia, para cada métrica.

Se puede observar claramente como en el caso de Querétaro, por ejemplo, al encontrarse prácticamente puras casas, se obtuvo un mejor desempeño en todas las métricas. En Distrito Federal, en cambio,

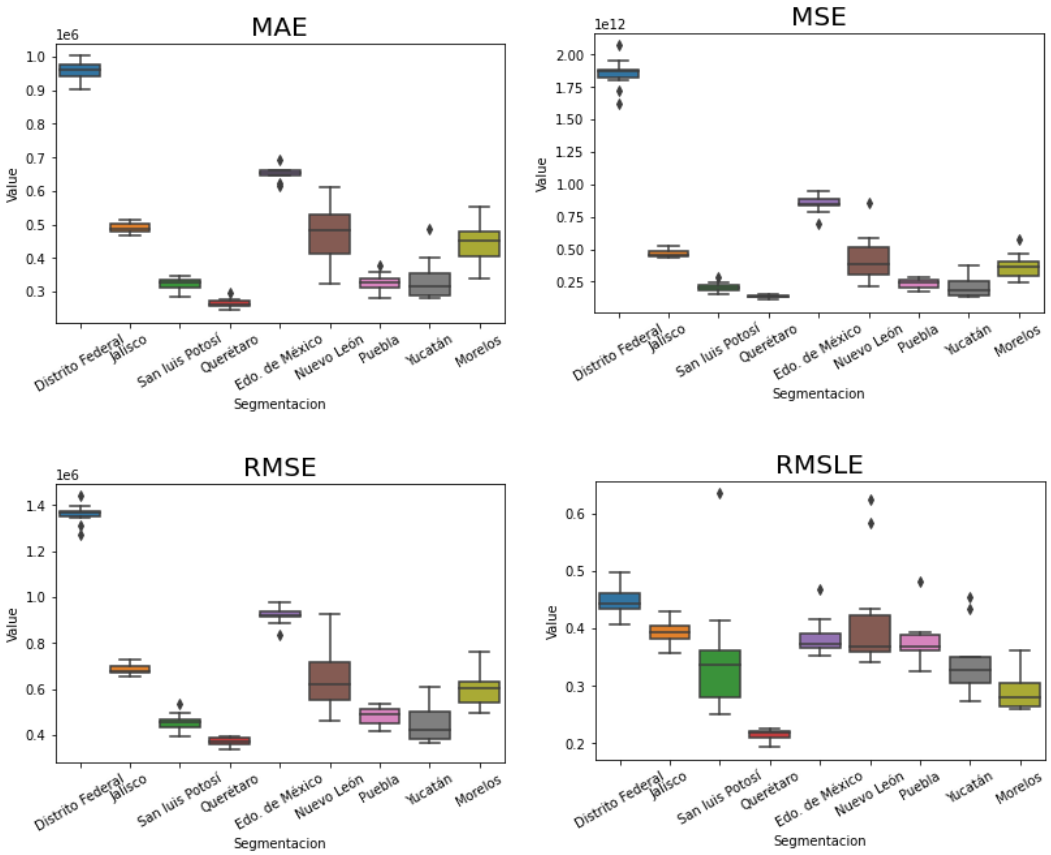

Figura 10: Los boxplots expresan las evaluaciones realizadas sobre la predicción de precio según una segmentación por provincias, para cada métricas.

al encontrarse una masa de tipos de propiedades mucho más heterogéneas, se observó un peor desempeño. Notar que estu sucedió independientemente de la cantidad de entradas, puesto que Distrito Federal era la que más tenía.

Ciudad	MSE	MAE	RMSE	RMSLE
Benito Juárez	1.119e+12	7.611e+05	1.055e+06	0.373
Zapopan	4.800e+11	4.876e+05	6.920e+05	0.366
Coyoacán	1.329e+12	8.177e+05	1.149e+06	0.389
San Luis Potosí	2.257e+11	3.311e+05	4.743e+05	0.486
Querétaro	1.325e+11	2.646e+05	3.636e+05	0.217
Naucalpan de Juárez	1.028e+12	7.560e+05	1.013e+06	0.312
Monterrey	4.078e+11	5.001e+05	6.364e+05	0.430
Puebla	1.683e+11	2.935e+05	4.079e+05	0.305
Miguel Hidalgo	3.311e+12	1.365e+06	1.817e+06	0.537
Mérida	1.768e+11	3.118e+05	4.175e+05	0.323
Huixquilucan	2.538e+12	1.194e+06	1.580e+06	0.383
Atizapán de Zaragoza	7.297e+11	6.013e+05	8.427e+05	0.357
Tlalpan	1.168e+12	7.661e+05	1.079e+06	0.376
Cuautitlán Izcalli	1.588e+11	2.545e+05	3.937e+05	0.319
Cuauhtémoc	2.042e+12	1.032e+06	1.420e+06	0.562
Alvaro Obregón	3.473e+12	1.372e+06	1.860e+06	0.520
Tlajomulco de Zúñiga	1.271e+11	2.474e+05	3.548e+05	0.356
San Andrés Cholula	2.462e+11	3.663e+05	4.924e+05	0.225
Cuernavaca	4.286e+11	4.679e+05	6.525e+05	0.295
Tlalnepantla de Baz	4.476e+11	5.083e+05	6.680e+05	0.347
Guadalajara	8.099e+11	6.777e+05	8.993e+05	0.447

Tabla que representa la evaluación de la predicción realizada sobre precio según una segmentación por ciudad, para cada métrica.

En el caso de las ciudades pudimos encontrar resultados similares a los provinciales. Aquí se reprodujo la lógica detallada anteriormente puesto que las ciudades con mayor cantidad de propiedades que correspondieran al mismo tipo de categoría de propiedad, presentaron mejores mejores resultados en las métricas de presición.

4.3.4. Predicción de metros cubiertos

Metros cubiertos según tipo de propiedad

Tipo de propiedad	MSE	MAE	RMSE	RMSLE
Urbano	1749.166	29.984	41.821	0.264
Comercial	10814.673	78.668	101.904	0.457
Casa	1890.567	31.426	43.478	0.257
Depto	265.887	7.888	16.274	0.137
Bajos	1518.333	28.080	38.964	0.262
Medios	3091.868	43.463	55.598	0.234
Altos	2363.245	38.087	48.496	0.188

Tabla que representa con mayor precisión la evaluación de la predicción realizada sobre metros cubiertos según una segmentación por tipos de propiedad, para cada métrica.

En líneas generales se observan buenos resultados en las distintas métricas, exceptuando el caso de la partición comercial. Las demás categorías presentan entradas más homogéneas en sus características o un mayor número de entradas (por ejemplo, en casas o bajos) que permite amalgamar las diferencias. Es notable el caso de la partición de departamentos que presenta los mejores resultados. Esto puede deberse

a que es la única partición que cuenta con un sólo tipo de propiedad, así como también suelen ser propiedades bastante similares entre sí (salvo casos extremos de departamentos de lujo). La partición comercial cuenta con dos desventajas, una variedad de propiedades que puede ir desde terrenos hasta oficinas comerciales y un menor número de representantes, lo que hace que las diferencias que puedan presentarse entre las propiedades contenidas no permitan generar un modelo lo suficientemente homogéneo.

Metros cubiertos según provincia

La siguiente tabla muestra la evaluación de desempeño de la predicción realizada de metros cubiertos según cada provincia, para distintas métricas.

Segmentacion	MSE	MAE	RMSE	RMSLE
Distrito Federal	1967.806	30.789	44.304	0.274
Jalisco	1249.736	24.912	35.328	0.230
San luis Potosí	1335.610	27.085	36.422	0.221
Querétaro	1029.574	23.319	31.983	0.182
Edo. de México	2242.000	35.189	47.323	0.254
Nuevo León	2835.882	40.105	53.127	0.274
Puebla	1671.114	30.123	40.527	0.242
Yucatán	1240.898	27.530	35.085	0.226
Morelos	2245.721	33.365	47.010	0.239

Como se puede observar las mediciones fueron variando en un rango no tan amplio para cada provincia. Dado que la tabla está en orden de la cantidad de entradas que posee cada provincia, se podría presuponer que los valores de las mediciones para Distrito Federal deberían ser mayores que los de Morelos sin embargo esto no es así. Se observa que todas las métricas tienen su menor valor en Querétaro, la cual posee una cantidad promedio de datos. Sin embargo cabe destacar que posee una mayor cantidad de entradas de tipo casa que las otras entradas, es decir tiene una distribución mas homogénea de los datos lo cual aporta un mejor desempeño para el modelo. Por otro lado una observación interesante a realizar es sobre las mediciones tomadas para RMSLE, la cual como vimos es una métrica muy robusta dada la ponderación por aproximación. Esta métrica alcanza su peor valor para provincias donde la distribución de los datos según el tipo de propiedad son mas heterogéneas logrando un peor desempeño del modelo. Esto nos hace señalar como impacta la calidad de los datos sobre la cantidad a la hora de evaluar un modelo. Dos características cruciales para tener en cuenta sobre el conjunto de datos sobre el que se está trabajando.

Metros cubiertos según ciudad

La siguiente tabla muestra la evaluación de desempeño de la predicción realizada de metros cubiertos según cada tipo de propiedad, para distintas métricas.

Segmentacion	MSE	MAE	RMSE	RMSLE
Benito Juárez	1025.968	20.899	31.494	0.219
Zapopan	1084.087	23.219	32.887	0.217
Coyoacán	3557.810	40.912	59.223	0.279
San Luis Potosí	1335.610	27.085	36.422	0.221
Querétaro	1029.574	23.319	31.982	0.182
Naucalpan de Juárez	2655.090	40.022	51.515	0.253
Monterrey	2835.882	40.105	53.127	0.274
Puebla	1470.340	26.626	37.629	0.238
Miguel Hidalgo	640.885	12.044	23.864	0.157
Mérida	1240.898	27.530	35.085	0.226
Huixquilucan	2024.534	31.586	44.865	0.197
Atizapán de Zaragoza	2034.778	33.247	44.954	0.246
Tlalpan	2915.666	40.851	53.862	0.293
Cuautitlán Izcalli	898.689	19.624	29.844	0.235
Cuauhtémoc	780.603	15.999	27.863	0.209
Alvaro Obregón	2957.843	41.889	54.322	0.296
Tlajomulco de Zúñiga	441.454	15.257	20.499	0.168
San Andrés Cholula	1496.561	30.532	38.498	0.203
Cuernavaca	2245.721	33.365	47.010	0.239
Tlalnepantla de Baz	2443.033	36.210	49.262	0.270
Guadalajara	2970.688	41.192	54.254	0.308

Finalmente, al observar los valores de error obtenidos en los modelos por ciudad, se observan en general valores similares a los presentados en la división por provincias. Nuevamente, las ciudades están ordenadas por cantidad de propiedades presentes en la partición. Parece observarse un aumento del error a medida que disminuye la cantidad de propiedades por ciudad con excepciones notables como Tlajomulco de Zúñiga que presenta arriba de un 90 % de propiedades de tipo casa, mientras que por el otro lado Coyoacán, pese a tener un gran número de entradas presenta una distribución aproximada de 50-50 entre departamentos y casas. Lo que lleva a observar que quizás es más importante la composición por tipo de propiedades de las distintas divisiones en los casos de divisiones geográficas, es decir, quizás hubiera sido más productivo subdividir dentro de provincias por tipos de propiedad para poder obtener un mejor modelo.

5. Conclusiones y trabajo futuro

A partir de los resultados obtenidos en la comparación de estrategias para tratar con los *outliers* y la falta de datos, se puede observar que esta puede tomarse según las características particulares del conjunto de datos con el que se está trabajando. Por ejemplo, sería interesante analizar el impacto de rellenar los datos faltantes con la mediana para un conjunto de datos con relativamente pocos datos faltantes. A su vez, sería interesante analizar qué estrategia sería conveniente para una base con pocos datos, puesto que la quita aquí podría generar cambios drásticos.

En cuanto los métodos analizados para seleccionar un modelo óptimo de predicción se encontró que métodos como **PCR**, que no requieren de datos adicionales, suelen ser muy prácticos y precisos cuando no se conoce mucho de la base o resulta complejo analizar las variables a utilizar. Sin embargo, si lo que buscamos es el mejor modelo posible para nuestro conjunto de datos y podemos obtener la información necesaria sobre los datos que se requiere, métodos como el **Forward Selection** en base a métricas como el **AIC** o el **R2** resultan mucho más precisos y logran obtener la mejor combinación de características para predecir un valor mediante un modelo de predicción.

A futuro se podría considerar utilizar diferentes métodos como **One Hot Encoding** para convertir variables categóricas como las provincias o ciudades, que vimos que tenían una gran incidencia como segmentación en las distintas predicciones, en variables numéricas. En el caso de One Hot Encoding utilizando una conversión a variables booleanas, que pueden ser tratadas como features numéricos.

En este trabajo se lograron conocer más en profunidad los diferentes elementos que pueden mejorar técnicas predictivas como la regresión lineal. Se pudo visualizar el impacto de las diferentes segmentaciones y la injerencia de los features (donde incluso cabe destacar que se pudo observar una fuerte injerencia de los features agregados con bases de datos externas) como piedra fundamental de simplificación y refinamiento de los métodos. A su vez, se pudo observar con claridad la radical diferencia que generaban los diferentes métodos como **Forward Selection** a la hora de generar modelos óptimos para cada segmentación en particular.

Referencias

[1] J. García Montalvo, "Perspectivas del precio de la vivienda en España," *Cuadernos de Información económica*, vol. 227, pp. 49–58, 2012.

[2] M. Mazerolle, "Improving data analysis in herpetology: using akaike's information criterion (AIC) to assess the strength of biological hypotheses," *Amphibia-Reptilia*, vol. 27, pp. 169–180, 03 2006.

[3] "Instituto nacional de estadística, geografía e informática (INEGI)." `http://en.www.inegi.org.mx/datos/`. Accedido: 11-07-2020.

6. Apéndice

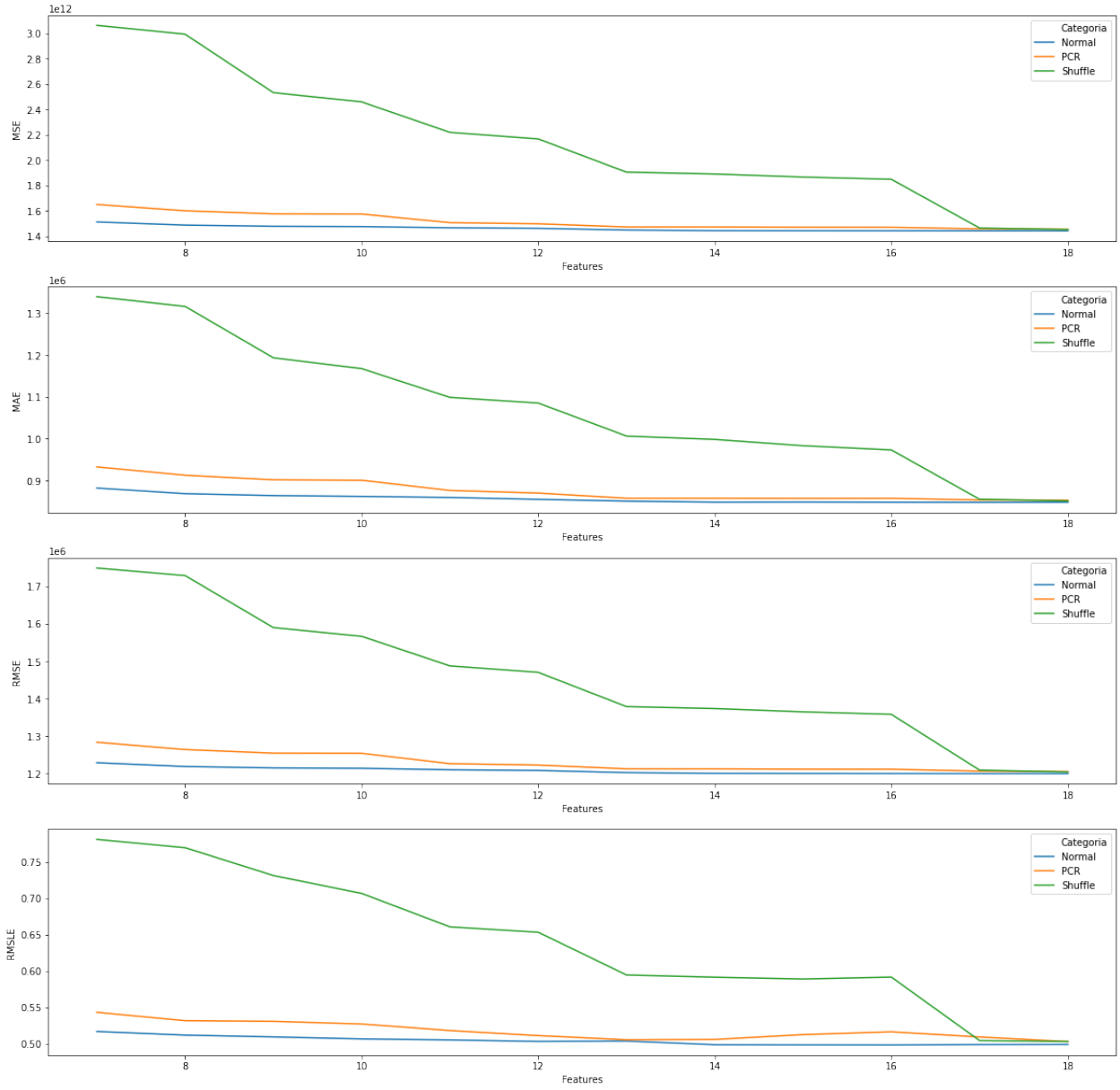

Figura 11: Análisis de las métricas para cada implementación de selección de Features en una predicción de precio para departamentos de todo México para más de 6 features.

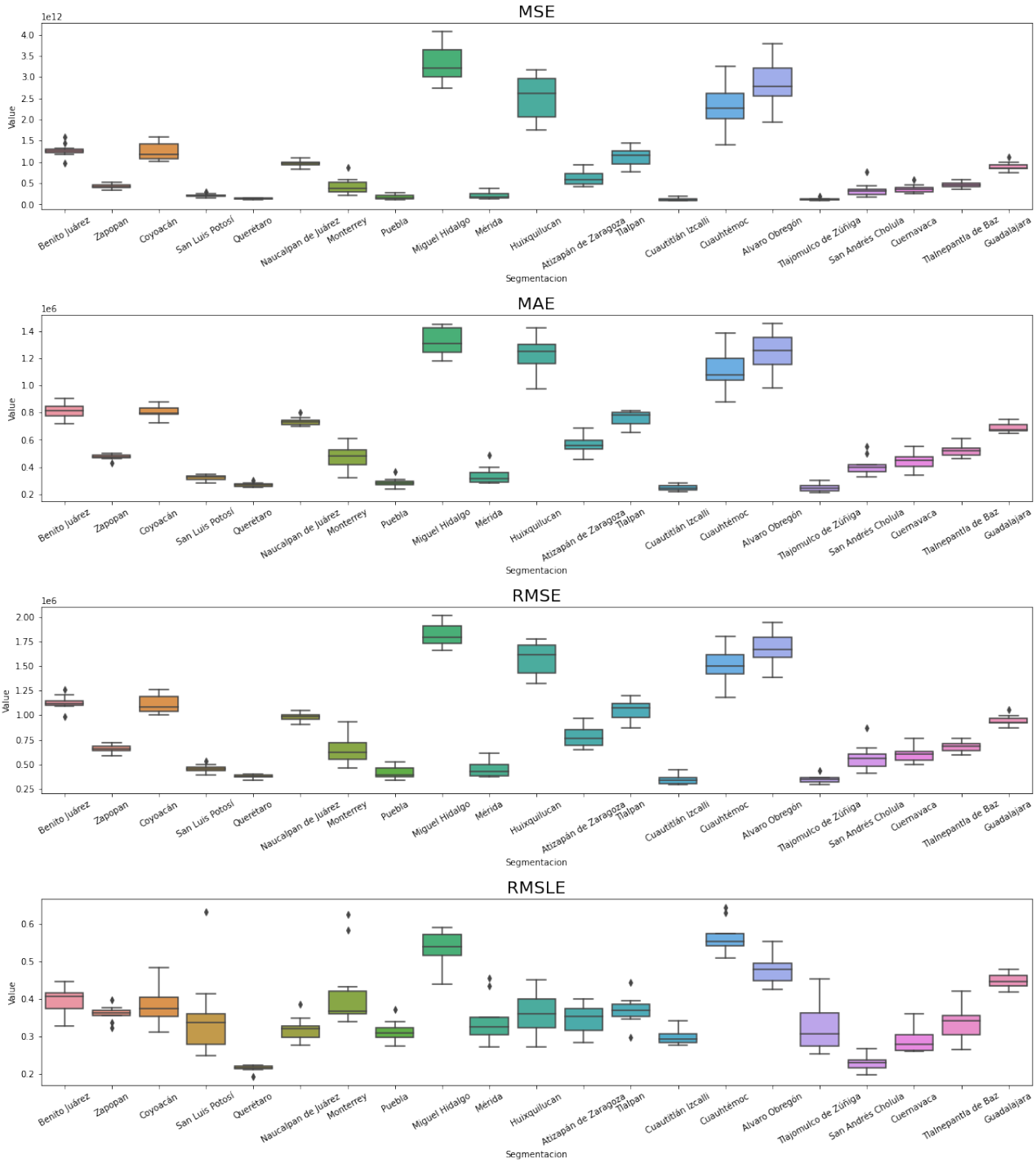

Figura 12: Boxplots correspondientes a las métricas obtenidas en la predicción de **precios** para cada ciudad.

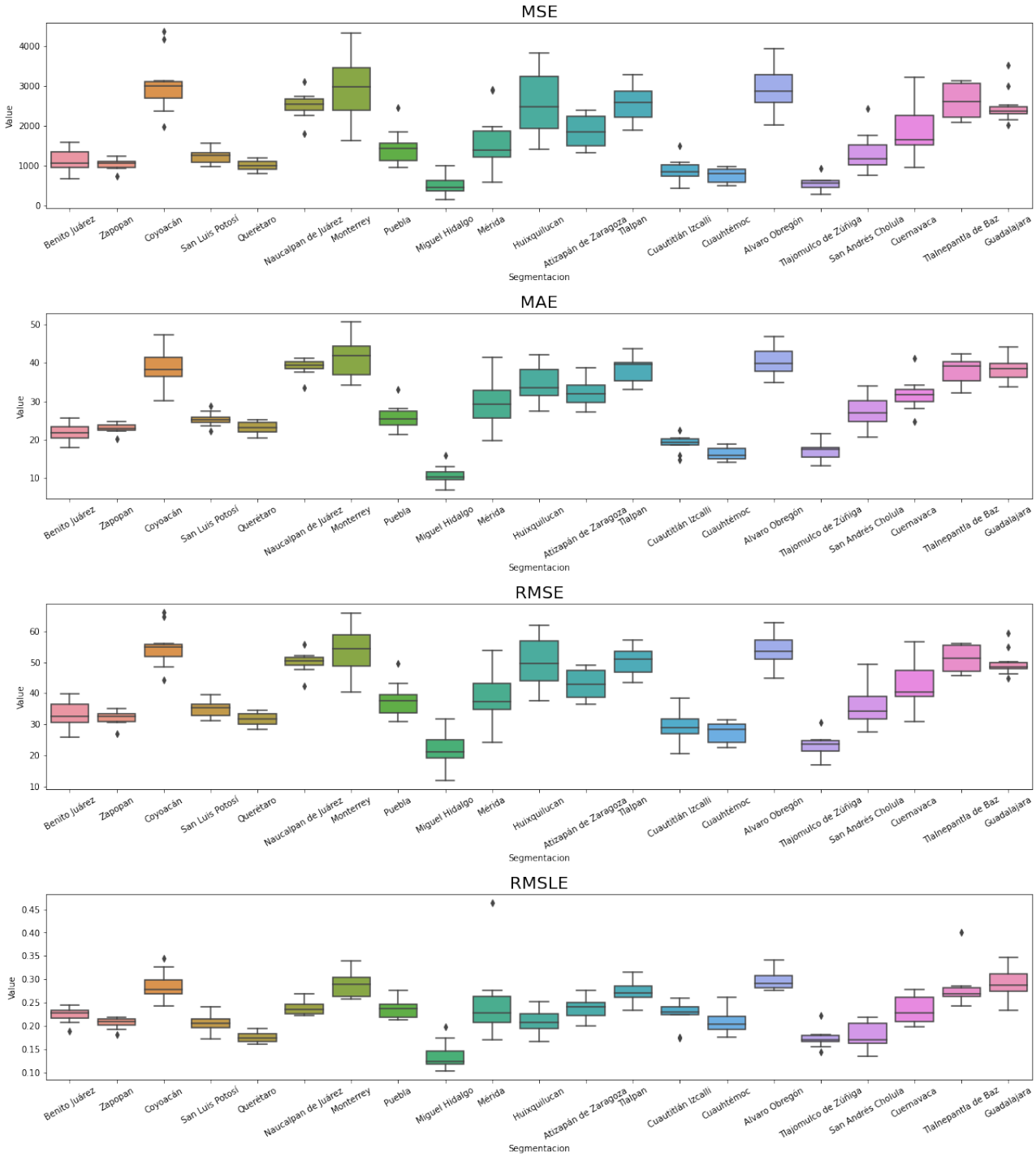

Figura 13: Boxplots correspondientes a las métricas obtenidas en la predicción de **metros cubiertos** para cada ciudad.

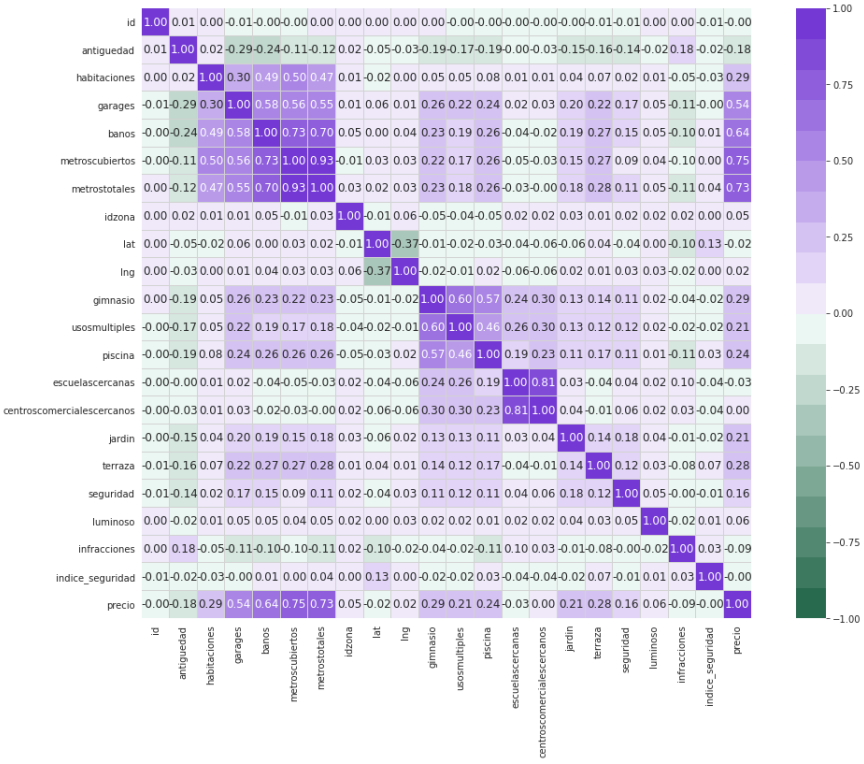

Figura 14: Matriz de correlación para segmentación por **departamento**

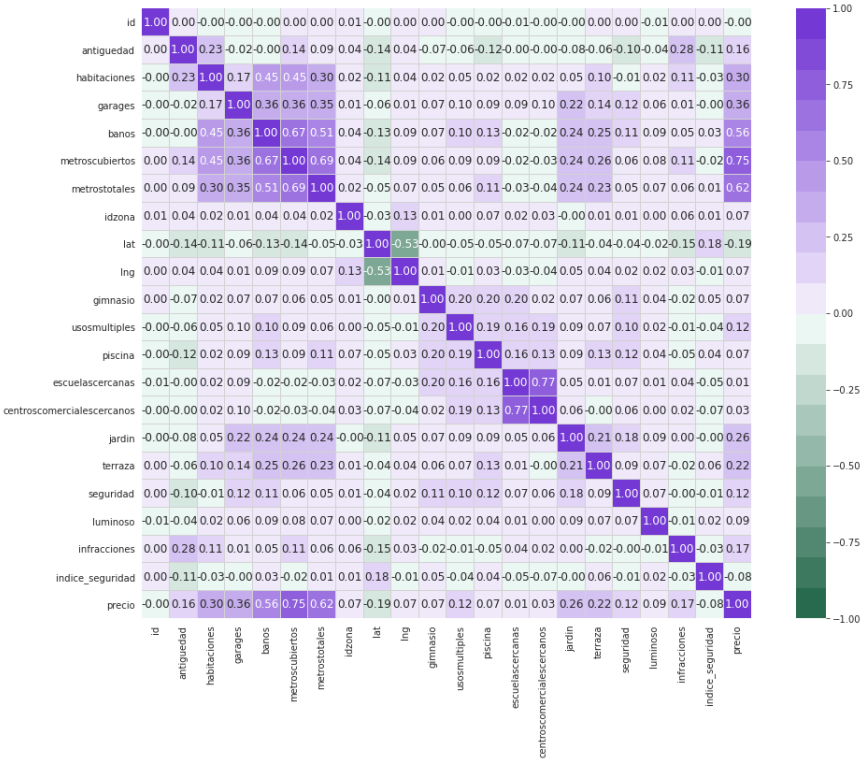

Figura 15: Matriz de correlación para segmentación por **casa**

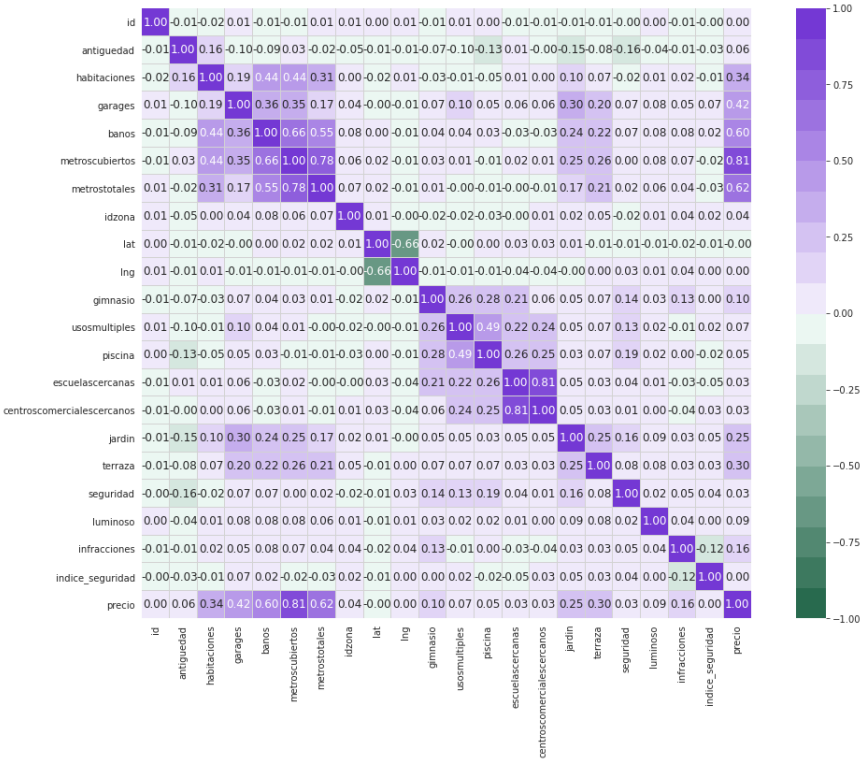

Figura 16: Matriz de correlación para segmentación por **ciudad**. En este caso en particular se utilizó a **Querétaro**, que presentó una gran proporción de casas.

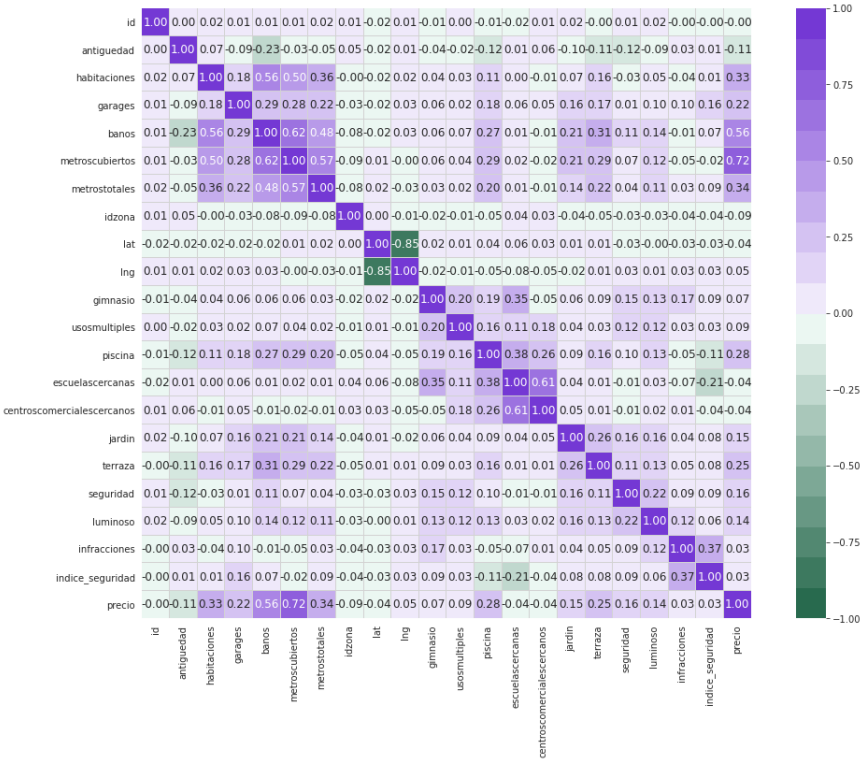

Figura 17: Matriz de correlación para segmentación por **provincia**. En este caso en particular se utilizó a **Yucatán**.