Predicciones del precio de la vivienda a partir de datos del registro de Reino Unido

Domingo Parrales de la Cruz

Octubre 2024

Índice

1. Objetivos del proyecto **3**

1.1. Situación . . . 3

1.2. Complicación . . . 3

1.3. Objetivos . . . 3

2. Principales logros obtenidos **4**

3. Metodología **4**

3.1. Descripción de los objetivos . . . 4

3.2. Extracción, transformación y carga de datos . . . 4

3.2.1. Datos de Kaggle . . . 4

3.2.2. Datos censales . . . 13

3.2.3. Mapa de Reino Unido . . . 15

3.2.4. Posición de pueblos y ciudades . . . 15

3.3. Resumen de las hipótesis de trabajo adoptadas . . . 17

4. Detalles de los modelos **17**

4.1. Modelos de regresión . . . 18

4.1.1. Variables del censo de Reino Unido . . . 18

4.1.2. Variables a partir de la información geográfica de Reino Unido . . . 18

4.1.3. Variables de los datos de Kaggle . . . 19

4.1.4. Escalado de las variables . . . 20

4.1.5. Selección de variables . . . 20

4.1.6. Información sobre los parámetros de los modelos utilizados . . . 23

4.2. Series temporales . . . 25

4.2.1. Análisis de la granularidad de la escala temporal . . . 26

4.2.2. Selección de hiperparámetros y construcción del modelo autorregresivo . . . 29

4.2.3. Predicción de la evolución de categorías específicas . . . 35

5. Resultados principales **38**

5.1. Modelo de regresión . . . 38

5.2. Series temporales: predicción de la tendencia del precio medio de las transacciones en Reino Unido . . . 39

6. Recomendaciones de uso **40**

7. Conclusión **40**

1. Objetivos del proyecto

1.1. Situación

Disponemos de datos sobre las ventas de viviendas en el Reino Unido desde enero de 1995 hasta abril de 2023. La mayor parte de la información se refiere a características geográficas, como el condado o el código postal correspondiente. También contamos con variables que describen características específicas de la vivienda, el tipo de contrato y la transacción realizada.

En la actualidad, estos datos podrían usarse para estimar el precio medio de una región en una fecha concreta. Sin embargo, aunque esto es de interés para la compra y venta de viviendas, puede no resultar suficiente. Por ejemplo, factores que facilitarían este tipo de transacciones incluirían la estimación del precio futuro de una vivienda o la predicción de las tendencias del mercado.

1.2. Complicación

Excepto el precio de venta de la vivienda, todas las variables son categóricas y la mayoría presentan una alta cardinalidad. Además, las variables que deberían proporcionar mayor precisión en la ubicación geográfica (más allá del pueblo o la ciudad) presentan una cantidad considerable de datos faltantes. Por ejemplo, hay un gran número de viviendas para las cuales no se especifica la calle.

Es necesario abordar este problema utilizando técnicas de ciencia de datos por varios motivos:

- En primer lugar, la construcción directa de un modelo es inviable debido a la alta cardinalidad de las variables categóricas. Es necesario encontrar otras formas de representar sus características.
- Además, resulta crucial identificar relaciones entre las variables para entender mejor el comportamiento del mercado.
- Finalmente, conocer los precios en función de la zona donde se encuentra la vivienda, junto con una predicción de su precio futuro, puede proporcionar una base sólida sin requerir un conocimiento previo y detallado de la zona.

1.3. Objetivos

- Predecir el precio de una vivienda según su ubicación y características básicas, así como de los precios medios en regiones más generales.
- Análisis de los principales factores que afectan al precio de la vivienda.
- Predicción de la tendencia del precio medio de las transacciones.
- Predicción del tipo de vivienda que proporcionará un mayor retorno de inversión.

2. Principales logros obtenidos

Se ha logrado una alta capacidad de predicción de los precios medios mensuales por distrito, que también se aplica, aunque en menor medida, a viviendas concretas.

Además, se ha logrado capturar la tendencia real de los precios a lo largo del tiempo, lo que permite estimar las fechas con mayor retorno de inversión. Además, se ha analizado qué tipo de viviendas son de interés para su compra-venta, alquiler o construcción.

Finalmente, a partir de las correlaciones entre las variables creadas, se han identificado las principales características que debe tener una vivienda para maximizar los beneficios.

3. Metodología

3.1. Descripción de los objetivos

En este proyecto afrontamos la predicción de precios con dos enfoques diferentes:

- Primero, hacemos un modelo de regresión, que podrá utilizarse tanto para estimar el precio futuro de viviendas concretas (especificando el tipo de vivienda, el tipo de contrato y la ciudad o pueblo al que pertenece) como para calcular el precio medio en regiones más amplias, especialmente a nivel de distrito.
- Por otro lado, utilizamos series temporales autorregresiva, es decir, que se modeliza la evolución del precio de las transacciones basándose únicamente en su comportamiento pasado. No obstante, con este modelo solo se trabaja a nivel de Reino Unido completo, aunque sí que estudiaremos distintas categorías por separado, como pisos, viviendas de alquiler o viviendas de segunda mano. De esta forma, se podrá concretar qué tipo de vivienda dará un mayor retorno a la inversión.

3.2. Extracción, transformación y carga de datos

Trabajamos con cuatro fuentes de datos principalmente: los datos de Kaggle [1], datos censales de Reino Unido [2], un mapa de Reino Unido [3] e información geográfica sobre los pueblos y ciudades [4][5]. Para cada uno se tiene una fase ETL específica.

3.2.1. Datos de Kaggle

Estos datos contienen información sobre la venta de viviendas y están distribuidos en varios archivos .csv según el año correspondiente. Esto resulta ventajoso, ya que, para evaluar la capacidad de predicción sobre casos futuros, se puede dividir el conjunto de datos en un conjunto de entrenamiento, que abarque desde enero de 1995 hasta diciembre de 2020, y un conjunto de prueba, que utilizará los datos desde enero de 2021 hasta abril de 2023.

Tomando una pequeña muestra de los datos y revisando la información proporcionada en la fuente original [6], podemos distinguir los siguientes tipos de variables en nuestro conjunto de datos:

- **Numéricas**: solo incluye el precio de venta de la vivienda. Se puede analizar su evolución y determinar si es necesario aplicar alguna transformación que mejore el rendimiento de los modelos.
- **Fechas**: la única variable de este tipo es la fecha de la transacción, de la cual se pueden extraer variables derivadas como el año, el mes y el día de la venta. Esto permitirá un análisis más detallado de las dependencias temporales en los modelos.
- **Categorías con baja cardinalidad**: las más relevantes son el tipo de vivienda, el tipo de tenencia de los inquilinos, la antigüedad de la vivienda y el tipo de transacción. Además, hay una variable que indica si fue necesario pagar tasas adicionales y otra que informa si la entrada fue modificada, añadida o eliminada de la base de datos.
- **Categorías con alta cardinalidad**: recogen la información espacial de los datos. Permiten un nivel de granularidad hasta el número de la vivienda, aunque presentan una gran cantidad de datos faltantes. También incluyen un identificador único para cada transacción.

Además, siguiendo la descripción proporcionada por los datos [6], se pueden modificar los nombres de las categorías para que sean más comprensibles en el análisis. A continuación, se detalla el tratamiento aplicado a cada tipo de variable.

Categorías con baja cardinalidad

Dado que estas variables son las únicas que proporcionan información específica sobre la vivienda, es fundamental que se representen de la mejor manera posible. Para lograr esto, se tomará como referencia la variable objetivo y se analizará cómo influye cada categoría en ella.

Podemos comenzar con la variable *Old.New*, que indica si la vivienda es nueva o de segunda mano. Como se muestra en la Figura 1, la evolución del precio de venta para cada categoría presenta una clara diferenciación al inicio, pero, conforme avanza el tiempo, ambas siguen una progresión similar. Cabe destacar que, a pesar del ruido presente, se mantiene una tendencia clara.

Además, está el tipo de tenencia de los inquilinos, representado por la variable *Duration*. En la Figura 2, se puede observar que, a diferencia del caso anterior, la categoría 'Desconocido' introduce ruido en la variable. También se aprecia que esta categoría solo es relevante para datos pasados, por lo que es seguro eliminar estos casos, ya que no afectarán las predicciones futuras.

También se ha estudiado la variable *PPD.Category.Type*, que indica en qué casos se ha pagado un precio adicional por la propiedad, ya sea por tasas de embargo, herencia, etc. Como se muestra en la Figura 3, los casos en los que se pagan tasas adicionales no presentan un patrón claro hasta los últimos años, y la dispersión es considerablemente alta. Sin embargo, de la descripción de

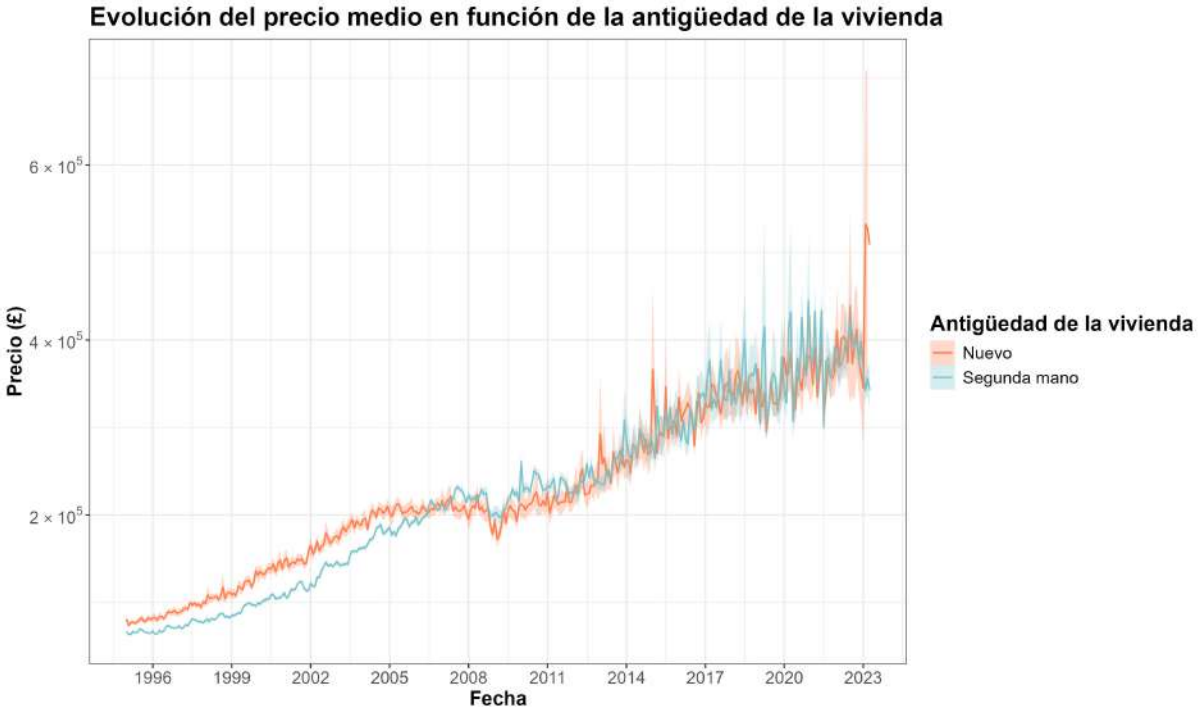

Figura 1: evolución del precio de la vivienda diferenciando entre el conjunto de propiedades nuevas y de segunda mano. Junto a la media se representa la dispersión del precio para esa fecha como un área semitransparente. En ambas se tiene una evolución similar.

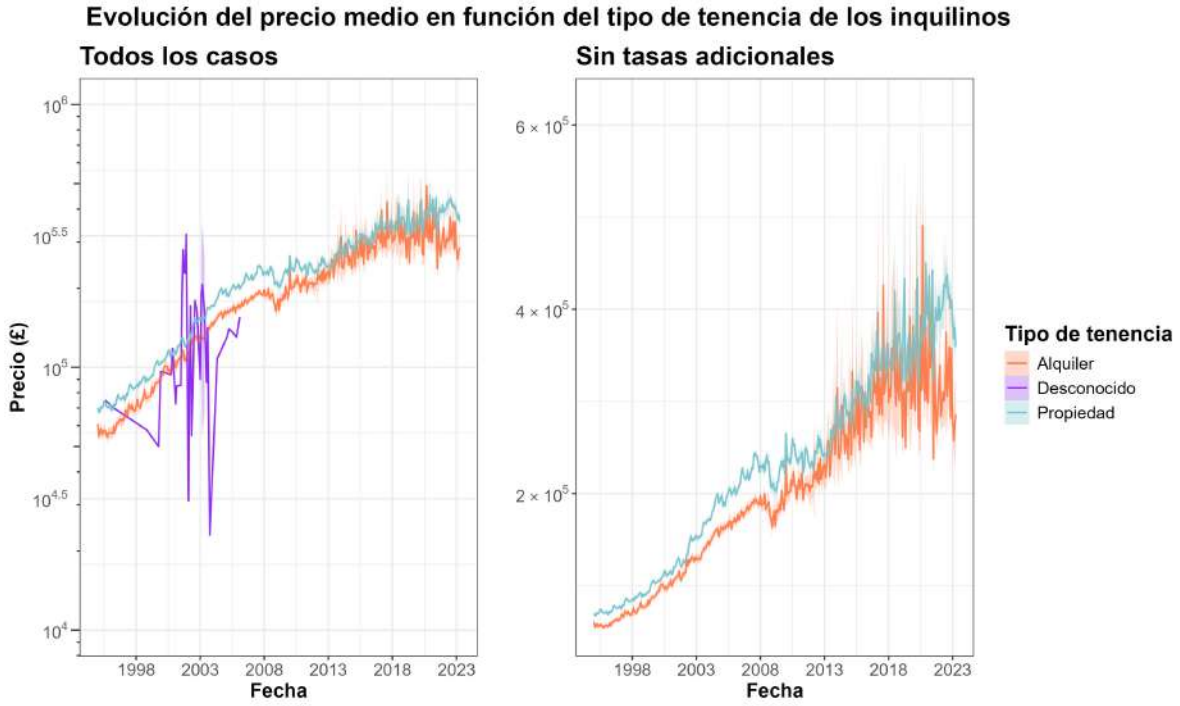

Figura 2: evolución del precio de la vivienda según el tipo de tenencia de los inquilinos. Podemos ver que la categoría 'Desconocido' solo ocurre en datos pasados. Se representa, junto a la línea del valor medio, la dispersión para esa fecha mediante el área semitransparente que la acompaña.

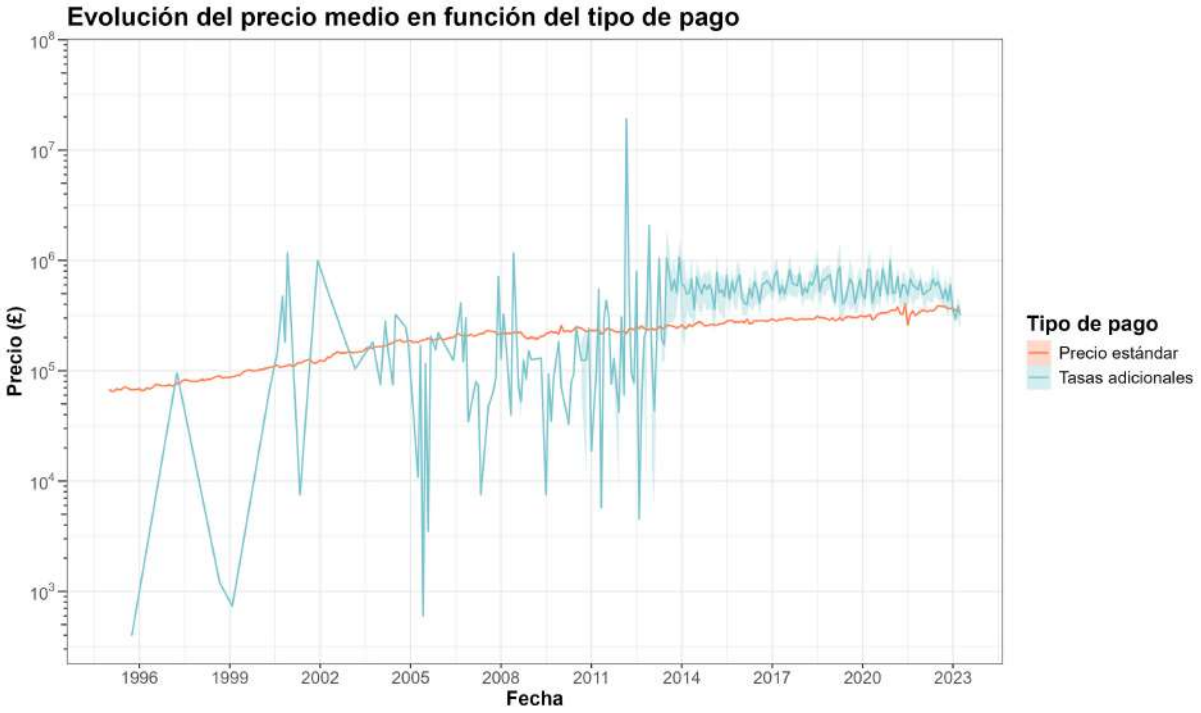

Figura 3: evolución del precio de la vivienda en función de si se pagaron o no tasas adicionales en la transacción. Junto al valor medio, se muestra una zona sombreada que representa la dispersión para esa fecha. Se aprecia que los casos en los que se pagó un precio adicional presentan una dispersión muy elevada.

los datos [6] se extrae que la categoría 'Otro' de la variable *Property.Type* corresponde siempre a casos de venta con tasas adicionales, por lo que debemos comprobar también la relación que tiene con esa variable antes de descartar estos casos.

Por último, está la variable *Property.Type*, que nos indica el tipo de propiedad. Como se muestra en la Figura 4, la variable 'Otro' presenta un problema idéntico al anterior. No obstante, se puede ver que son una minoría al fijarnos en la Figura 5 (el tamaño de los cuadrados representa el número de casos que hay para cada pareja). En consecuencia, se puede descartar la categoría para conservar la estabilidad del modelo, como se observa a la derecha en la Figura 4.

Categorías con alta cardinalidad

Especificar la ubicación geográfica de una vivienda mediante nombres tiene la ventaja de que se puede comparar con otras bases de datos con relativa facilidad. Sin embargo, pueden existir diferencias, como en el uso de puntuación, la ausencia de *stopwords* (como 'the' u 'of') o incluso el uso de abreviaturas ('St.' en lugar de 'Saint').

Esto plantea un preprocesamiento que puede volverse inabarcable dependiendo del nivel de precisión o granularidad deseado. Para tomar una decisión, se ha calculado el número de categorías únicas que presenta cada variable geográfica, como se muestra en el Cuadro 1. El número de categorías aumenta progresivamente hasta llegar a la localidad, donde se observa un incremento

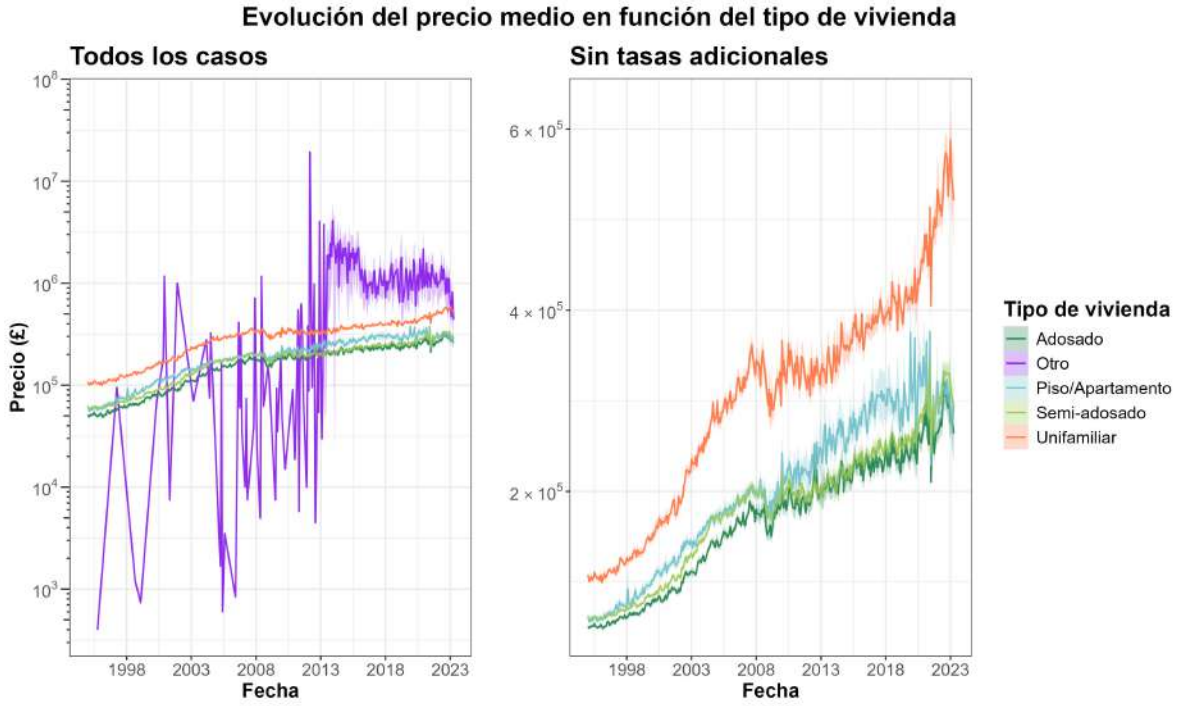

Figura 4: evolución del precio de la vivienda segregado según el tipo de propiedad. En la imagen izquierda vemos que la categoría 'Otro' introduce una gran dispersión, interfiriendo en la caracterización de la tendencia de las otras categorías, que se muestran en la imagen derecha.

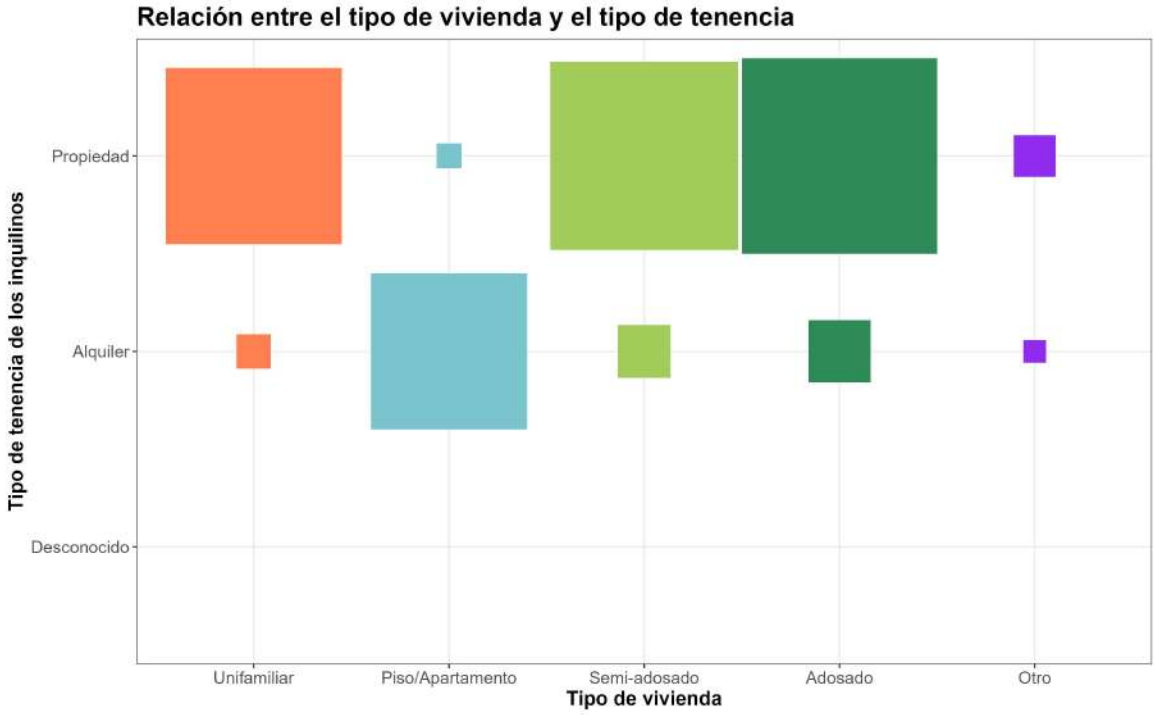

Figura 5: representación del número de casos para cada par *Property.Type* - *Duration*. Esta cualidad queda representada por el tamaño del cuadrado. Podemos ver que la categoría 'Otro' de *Property.Type* y 'Desconocido' de *Duration* son una minoría en los datos.

Variable	Número de categorías únicas
Condado	132
Distrito	467
Pueblo/Ciudad	1172
Barrio	23862
Calle	326400

Cuadro 1: número de categorías únicas de algunas de las variables geográficas disponibles. Las ausentes hacen referencia al código postal y el número de la vivienda, pero tienen una cantidad de datos faltantes excesiva.

drástico. Considerando esto, se ha decidido establecer como nivel máximo de granularidad los pueblos y ciudades, aunque el análisis se centrará principalmente en los distritos.

Además, es importante señalar que los resultados de el Cuadro 1 no consideran el número de combinaciones posibles, derivadas de la estructura jerárquica de estas variables[1]. En concreto, si se considera la multiplicidad de combinaciones en los datos de Condado - Distrito - Pueblo o Ciudad, se obtienen 7563 casos distintos.

En definitiva, se descartará cualquier variable que proporcione una precisión mayor que la de pueblos o ciudades. Esto también evita el uso de variables con datos faltantes. En la Figura 6 se muestran las relaciones entre los datos faltantes de las variables. A partir de esta figura, podemos deducir lo siguiente:

- A mayor profundidad geográfica, se observa una mayor proporción de datos faltantes.
- La variable *SAON* es la principal fuente de datos faltantes. Esto es comprensible, ya que solo se utiliza cuando es necesario añadir información adicional al número de la casa (*PAON*), como en el caso de indicar el número de piso.
- La ausencia de datos en *Locality* sugiere que hay propiedades sin un barrio asignado. Esto es posible, especialmente en áreas rurales o aisladas, aunque no debería ser tan frecuente.
- Existen casos en los que faltan exclusivamente el código postal o el nombre de la calle. A diferencia del caso anterior, aquí realmente falta información esencial que será difícil de imputar.

Variable objetivo: precio de venta

Al construir un modelo de regresión, uno de los factores más importantes es considerar la normalidad de la variable objetivo, ya que muchos modelos suponen esta característica. Para comprobar si este es nuestro caso, se ha representado a la izquierda de la Figura 7 la distribución de los valores de la variable objetivo.

[1]Un condado contiene distritos. Un distrito contiene pueblos y ciudades. Los pueblos y ciudades contienen barrios, y los barrios contienen calles.

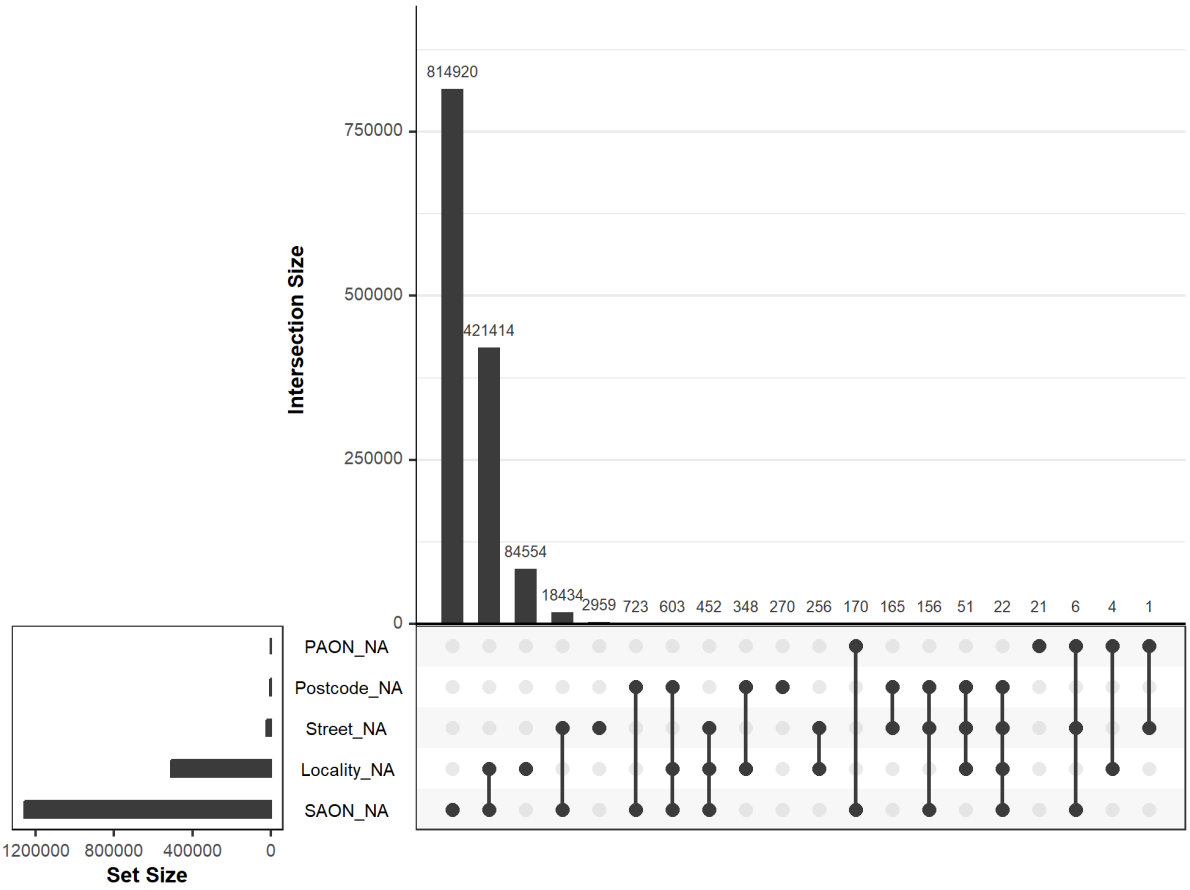

Figura 6: distribución de datos faltantes entre las variables de los datos de Kaggle creado con la librería *naniar* de *R*. Con las barras horizontales de la izquierda, podemos ver la cantidad de datos faltantes en las variables afectadas. Con los puntos en la zona inferior derecha se representan las conexiones entre los datos faltantes. Las barras verticales representan el número de casos en los que se cada cada relación marcada con los puntos.

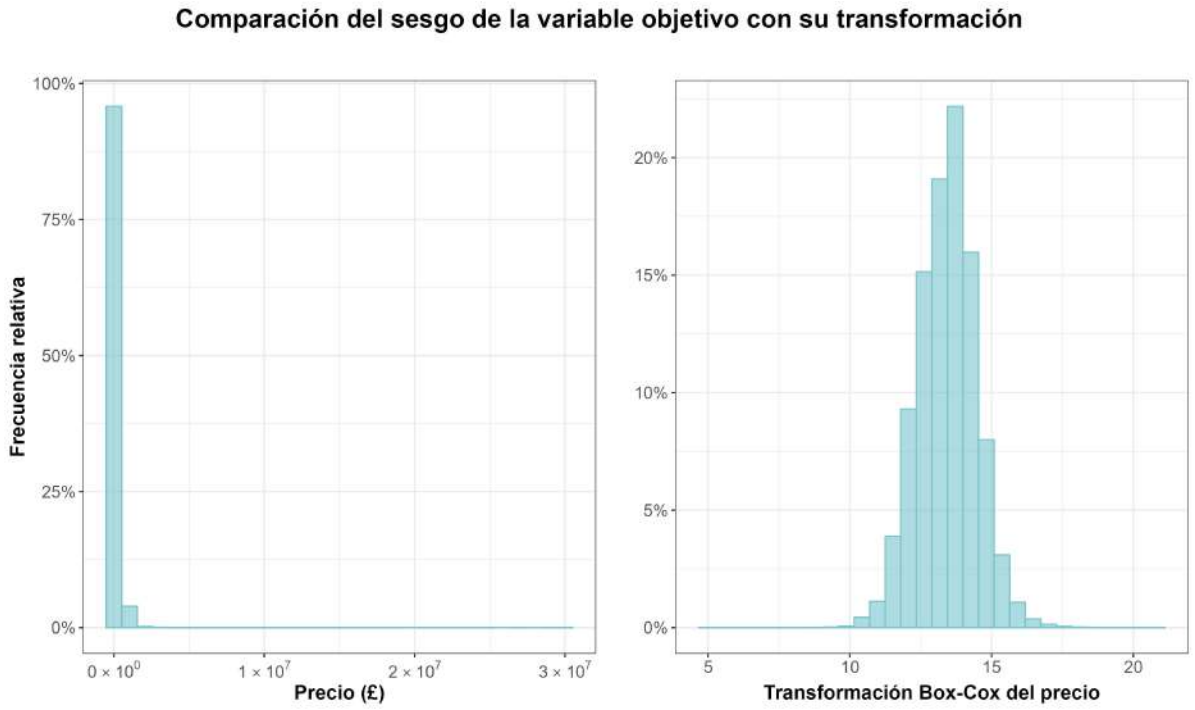

Figura 7: A la izquierda representamos el número de casos en la horquilla de precios que representa el ancho de cada barra. A la derecha, se tiene una representación análoga, pero con la variable objetivo transformada (mediante una transformación Box-Cox).

Es evidente que la distribución de precios está fuertemente sesgada a la izquierda, lo que indica que la mayoría de los precios están concentrados en valores bajos, mientras que las propiedades con precios de venta más altos son una minoría.

Una forma de corregir este sesgo es mediante la transformación de la variable. En particular, se ha optado por aplicar la transformación Box-Cox a los datos y, como se muestra a la derecha en la Figura 7, la distribución resultante se ajusta de manera más cercana a una distribución gaussiana.

Outliers

A partir del análisis de normalidad de la variable objetivo, se ha identificado que los precios más altos representan una minoría en el conjunto de datos. Sin embargo, se puede extraer más información comparando los *outliers* de los datos históricos (desde enero de 1995 hasta abril de 2023) con los de los más recientes (desde el 2020 hasta abril del 2023).

Para ello, se analizan las Figuras 8, 9 y 10, y, de ellas, se extraen varias características comunes a las tres:

- La mediana de los datos más recientes es consistentemente mayor que la mediana de los precios históricos. Esto indica un crecimiento económico en el sector inmobiliario del Reino Unido.
- Aunque la escala logarítmica del eje horizontal dificulta su visualización, los *outliers* infe-

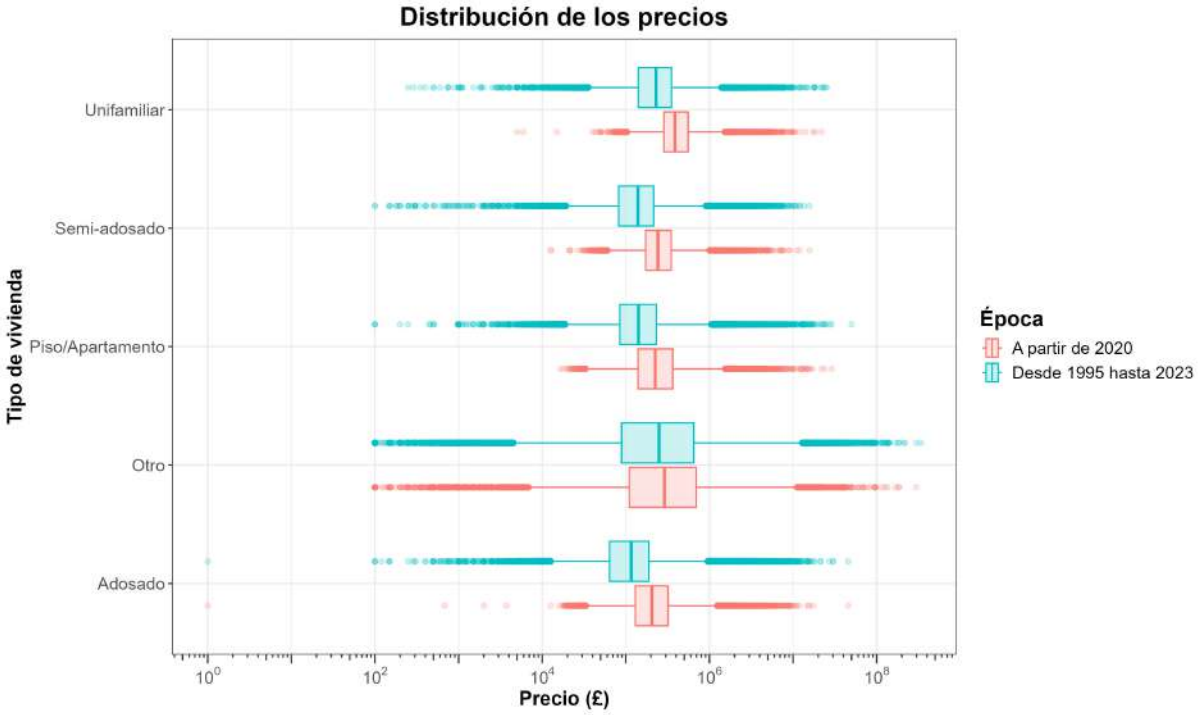

Figura 8: boxplot del precio de venta de las propiedades, donde se han dividido los datos en categorías. Se representan conjuntamente los datos más recientes (desde enero de 2020 hasta abril de 2023) junto a los históricos (desde enero de 1995 hasta abril de 2023). Se utiliza escala logarítmica en el precio para facilitar la visualización. Se observa que, en la mayoría de casos, los precios actuales superan a los históricos.

riores se alejan menos de la mediana que los *outliers* superiores.

- Los *outliers* con precios más elevados no se encuentran necesariamente en el rango de valores desde 2020, aunque en promedio las propiedades se vendan a un precio mayor. Esto se aprecia especialmente en la Figura 10, por ejemplo, para las viviendas con contratos de alquiler, en la que los *outliers* más altos se dan para los precios históricos.

Además, se puede extraer información relevante sobre la evolución del mercado:

- En la Figura 8 se observa que, sin tener en cuenta la categoría *Otro*, los precios más elevados suelen involucrar viviendas unifamiliares. No obstante, si tenemos en cuenta los outliers, estos precios son comparables al de algunos pisos/apartamentos y viviendas adosadas, aunque en menor medida.

- De la Figura 9, se destaca que los precios de las viviendas nuevas están más acotados que los de las viviendas de segunda mano. Esto es especialmente evidente en los precios de 2020.

- Finalmente, la Figura 10 muestra que, en general, los contratos de propiedad son más caros que los de alquiler, aunque la diferencia no es tan grande.

Para manejar los *outliers*, se añadirán cuatro nuevas variables al conjunto de datos, en lugar de eliminar estos casos. Las primeras dos variables serán la media diaria del primer y tercer cuartil

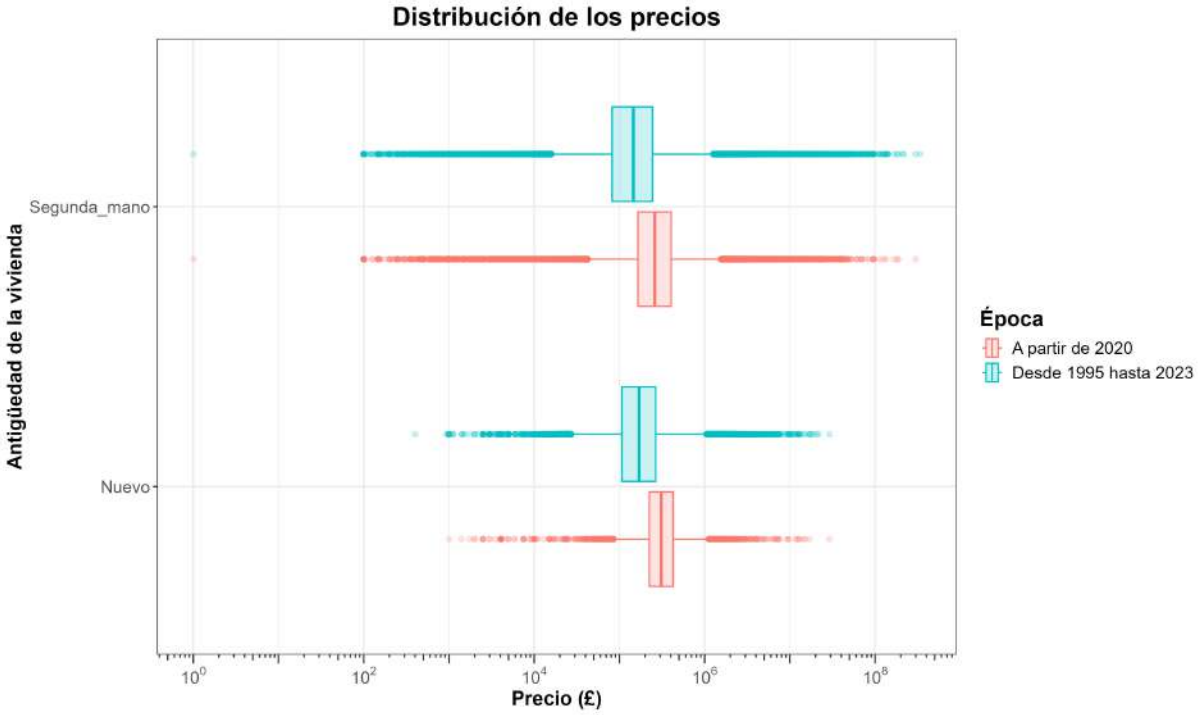

Figura 9: boxplot del precio de venta de las propiedades, donde se han dividido los datos en categorías. Se representan conjuntamente los datos más recientes (desde enero de 2020 hasta abril de 2023) junto a los históricos (desde enero de 1995 hasta abril de 2023). Se utiliza escala logarítmica en el precio para facilitar la visualización. Se observa que, en la mayoría de casos, los precios actuales superan a los históricos.

de la variable *Price*, y otras dos indicarán si el caso es un *outlier* superior o inferior (o ninguno si ambas son cero).

Este tratamiento es sencillo de aplicar en el conjunto de entrenamiento, pero no es posible construir variables basadas en el precio para el conjunto de prueba. Por lo tanto, se ha utilizado un modelo de series temporales para extrapolar los valores de los cuartiles. Además, para mejorar la precisión, se ha optado por enfocarse en los cuartiles mensuales en lugar de los cuartiles diarios[2].

3.2.2. Datos censales

Esta es la principal fuente de enriquecimiento de los datos originales. Cuenta con varias categorías relevantes como la densidad de trabajo, la cantidad de casas o el número de estudiantes, todas ellas referente a los distritos de las regiones de Inglaterra y Gales, que son las únicas que se tienen en los datos originales.

La principal ventaja de estos datos es que proporcionan una representación numérica de los distritos, lo que permite representarlos sin necesidad de codificarlos como una variable categórica.

[2]Es complicado replicar la evolución del precio diaria, dada las oscilaciones. No obstante, las medias mensuales reducen este ruido, por lo que se vuelven más sencillas de modelar.

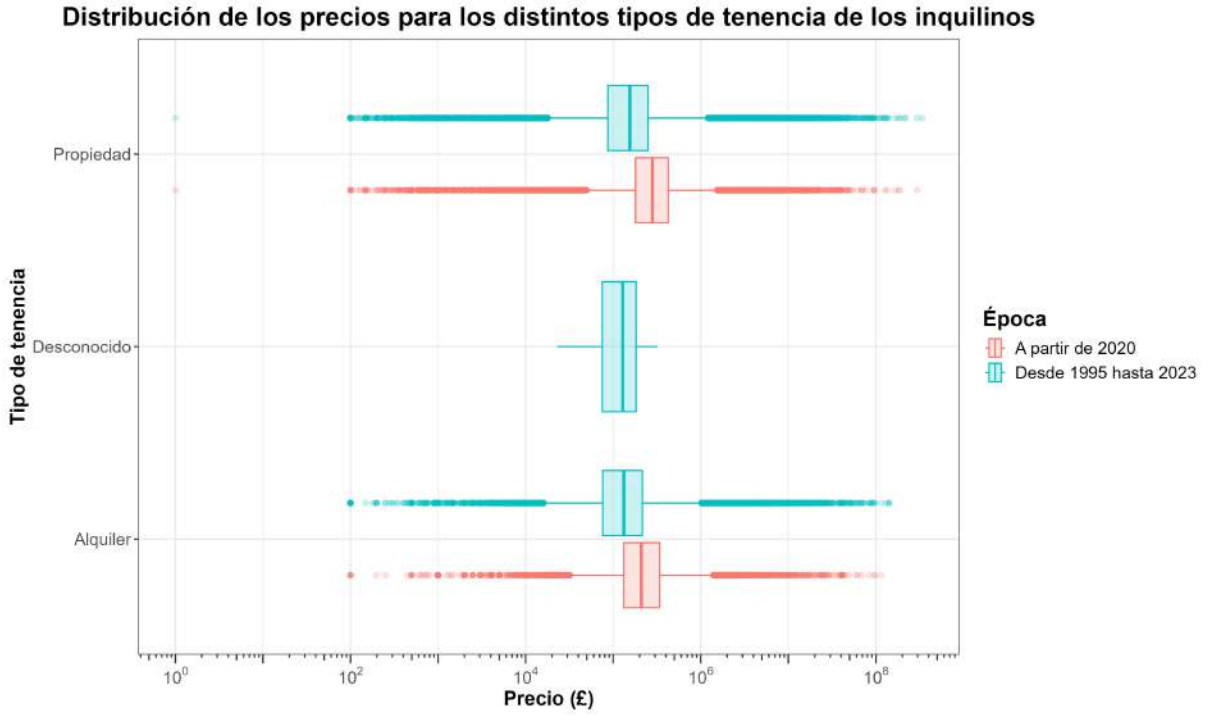

Figura 10: boxplot del precio de venta de las propiedades, donde se han dividido los datos en categorías. Se representan conjuntamente los datos más recientes (desde enero de 2020 hasta abril de 2023) junto a los históricos (desde enero de 1995 hasta abril de 2023). Se utiliza escala logarítmica en el precio para facilitar la visualización. Se observa que, en la mayoría de casos, los precios actuales superan a los históricos.

Sin embargo, la incorporación de estos datos no es completamente directa, ya que los nombres de algunos distritos no coinciden con los de los datos de Kaggle. En el Cuadro 2 se muestran algunos de los cambios realizados.

Otro factor a tener en cuenta es que los datos extraídos corresponden a un año concreto. Se ha optado por utilizar los más recientes, lo que, en principio, favorecerá las predicciones futuras y podría afectar negativamente a las predicciones pasadas. Además, para los distritos que desaparecieron antes del registro, no habrá ningún valor asociado, por lo que se han imputado estos valores faltantes utilizando la media de la variable correspondiente.

Distritos de los datos censales	
Original	**Corrección**
County Durham	Durham
St. Helens	St Helens
Bristol, City of	Bristol

Distritos del mapa de Reino Unido	
Original	**Corrección**
Rhondda, Cynon, Taff	Rhondda Cynon Taff
St. Helens	St Helens
Telford and Wrekin	Wrekin

Cuadro 2: ejemplos de correcciones realizadas a los nombres de distritos en el conjuntos de datos del censo de Reino Unido y los de su mapa. El objetivo de dichos cambios es que coincidan con los del conjunto de datos original. Destacan la eliminación de puntuación y de términos como 'County' o 'City'.

3.2.3. Mapa de Reino Unido

Se ha utilizado una representación en polígonos de los distritos actuales en el Reino Unido para generar visualizaciones geográficas de los datos.

La principal dificultad en el uso de estos datos es similar a la que presentan los datos censales: es necesario corregir algunos nombres de distritos y, dependiendo de lo reciente que sea el mapa, los distritos antiguos de nuestros datos no estarán representados. Algunos ejemplos de las correcciones realizadas en este conjunto de datos se encuentran a la derecha en el Cuadro 2.

Utilizando estos datos, se ha elaborado en la Figura 11 una representación espacial de la media de precios por distrito. A partir de esta representación, se pueden observar tres características fundamentales:

- En primer lugar, se confirma nuevamente que, en general, los precios de las transacciones han aumentado en los últimos años, ya que se tienen más colores cercanos al amarillo.
- En segundo lugar, se observa que, a medida que nos acercamos a Londres, los precios aumentan rápidamente. Esto es visible tanto en los precios recientes como al considerar todos los precios registrados. Este patrón puede incluirse como variable en el modelo para capturar parte de la relación espacial entre los distritos.
- Finalmente, se aprecia que, en general, hay agrupaciones de distritos con color similar. Por lo tanto, parece razonable introducir una variable adicional que agrupe los pueblos en función de su proximidad espacial.

3.2.4. Posición de pueblos y ciudades

Hasta ahora, se ha limitado la precisión geográfica a los distritos para reducir las correcciones necesarias al adjuntar otros conjuntos de datos. Sin embargo, si se utilizan los nombres presentes en nuestros datos, no será necesario este tratamiento.

Teniendo esto en cuenta, se ha decidido emplear los nombres de pueblos y ciudades para extraer su posición geográfica (latitud y longitud). Para ello, en primer lugar, se ha utilizado la API de OpenCage [4], y posteriormente se revisaron los casos con valores atípicos o que quedaron sin asignar mediante la API de Nominatim [5][3].

Para hacer las peticiones, inicialmente se proporcionaron direcciones a la API de OpenCage con la estructura 'Condado, Distrito, Pueblo/Ciudad'. Sin embargo, algunas coincidían con lugares en Australia. Posteriormente, con la API de Nominatim, se empleó la estructura 'Pueblo/Ciudad, UK', dando mayor importancia al nombre del pueblo o ciudad y especificando que pertenecen al Reino Unido.

[3]La API de OpenCage tiene un límite de uso diario generoso en su versión de prueba, mientras que Nominatim es Open Source, por lo que se redujo el ritmo de peticiones al usarla. Existen otras opciones como la API de Google [7], que permite un gran número de peticiones y podría resultar más atractivo para proyectos a gran escala.

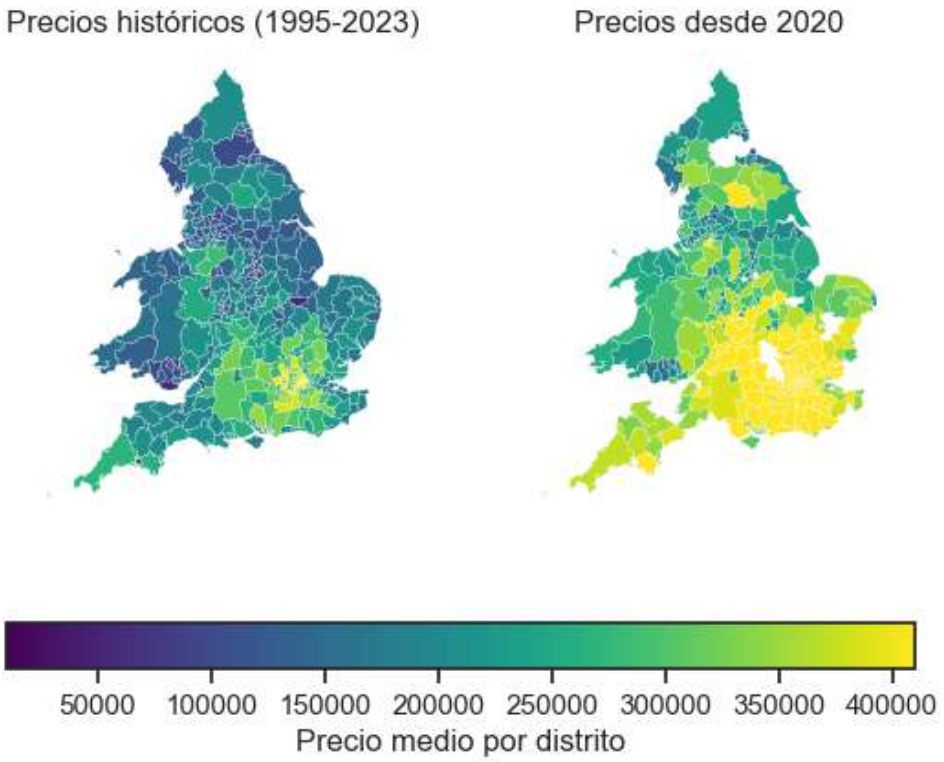

Figura 11: mapa de Reino Unido de las regiones de Inglaterra y Gales. Se muestra, según la escala de color inferior, la distribución geográfica del precio medio de venta de la vivienda por distrito. Para la imagen izquierda, usamos todos los casos disponibles, mientras que en la derecha solo los que se registran tras 2020.

3.3. Resumen de las hipótesis de trabajo adoptadas

Durante el análisis inicial de las variables del modelo, se ha planteado la hipótesis de que el tipo de tenencia de los inquilinos debe ser conocido, que la vivienda debe ser adosada, semi-adosada, unifamiliar o un piso/apartamento, y que las transacciones se realizan sin pagos de tasas adicionales.

Además, existe una hipótesis implícita en la elección de la granularidad de los datos censales: se asume que las características de las ubicaciones de mayor precisión dentro del distrito no presentan variaciones significativas, de modo que pueden ser representadas adecuadamente por el valor medio del distrito al que pertenecen. No obstante, sabemos que esta última hipótesis no es completamente precisa, ya que variables como la densidad laboral varían notablemente entre pueblos y ciudades, y otras más importantes, como el precio medio de las transacciones, varían incluso entre viviendas de una misma calle. Aun así, esta es la representación más confiable que se ha podido alcanzar dadas las limitaciones computacionales y los propios datos disponibles[4].

4. Detalles de los modelos

Tanto para los modelos de regresión como para los de series temporales, el objetivo es obtener un modelo que pueda predecir precios futuros. Por tanto, como los datos utilizados abarcan desde enero de 1995 hasta abril de 2023, hemos decidido mantener los datos a partir de ciertos años como conjunto de test. En concreto:

- Para los modelos de regresión tomamos como conjunto de entrenamiento el rango entre enero de 1995 y diciembre de 2020. Por tanto, se tienen como conjunto de test los datos entre enero de 2021 y abril de 2023. Además, se toma una muestra aleatoria del 5 % de los datos de test como conjunto de validación.
- Para los modelos de series temporales se toma el mismo rango para realizar el estudio inicial. Sin embargo, para el ajuste de hiperparámetros, se pasa a tener tres conjuntos: un conjunto de entrenamiento, con datos entre enero de 1995 y diciembre de 2005; un conjunto de validación, con datos entre enero de 2006 y diciembre de 2015; y un conjunto de test, con datos entre enero de 2016 y abril de 2023. Finalmente, se ha usado *backtesting*[5] para verificar que el modelo seleccionado es capaz de predecir correctamente los datos.

También, se contaba inicialmente con aproximadamente 30 millones de registros, restringidos a las regiones de Inglaterra y Gales. Sin embargo, en base a la información extraída del análisis inicial, se han filtrado los casos para centrarse en:

- Ventas correspondientes a viviendas adosadas, semi-adosadas, unifamiliares o pisos/apartamentos.
- Contratos de venta o alquiler.

[4] La máxima precisión que se consiguen para los datos censales suele ser la de pueblos y ciudades, aunque en este proyecto se toman los datos por distritos por los motivos expuestos anteriormente.

[5] El *backtesting* consiste en evaluar el rendimiento de un modelo para predecir datos futuros no utilizados para el entrenamiento del mismo. De esta forma, se puede verificar su robustez de cara al uso real.

- Viviendas nuevas o de segunda mano.
- Casos que no son considerados outliers.

4.1. Modelos de regresión

Para mejorar la capacidad predictiva de los modelos de regresión, se han añadido múltiples variables y, posteriormente, se ha realizado un cribado considerando factores como las correlaciones y las importancias de las mismas. Veamos en primer lugar las variables añadidas a partir de los datos censales y geográficos.

4.1.1. Variables del censo de Reino Unido

Del censo se ha extraído información sobre los distritos de diversa índole, que puede agruparse en dos categorías:

- **Información socio-demográfica**: número total de personas por distrito y cantidad de ellas que son trabajadoras o estudiantes, la densidad de empleo por habitante y la densidad de población (por hectárea).
- **Información de las viviendas**: número total de viviendas en cada distrito y el porcentaje de estas en cada nivel de eficiencia energética (nivel A, nivel B, etc.).

Todas las variables seleccionadas son numéricas, lo que permite representar los distritos en función de algunas de sus características clave. Para su selección, se consideraron los factores que, en principio, se cree que pueden influir en el precio de la vivienda.

4.1.2. Variables a partir de la información geográfica de Reino Unido

De manera directa, no se disponen de nuevas variables para añadir a los datos disponibles. Sin embargo, es posible crear algunas que aporten información adicional:

- **Zonas costeras**: dado que se cuenta con los polígonos que conforman el mapa del Reino Unido, se puede utilizar la librería *geopandas* de Python para seleccionar los bordes, que en este caso corresponden a las zonas que limitan con el mar. A continuación, se busca la intersección de estos bordes con los polígonos del mapa y se asigna la propiedad *is_coastal*, representando que son distritos costeros.
- **Islas**: utilizando nuevamente los datos del mapa del Reino Unido, es posible contar el número de polígonos adyacentes a cada uno. Aquellos que no tengan ninguno serán considerados islas y quedarán representados en la variable *is_isle*.
- **Agrupaciones de pueblos y ciudades**: a partir de la latitud y la longitud de los pueblos y ciudades, se pueden agrupar en clusters por cercanía. Esto se ha realizado con el algoritmo *HDBSCAN*, que es capaz de detectar clusters de diversas formas y no se ve muy afectado por el ruido. El grupo asignado se almacena en la variable *spatial_cluster*.

- **Distancia a Ciudad de Londres**: se ha observado que, a medida que nos alejamos de Londres, los precios medios por distrito tienden a bajar. Por tanto, utilizando la latitud y la longitud de los pueblos, se han calculado las distancias a la Ciudad de Londres. No obstante, las coordenadas utilizadas corresponden a una geometría esférica, por lo que fue necesario adaptar la fórmula para calcular la distancia [8]. Esta información se almacena en la variable *distance_to_london*.

4.1.3. Variables de los datos de Kaggle

Las variables de este conjunto de datos que ya están presentes y se pueden utilizar para el entrenamiento de modelos son:

- **Duration**: indica el tipo de tenencia de los inquilinos, es decir, alquiler, propiedad o desconocido. Esta variable se codifica asignando un valor numérico a cada categoría. Cabe destacar que la categoría 'desconocido' será eliminada antes de entrenar el modelo, pero resulta útil en el análisis preliminar de las variables.

- **Property.Type**: indica el tipo de vivienda en la transacción, es decir, unifamiliar, adosada, semi-adosada, piso/apartamento u otra. Al igual que en el caso anterior, se codifican los valores, y la categoría 'Otro' se eliminará antes del entrenamiento del modelo.

- **Old.New**: indica si la vivienda es nueva o de segunda mano, permitiendo una codificación binaria. De este modo, la variable pasa a ser *First_hand*, indicando si la vivienda es nueva.

- **PPD.Category.Type**: indica si la transacción fue realizada con tasas adicionales o no. De nuevo, se realiza una codificación binaria para crear la variable *paid_additional_price*. Sin embargo, esta variable será eliminada antes del entrenamiento, ya que todos los casos con tasas adicionales serán filtrados.

- **Date.of.Transfer**: indica la fecha de la transferencia. A partir de esta, se crean tres variables: el año, el mes y el día en que se realiza la transacción. Estas tres variables se utilizarán en el modelo de regresión, mientras que la fecha original es más útil para los modelos de series temporales.

- **Price**: indica el precio de cada transacción. Como se mostró en el análisis inicial de las variables, está fuertemente sesgada a la izquierda, por lo que es recomendable transformarla. En este caso, se ha aplicado la transformación Box-Cox, almacenando los nuevos valores en la variable *price_boxcox*. Esta será la nueva variable objetivo del modelo de regresión.

Además, se han creado variables para representar las relaciones económicas a partir de la variable objetivo, *Price*. Es importante señalar que no se cuenta con esta variable en el conjunto de test. A continuación, se describen las nuevas variables añadidas y los métodos utilizados para extrapolar esta información:

- **Cuartiles**: se estiman los cuartiles primero y tercero de los precios diarios en el conjunto de test. Posteriormente, se aplica el logaritmo de estos valores y se añaden como variables adicionales. Finalmente, se realiza un pronóstico mediante un modelo Random Forest,

como se menciona en [9]. La información del primer cuartil se almacena en la variable numérica *q1* y la del tercer cuartil en *q3*, y sus logaritmos en *log_q1* y *log_q3*.

- **Outliers**: a partir de los cuartiles, se calcula el rango intercuartílico, lo que permite identificar si un valor es un outlier superior o inferior. Si es un outlier superior, se indica en la variable binaria *is_top_outlier*, y si es un outlier inferior, en *is_bottom_outlier*.
- **Precio medio**: tras calcular el precio medio diario en el conjunto de entrenamiento y su logaritmo, se realiza un pronóstico para extrapolar los valores en el conjunto de test [9]. Esta información se almacena en *mean_price* y *log_mean_price*.
- **Evolución anual del precio**: en el conjunto de entrenamiento se calcula el precio medio anual por distrito y tipo de vivienda. Luego, se calcula la variación porcentual entre cada año y el anterior, imputando el valor medio en el primer año. Para el conjunto de test, se agrupa la variable *mean_price* por año, siguiendo el mismo proceso. Además, al disponer de datos de años anteriores, se pueden calcular los cambios a partir del primer año (2021). La información se almacena en *year_to_year_change*, incluyendo su logaritmo.
- **Comparación del precio con Londres**: se calcula la media de los precios por distrito y se compara con los valores de la Ciudad de Londres. Para el conjunto de test, se utilizan los mismos valores que en el conjunto de entrenamiento, y los valores faltantes se imputan con la media. Esta información se almacena en *price_compare_london* y se incluye su logaritmo.
- **Corrección de efectos espaciales**: como se menciona en el apartado *Spatial fixed effects* de [10], se puede construir un modelo de mínimos cuadrados ordinarios (OLS) para estimar el efecto de cada distrito sobre la variable objetivo (*Price_boxcox*). Una vez estimado, esta información se añade a la variable *fixed_effects* tanto en el conjunto de entrenamiento como en el de prueba.

4.1.4. Escalado de las variables

Se ha utilizado el escalador de la librería *sklearn* para ajustar la mayoría de las variables[6], de manera que tengan media cero y desviación estándar unidad. Esto favorece que el modelo otorgue la misma importancia inicial a todas las variables.

4.1.5. Selección de variables

Para seleccionar las variables de interés para el entrenamiento de los modelos, se comenzará con un análisis de los coeficientes de correlación de Pearson. Podemos hacer un estudio preliminar observando la Figura 12, en la que se tienen múltiples casos de correlación alta entre las variables. En ella, se tiene como primera columna la variable objetivo *Price_boxcox*, mientras que el resto son las variables explicativas disponibles. Se observa que, mientras que tenemos múltiples variables con una correlación importante con la variable objetivo, también existen casos de multicolinealidad, es decir, de alta correlación entre variables explicativas. Esto es un indicativo

[6]Las que no han sido escaladas son las codificadas (ya sea por *label encoding* o *binary encoding*), ni las variables referentes a la fecha de la transacción.

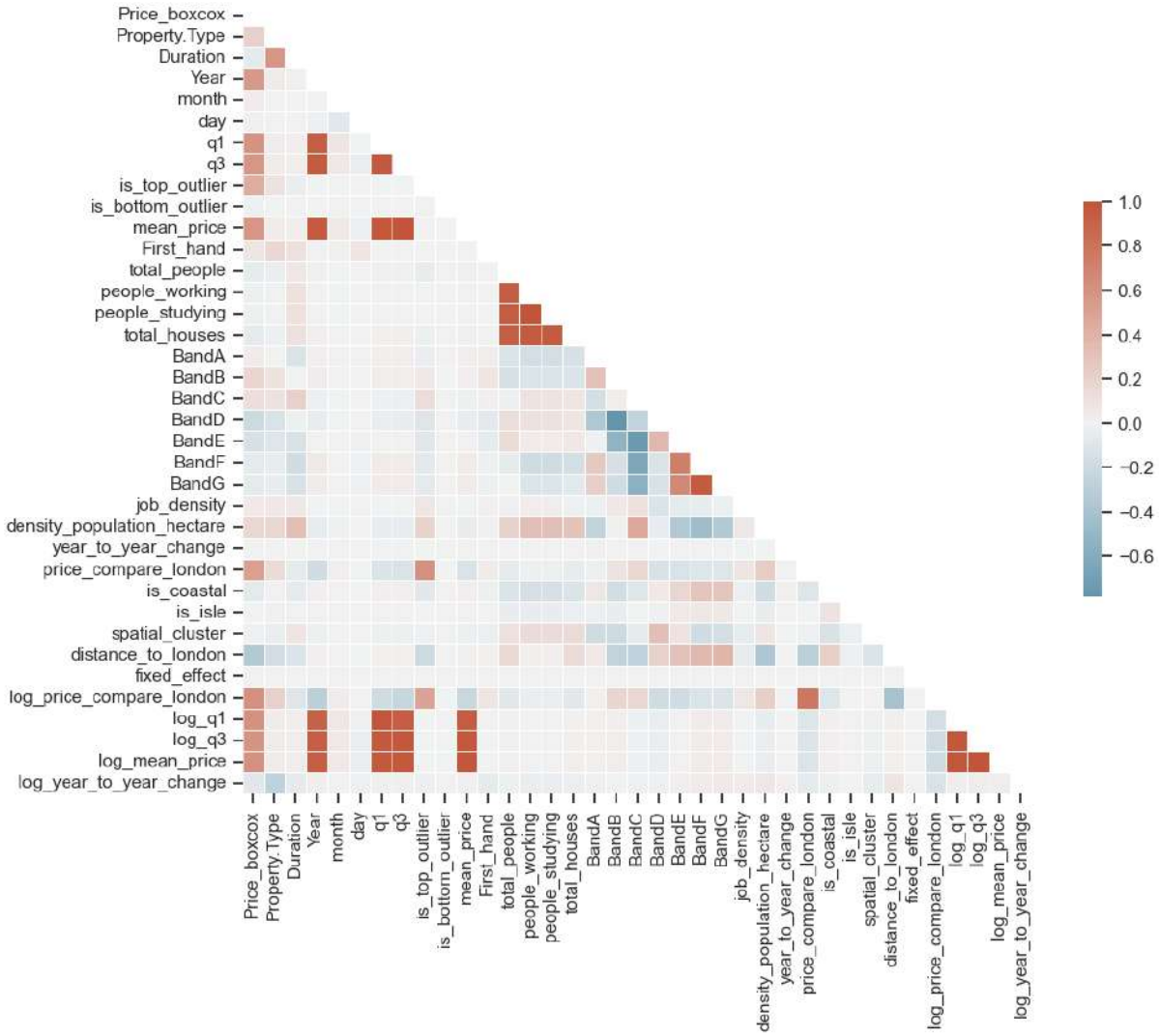

Figura 12: gráfica de correlación entre las variables. La correlación positiva se representa en rojo y la negativa en azul. Se ha añadido la variable objetivo en la primera columna para descartar las variables que no la expliquen, es decir, que muestren un cuadrado blanco.

de redundancia en los datos y pueden perjudicar enormemente el rendimiento de los modelos lineales que se construyan.

A continuación, estudiando la correlación con la variable objetivo, como podemos ver en la Figura 13, se decide qué variables se utilizan para el entrenamiento de los modelos de regresión. Se aprecia que existen dos umbrales en cuanto al valor del coeficiente, uno para las variables correladas negativamente en *BandE* y otro para las variables correladas positivamente en *is_top_outlier*.

No obstante, este análisis se restringe a las relaciones lineales entre las variables, por lo que no aportan información suficiente para decidir qué variables usar en caso de entrenar modelos no lineales. Por tanto, para abordar este problema se hace un estudio de la importancia de las variables con un modelo de Random Forest, como puede verse en la Figura 14. Las conclusiones extraídas de este estudio son:

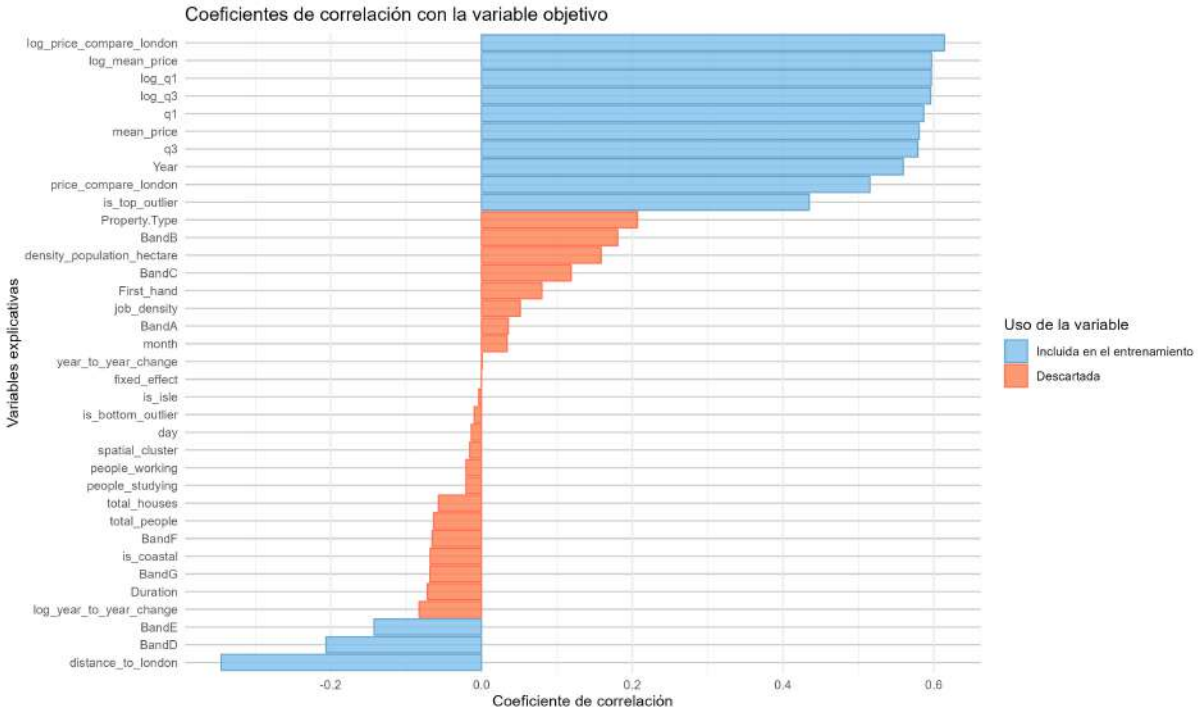

Figura 13: coeficiente de correlación de Pearson de las variables listadas en el eje vertical con la variable objetivo, *Price_boxcox*. Se aprecia que para los valores de correlación positiva se tiene un descenso pronunciado en *is_top_outlier* que podemos tomar como umbral y, para la correlación negativa, ocurre algo análogo a partir de *BandE*.

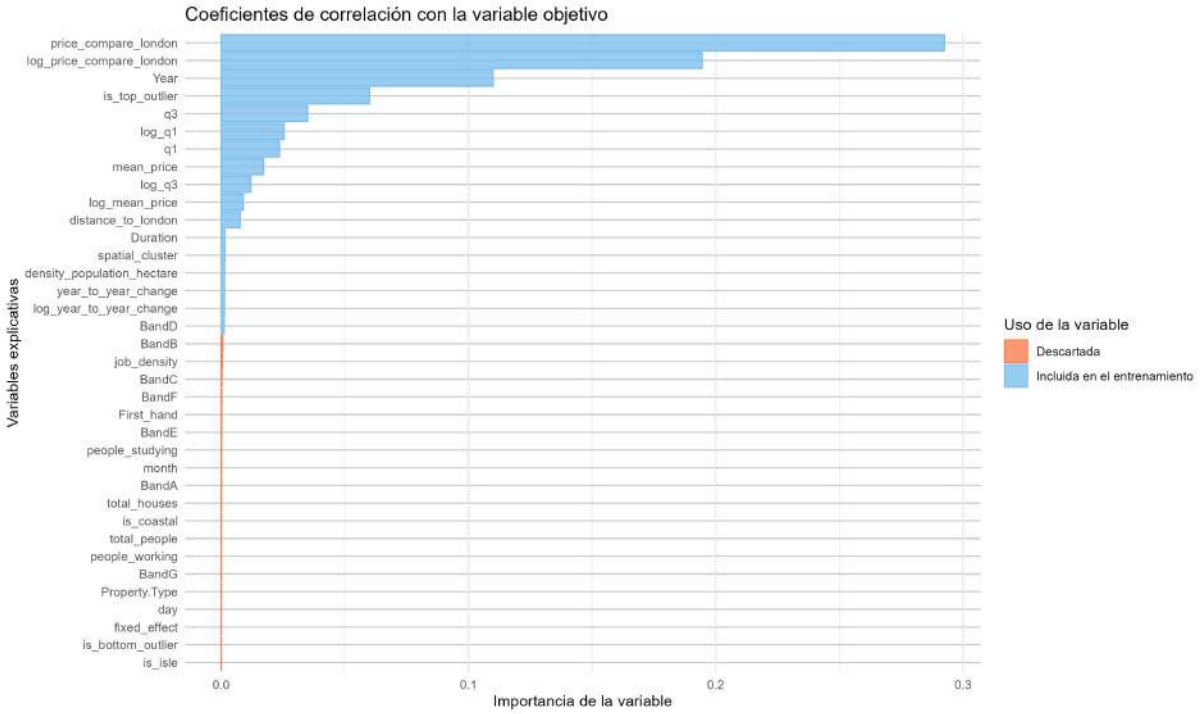

Figura 14: cálculo de la importancia de las variables realizado con un modelo de Random Forest. La principal aportación al modelo la hacen las variables con información económica.

- Las variables con mayor importancia son aquellas que almacenan información económica y el año de la transacción.
- En la Figura 12, se observa que las variables referentes a los cuartiles de los precios y al precio medio, están altamente correladas. Más adelante, en la Figura 13, se muestra que las variables *q1*, *log_q1* y *log_q3*, están altamente correladas con la variable objetivo. Y, finalmente, con este último resultado se aprecia que, de las tres anteriores, la variable *log_q1* es la de mayor importancia. Por tanto, para la construcción de modelos lineales, se usará *log_q1*, descartando, por tanto, las variables *q1*, *q3*, *log_q3*, *mean_price*, *log_mean_price*.
- Las variable *price_compare_london* y su logaritmo están altamente correladas, por lo que, para modelos lineales, se usará únicamente *price_compare_london*, porque tiene una correlación similar a su logaritmo con la variable objetivo, pero presenta una importancia mucho mayor.

Cabe destacar que el método para evaluar la importancia consistía en entrenar un modelo Random Forest con todas las variables disponibles[7] y evaluar el error de las predicciones con el propio conjunto de entrenamiento (MSE) al mezclar aleatoriamente cada una de las variables por separado. De esta forma, un aumento considerable del error nos indicaría que la variable mezclada aleatoriamente era fundamental para el rendimiento del modelo.

En adición a lo anterior, se pueden deducir ciertas propiedades de los datos a partir de los resultados obtenidos:

- Como se sospechaba, la distancia a Londres representa relativamente bien el descenso del precio de las transacciones. De ahí que sea la que mayor correlación negativa presenta.
- Las correlaciones de los niveles de eficiencia energética con la variable objetivo, denotan que el cambio más favorable para una vivienda es del nivel 'D' al nivel 'C', ya que se pasa de una correlación negativa a una positiva.

Por último, en caso de usar modelos no lineales, como Random Forest o XGBoost, la selección de variables se basa únicamente en su importancia. Concretamente, solo mantenemos las variables desde *price_compare_london* hasta *job_density*, ya que a partir de esta, se tiene una importancia prácticamente nula. Además, para este tipo de modelos la multicolinealidad no es problemática y, de hecho, son capaces de extraer información de todas estas variables, potenciando así su rendimiento.

4.1.6. Información sobre los parámetros de los modelos utilizados

El método utilizado para estimar los hiperparámetros comienza con la construcción de un modelo con sus parámetros por defecto para tomar el error obtenido de evaluar el conjunto de validación como referencia. A continuación, se hace una búsqueda en rejilla (o *grid search*) y

[7] Concretamente, se ha usado un modelo de Random Forest de la librería *cuml* con los hiperparámetros n_estimators = 100, max_depth = 10.

seleccionamos el modelo con menor error en validación registrado.

Los modelos que se presentan a continuación, fueron entrenados en GPU mediante la librería de *RAPIDs*y *XGBoost* en Python, junto con otras necesarias para trabajar en GPU como son *cupy*, *cudf* y *cuml*.

Además, dadas las limitaciones de almacenamiento que emergen del uso de la GPU, el entrenamiento se ha realizado utilizando una muestra aleatoria del 5 % de los datos, salvo en el caso de algunos de los modelos de Random Forest, para los que se ha necesitado reducir aún más el número de registros. No obstante, se ha evaluado tanto en el conjunto de validación como la totalidad de los datos de test[8].

Modelo de regresión lineal
Los hiperparámetros ajustados de los modelos de regresión lineal fueron[11]:

- **fit_intercept**: añadir un término independiente o no. Es importante en caso de no haber centrado los datos.
- **normalize**: aplicar o no un escalado previo de las variables explicativas.
- **algorithm**: método computacional para resolver el problema de regresión lineal.

Modelo de regresión Elastic-Net
Ampliación del modelo de regresión lineal básico que incorpora parámetros de penalización L1 y L2. Los hiperparámetros ajustados de los modelos de regresión lineal fueron[12]:

- **alpha**: constante que multiplica al término L1. Si se toma como 0, el modelo es equivalente a uno de mínimos cuadrados ordinarios.
- **l1_ratio**: parámetro de mezcla del modelo Elastic-Net entre la penalización L1 y L2. Si se toma como 0, la penalización es L2 y si se toma como 1, la penalización es L1.
- **tol**: tolerancia de la optimización.

Modelo de regresión Random Forest
Modelo no lineal. Los hiperparámetros ajustados de los modelos de Random Forest fueron[13]:

- **n_estimators**: número de árboles de regresión.
- **max_depth**: máxima profundidad de los árboles.

[8]Para ello, se particionan los datos y evaluamos cada una de ellas. Después tomamos la media de las métricas obtenidas.

Mínimo error sobre validación (MSE)			
Modelos lineales		**Modelos no lineales**	
Regresión lineal	**Elastic-Net**	**Random Forest**	**XGBoost**
1,24	0,71	0,38	0,39

Cuadro 3: mínimo error obtenido tras la optimización de hiperparámetros para los modelos de regresión.

Modelo de regresión XGBoost
Modelo no lineal. Para encontrar el modelo XGBoost óptimo, se realizó una búsqueda de parámetros. En particular, se ajustaron los siguientes parámetros[14]:

- **subsample**: porcentaje de los datos de entrenamiento utilizado en cada iteración de boosting para evitar el sobreajuste.
- **colsample_bytree**: porcentaje de características seleccionadas para el entrenamiento de cada árbol.
- **max_depth**: máxima profundidad de los árboles.
- **learning_rate**: tasa de aprendizaje, utilizada para reducir el tamaño del paso del gradiente en cada iteración y prevenir el sobreajuste.

Conclusiones
A modo de resumen, se ha registrado en el Cuadro 3 el MSE mínimo obtenido tras la optimización de los hiperparámetros de los modelos de regresión. Se observa una mejora significativa al considerar efectos no lineales en el modelo, obteniéndose para el modelo Random Forest y XGBoost resultados muy similares. No obstante, la memoria que ocupa el modelo de Random Forest es muy superior, por lo que optamos finalmente con el modelo XGBoost.

Finalmente, vamos a evaluar el modelo elegido (XGBoost) sobre el conjunto de test completo. Además, como podemos ver en el Cuadro 4 en caso de limitar la precisión geográfica o la temporal, se aprecia que el mejor rendimiento siempre ocurre cuando nos restringimos a los distritos. Además, la diferencia entre el rendimiento mensual y el anual es tan pequeño que no merece la pena restringirse a una predicción anual de los precios.

4.2. Series temporales

Uno de los factores principales al trabajar con series temporales es la elección de la precisión temporal con la que queremos trabajar. En este proyecto, los datos tienen granularidad diaria, pero hay que estudiar si esto afecta negativamente al modelo (por ejemplo, podría introducirse un ruido que dificulte la detección de tendencias estacionales). Además, si por el contrario se elige una granularidad demasiado alta, también se perdería información.

Modelo XGBoost evaluado en test		
	R^2 **medio**	MSE **medio**
Viviendas concretas	0,56	0,32
Pueblos y ciudades (diario)	0,57	0,14
Distritos (diario)	0,67	0,14
Condados (diario)	0,53	0,27
Pueblos y ciudades (mensual)	0,55	0,32
Distritos (mensual)	0,83	0,20
Condados (mensual)	0,71	0,07
Pueblos y ciudades (anual)	0,56	0,19
Distritos (anual)	0,86	0,04
Condados (anual)	0,74	0,06

Cuadro 4: resultados de evaluar el modelo XGBoost de regresión sobre el conjunto de test completo, es decir, los datos entre inicios de 2021 y abril de 2023. El primer caso es la evaluación del acierto en la predicción del precio de viviendas concretas. Los siguientes son el acierto en la predicción del precio medio para un cierto grado de precisión geográfica y temporal.

4.2.1. Análisis de la granularidad de la escala temporal

Para comenzar el estudio, se muestra en la Figura 15 una comparación entre las tres posibles escalas temporales. En ella, resalta la gran dispersión que presenta la escala diaria frente a la mensual y la anual, mientras que la escala anual destaca por capturar vagamente la tendencia, lo que podría ser indicativo de que se está perdiendo información.

Mediante el cálculo de la media, varianza y desviación estándar de los datos, podemos denotar la diferencia en la dispersión de los datos. Esto queda representado en la Figura 16, donde vemos que, aunque el valor medio es similar en los tres casos, la desviación estándar y la varianza presentan una diferencia apreciable, a pesar de la escala logarítmica. No obstante, no es tan elevado como para descartar el uso de la escala diaria.

Otro factor a tener en cuenta para elegir la escala adecuada es que los patrones temporales de los datos queden representados correctamente. Esto puede estudiarse mediante el análisis de las autocorrelaciones y de las autocorrelaciones de la serie temporal diaria, ya que nos indicará algunos de los patrones existentes. Esto queda representado en la Figura 17, de la que se extrae la siguiente información:

- Se aprecian oscilaciones en la parte ACF (autocorrelación). Si contamos el número de picos en el intervalo entre el cero (podemos considerarlo el primer pico aproximadamente) y el retraso 90 (el último) se obtienen 14 picos, de lo que extraemos que hay uno cada 6 o 7 días, denotando una tendencia semanal en los datos.
- No obstante, también vemos que el decrecimiento de la parte ACF es muy lento, denotando que existen tendencias a largo plazo.
- Para la parte PACF (autocorrelación parcial) se tiene un valor especialmente alto en el retraso 1, lo que significa que el precio medio de un día concreto está altamente correlado con el del día anterior.

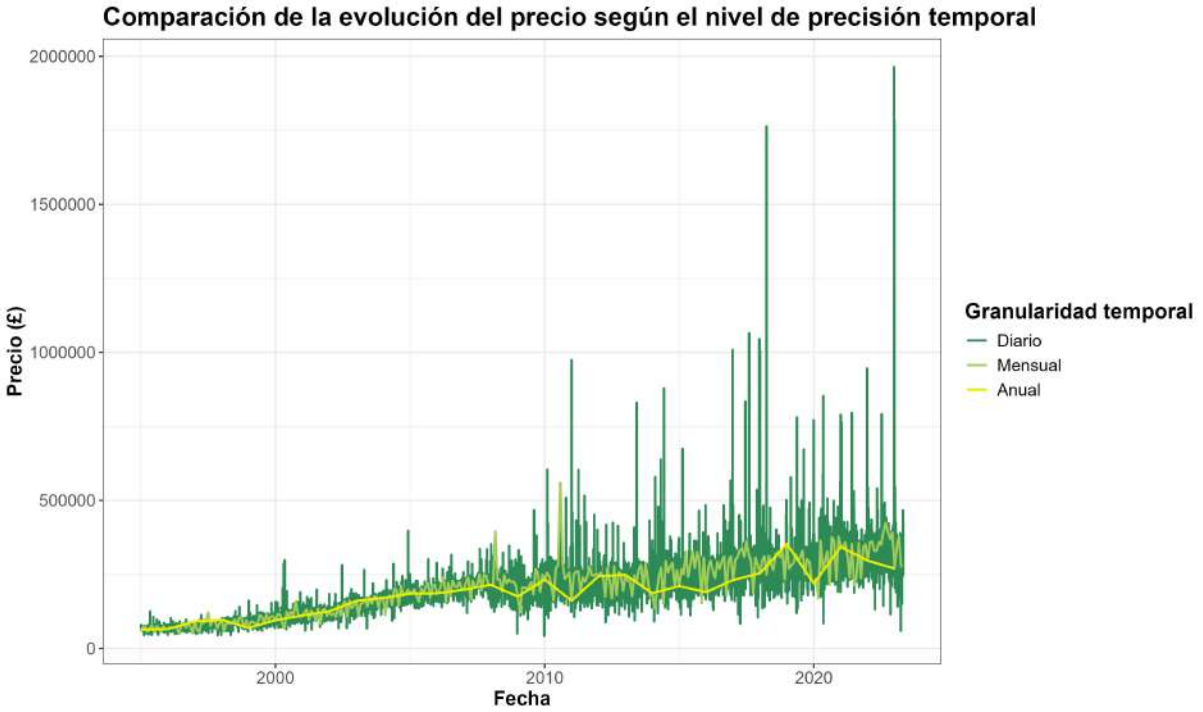

Figura 15: comparación de la evolución del precio en función de la precisión temporal. Podemos ver que la dispersión de los datos diarios es muy superior a la de los datos mensuales y anuales.

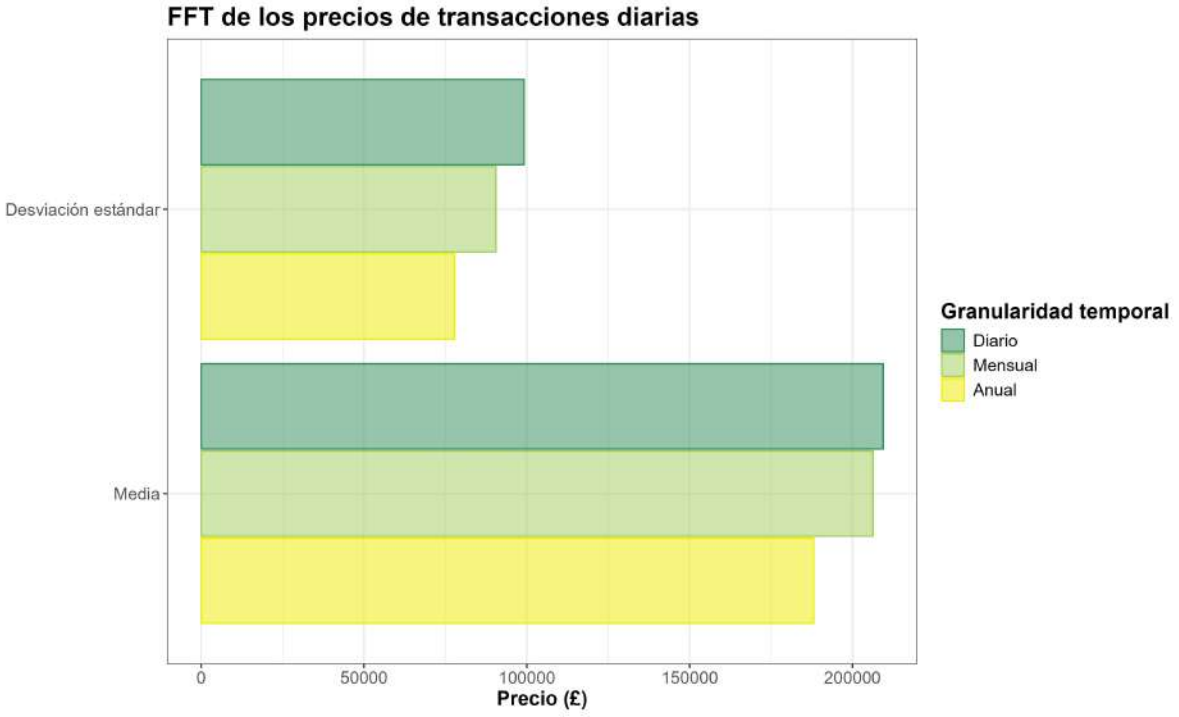

Figura 16: comparación de la media, la varianza y la desviación estándar de los datos en las distintas precisiones temporales. En los tres casos se obtienen resultados similares, aunque a menor precisión, menor dispersión se observa.

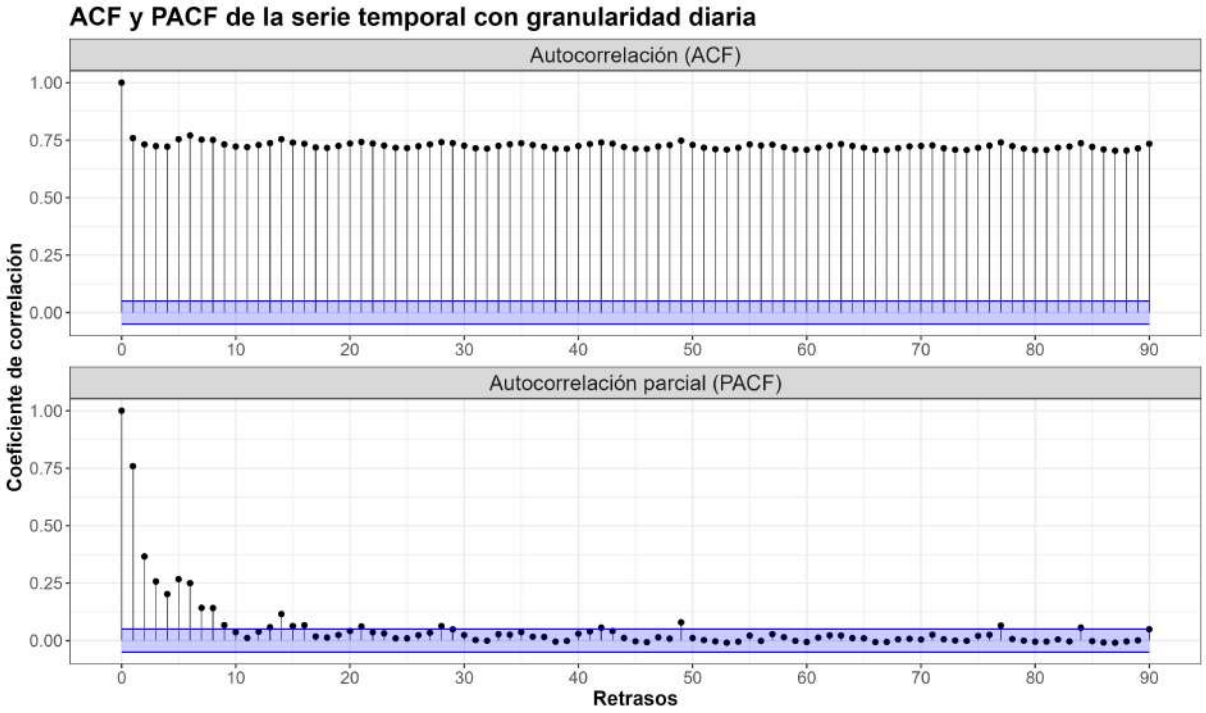

Figura 17: gráficos de autocorrelación y autocorrelación parcial para la serie temporal de los precios diarios. Se observa un decrecimiento lento en la función de autocorrelación y oscilaciones en ambas.

En definitiva, dada la ausencia de una periodicidad marcada en el término ACF y la existencia del pico en el retraso 1 para la parte PACF, se deduce que la escala temporal más adecuada para explicar patrones a corto plazo es la diaria. Sin embargo, también podríamos considerar la mensual, ya que suavizará los valores de la escala diaria, permitiendo dar mayor importancia a los efectos a largo plazo. Al mismo tiempo, se espera que el suavizado que introduce la escala anual limite la capacidad de predicción del modelo, especialmente, si existen otros efectos a corto plazo similares al patrón semanal encontrado.

Finalmente, se complementa este estudio con el análisis de la serie de Fourier que corresponde a esta distribución de precios. De esta forma, se obtendrán picos cuya posición en el eje horizontal indica la frecuencia correspondiente al periodo de tiempo del patrón al que corresponda según la expresión

$$\textbf{Periodo} = \frac{1}{\textbf{Frecuencia}},$$

y su altura nos indica la importancia que tiene este patrón en nuestros datos, es decir, a mayor altura del pico, mayor información aporta a la serie temporal. De esta forma, con la Figura 18 se obtiene la posición en el eje X de los picos más altos y, aplicando la expresión anterior, se deduce que los patrones más importantes de la serie temporal son: diarios (cada 2 o 3 días), semanales (cada 7 días), anuales (cada aproximadamente 365 días) y otros más tardíos, que aparecen debido a la existencia de tendencias a largo plazo, no a ciclos. Por tanto, para poder representar los patrones a corto plazo sin que el ruido que se introduce impida incluir los ciclos

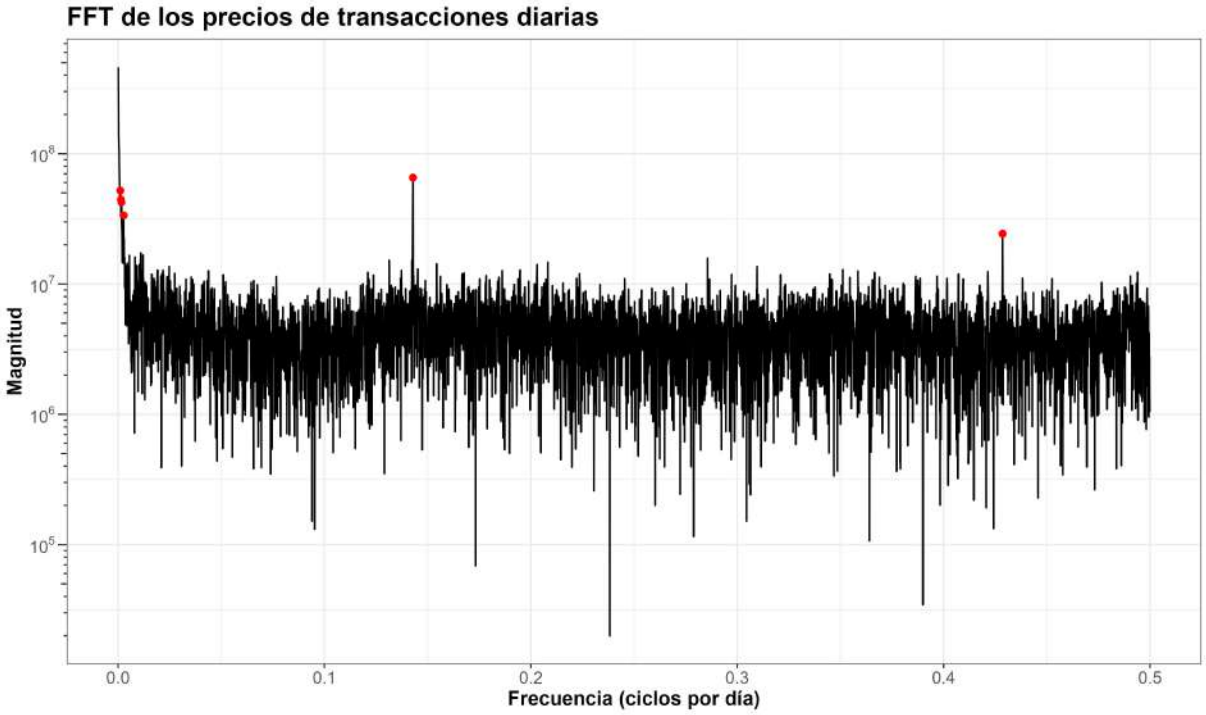

Figura 18: representación de las frecuencias más comunes en los datos a partir de su transformada de Fourier. El eje horizontal nos indica la frecuencia de los patrones y la altura del pico, lo común que es en los datos.

anuales, se opta por una precisión mensual de los datos.

4.2.2. Selección de hiperparámetros y construcción del modelo autorregresivo

Ahora que se ha concretado la precisión mensual de los datos, se van a proponer los hiperparámetros con los que construir un modelo autorregresivo[9]. Además, se busca incluir la estacionalidad que se ha observado, por lo que vamos a entrenar un modelo SARIMA. Por tanto, será relevante dar una estimación inicial de los siguientes parámetros:

- p: orden (número de retrasos o lags) de la parte autorregresiva del modelo.
- d: grado de diferenciación, es decir, número de veces que se restan los valores pasados a los datos.
- q: orden de la parte de la media móvil del modelo.
- P: orden de la parte estacional del modelo.
- D: grado de diferenciación de la parte estacional del modelo.
- Q: orden de la parte de la media móvil de la parte estacional del modelo.
- m: número de periodos en cada estación. Por ejemplo, para datos mensuales, si se una estacionalidad anual se elegiría $m = 12$.

[9]Un modelo autorregresivo se basa únicamente en el comportamiento pasado de la serie temporal para predecir el futuro.

Análisis de estacionariedad		
Serie temporal	**ADF test (p-valor)**	**KPSS test (p-valor)**
Original	0,83	0,01
Diferenciada 1 vez	$1{,}1 \cdot 10^{-12}$	0,10
Diferenciada 2 veces	$2{,}1 \cdot 10^{-14}$	0,10

Cuadro 5: estudio de la estacionariedad de la serie temporal original y despues de diferenciarla (1 y 2 veces). La hipótesis nula del test ADF se relaciona con la no estacionariedad de la serie temporal (y su alternativa con la estacionariedad). Para el test KPSS ocurre lo contrario, su hipótesis nula indica estacionariedad (y la alternativa no estacionariedad).

La estimación inicial de hiperparámetros se realiza con el conjunto completo de los precios mensuales disponibles (desde 1995 hasta 2023).

Análisis de estacionariedad: estimación del parámetro d
Una de las hipótesis con las que trabaja el modelo SARIMA es que la serie temporal sea estacionaria[10], lo que permite que los patrones predichos por el modelo se mantengan en el tiempo. No obstante, una serie temporal con tendencia o estacionalidad no puede ser estacionaria.

En caso de que la serie temporal no sea estacionaria, se diferencia la serie hasta conseguirlo, lo que nos ayuda a identificar un valor adecuado para d. En el Cuadro 5, se muestran los p-valores del *Augmented Dickey-Fuller test* (ADF test) y del *Kwiatkowski-Phillips-Schmidt-Shin test* (KPSS test).

El test ADF toma como hipótesis nula que la serie temporal tiene una raíz unidad[11] (lo que indica que la serie temporal no es estacionaria) y como hipótesis alternativa que no la tiene. Por tanto, se observa que se rechaza que la serie temporal original sea estacionaria, pero, al diferenciarla al menos una vez, sí que se acepta.

Por otro lado, el test KPSS toma como hipótesis nula que la serie temporal presenta una tendencia estacionaria, mientras que la alternativa es que presenta una raíz unitaria. En consecuencia, se vuelve a rechazar la estacionariedad de la serie original y se acepta para las series diferenciadas.

En conclusión, basándonos en ambos resultados, es evidente la necesidad de diferenciar la serie temporal para garantizar su estacionariedad, que es un requisito para los modelos SARIMA. Además, se observa que una única diferenciación es suficiente, aunque aún no podemos descartar que una segunda diferenciación pueda mejorar los resultados.

Análisis de autocorrelación: estimación de los parámetros p y q
Vamos a analizar la función de autocorrelación (ACF) de la serie temporal para estimar el parámetro q y la de autocorrelación parcial (PACF) para el parámetro p. Empezando por la

[10]Una serie temporal estacionaria es aquella cuyas propiedades estadísticas (como la media, la varianza o la autocorrelación) se mantienen constantes en el tiempo.

[11]La presencia de una raíz unidad es indicativo de que la serie no es estacionaria. Esto se debe a que, al igual que ocurre con el *paseo aleatorio* [15], se acumula la incertidumbre en cada paso, provocando una desviación progresiva de la media original, en lugar de oscilar entorno a ella.

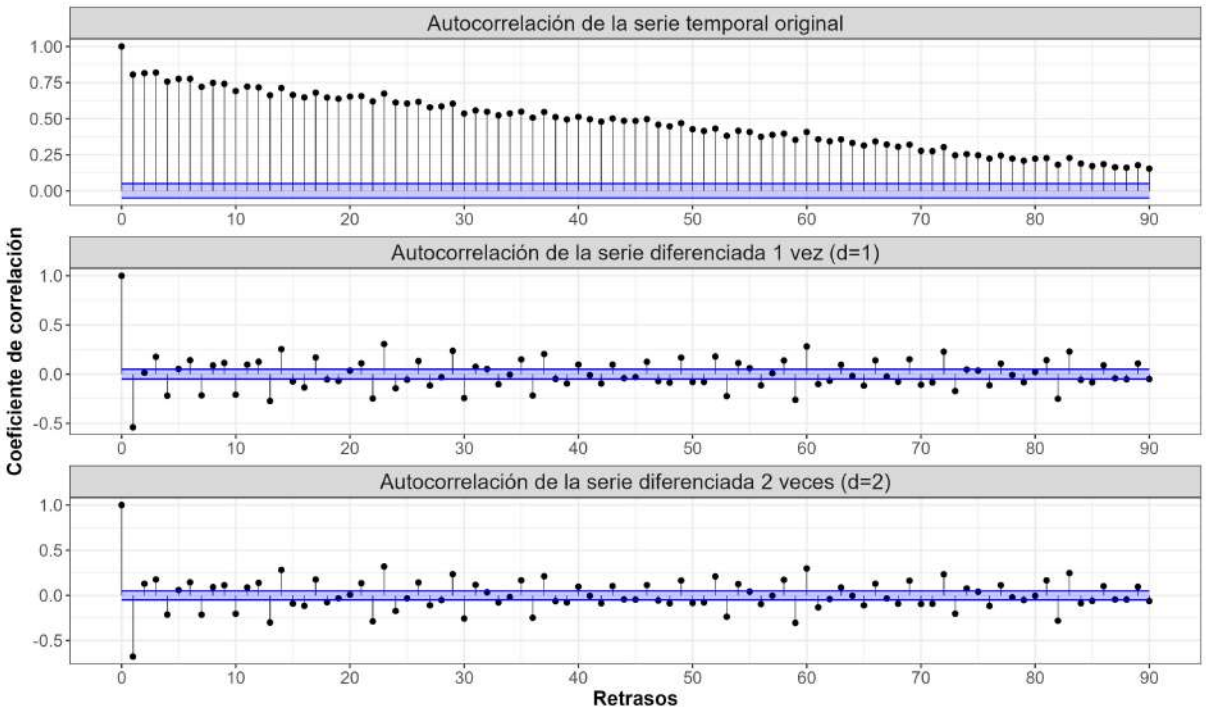

Figura 19: función de autocorrelación para la serie temporal original y diferenciada (1 y 2 veces). Se aprecia que el lento descenso de la serie original puede remediarse con una diferenciación. La segunda diferenciación hace más dispersos los valores, en comparación con la serie diferenciada una única vez.

autocorrelación de la serie temporal original y con 1 y 2 diferenciaciones, representada en la Figura 19, se aprecia que:

- El valor de $d = 1$ parece ser el correcto, ya que, al aplicar la primera diferenciación, los primeros retrasos (o *lags*) toman valores más cercanos a cero (a pesar de volverse negativos), pero al pasar a $d = 2$ los retrasos se alejan de cero, tomando valores más negativos.
- En la gráfica de la serie temporal con 1 diferenciación ($d = 1$), se aprecia que a partir del retraso 1 se toman valores aproximadamente nulos, lo que indica que $q = 1$ se adecúa a los datos.

Por otro lado, de la autocorrelación parcial, representada en la Figura 20, se extrae que:

- De nuevo, se observa la mejora para $d = 1$, mientras que para $d = 2$ los valores vuelven a alejarse de cero.
- Para $d = 1$, vemos un decaimiento pronunciado tras el segundo retraso, por lo que se estima que $p = 2$.

Además, en ambos casos se observa un decaimiento oscilante y valores que puntualmente toman valores más elevado que los de su entorno. Esto podría ser indicativo de que hay una parte estacional o una tendencia a largo plazo en los datos.

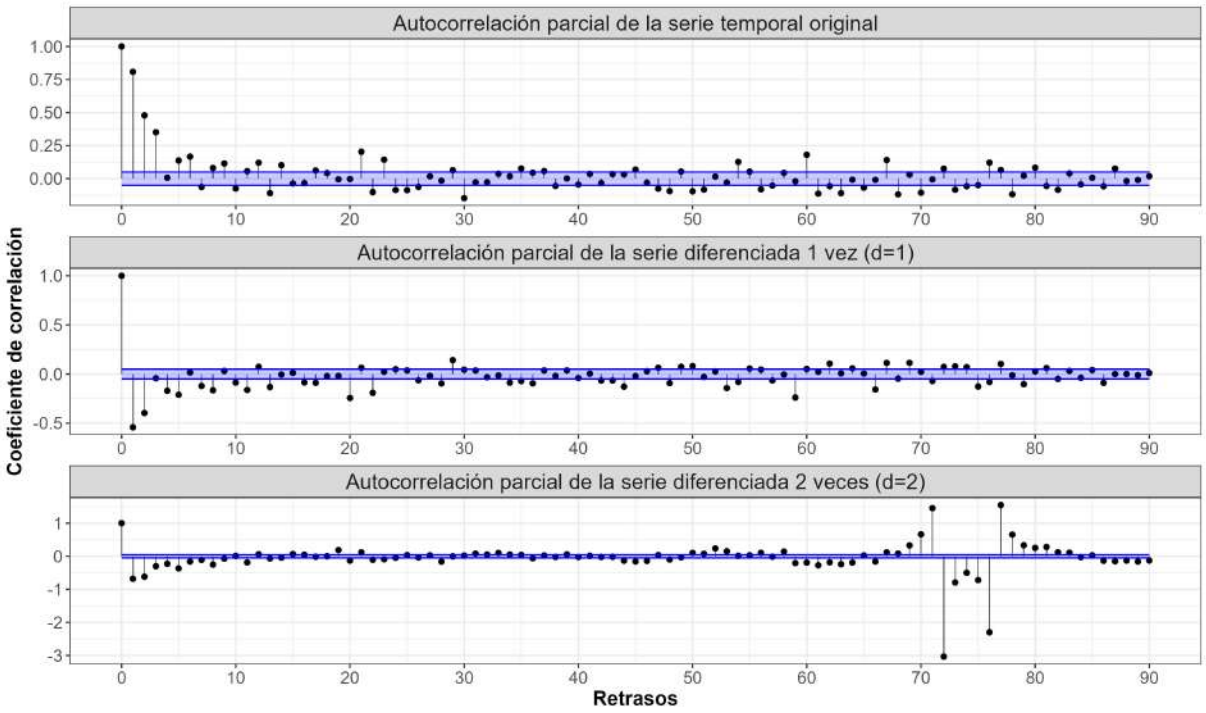

Figura 20: función de autocorrelación parcial para la serie temporal original y diferenciada (1 y 2 veces). Se aprecia que el descenso inicial de la serie original se vuelve más pronunciado con una diferenciación. La segunda diferenciación hace más dispersos los valores e introduce irregularidades, como se aprecia entre los retrasos 70 y 80.

Descomposición de la serie temporal: parámetros estacionales

A continuación, se descompone la serie temporal con diferenciación $d = 1$ en su tendencia, estacionalidad y residuos. Además, como es estacionaria, se realiza la descomposición aditiva (se tiene una varianza constante a lo largo del tiempo). El resultado puede observarse en la Figura 21, donde resulta especialmente llamativo el patrón que sigue la parte estacional de la serie estacional. Concretamente, se observa que la periodicidad es de aproximadamente 12 meses, que es la tendencia a largo plazo que se detectó mediante la descomposición en series de Fourier. Por tanto, para reflejar este comportamiento en el modelo, se elige $m = 12$.

De esta forma, solo quedan por encontrar las estimaciones de P, D y Q. Sin embargo, resulta más sencillo tomarlos como 1 y variarlos en función del rendimiento del modelo.

Además, antes de pasar a la búsqueda de hiperparámetros mediante la evaluación de modelos, podemos comprobar con los test de estacionariedad anteriores si el modelo es estacionario al diferenciar primero respecto del mes anterior ($d = 1$) y después respecto del año anterior ($m = 12$). Los resultados son:

$$\text{p-valor ADF test} = 2{,}1 \cdot 10^{-11} < 0{,}5 \rightarrow \textbf{Estacionaria},$$
$$\text{p-valor KPSS test} = 0{,}10 > 0{,}5 \rightarrow \textbf{Estacionaria}.$$

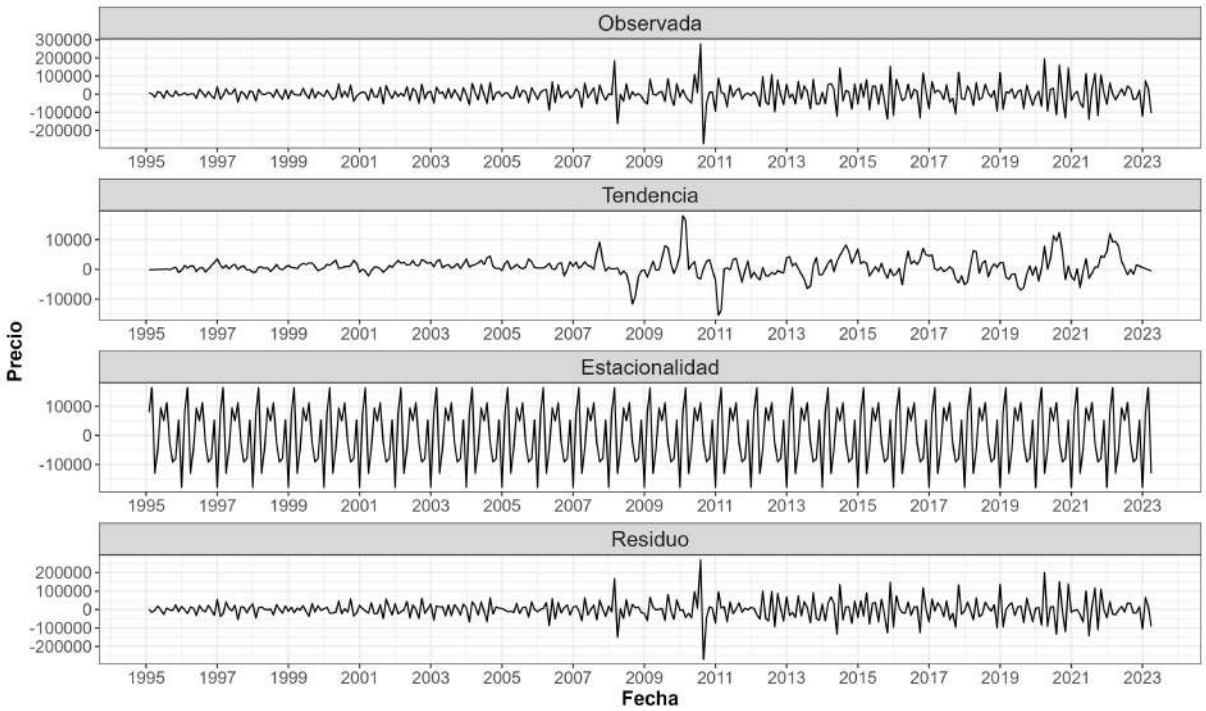

Figura 21: descomposición de la serie temporal diferenciada 1 vez en su tendencia, estacionalidad y residuo. Se aprecia un patrón de estacionalidad anual.

Optimización de hiperparámetros

Para continuar, se divide el conjunto de datos completo en un conjunto de entrenamiento (desde 1995 hasta diciembre de 2005), un conjunto de validación (desde 2006 hasta diciembre de 2015) y un conjunto de test (desde 2016 hasta abril de 2023). Con el conjunto de entrenamiento y el de validación se evaluarán modelos SARIMA de diferentes parámetros y se compararán según AIC y, de forma separada, según su MSE, para decidir cuál es el modelo óptimo.

Los parámetros de los modelos se dispondrán en un mallado de valores, que, gracias al estudio anterior quedan restringidos a un número manejable de casos.

Para la búsqueda de los hiperparámetros óptimos según el MSE del modelo, se ha usado la función *grid_search_sarimax*, de la librería *skforecast*. Se ha mantenido para todos los casos $d = 1$ y $m = 12$, salvo para aquellos en los que se han entrenado modelos sin estacionalidad. Por otro lado: para p se prueban valores entre 1 y 3 y para q, D, P y Q se prueban 0 y 1. De esta forma, el modelo con menor error encontrado con este método tiene:

$$(p, d, q) = (2, 1, 0) \text{ y } (P, D, Q, m) = (0, 0, 0, 0),$$

es decir, se ha descartado la componente estacional por completo.

Para la búsqueda de los hiperparámetros óptimos según el AIC del modelo, se ha usado la función *auto_arima*, de la librería *pmdarima*. Esta aproximación es mucho más rápida de entrenar, por lo que podemos proponer más opciones. En concreto: se ha dado libertad en la elección de d, D, P y Q, se ha fijado $m = 12$ y se prueban valores de p y q entre 0 y 3. De esta forma, el modelo

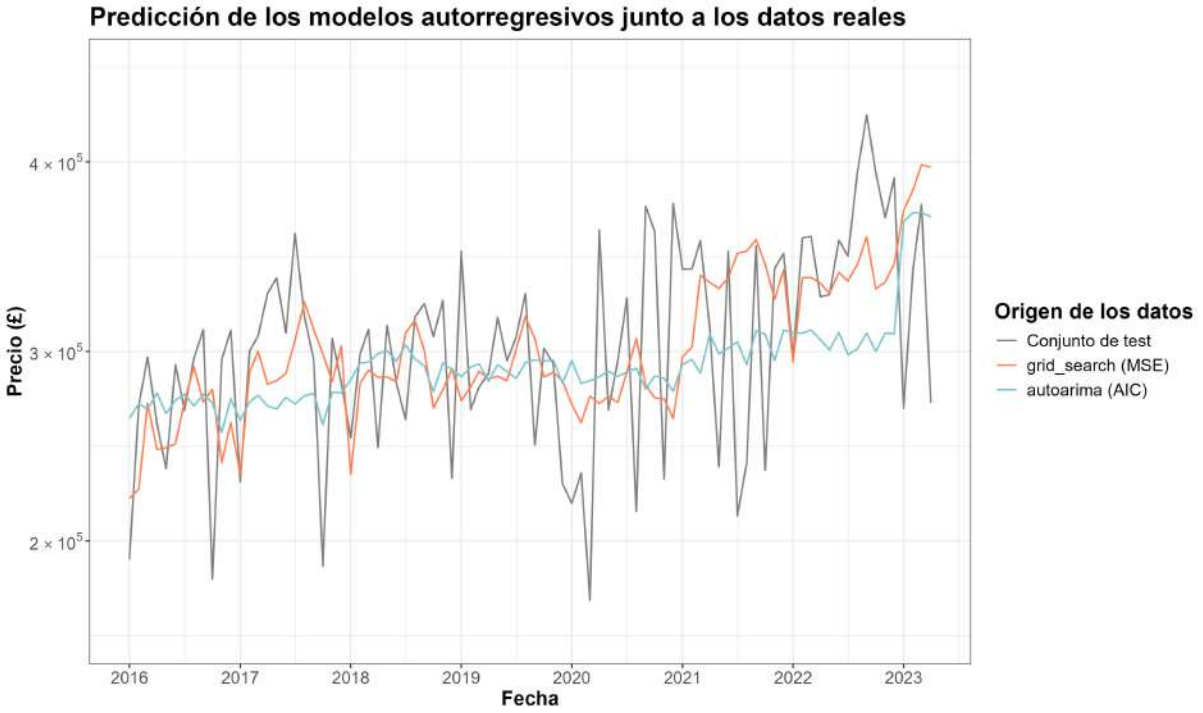

Figura 22: comparación del modelo obtenido mediante minimización de MSE (*grid_search*) y de AIC (*autoarima*) con los datos reales. Los resultados del *grid_search* se ajustan mejor en general a la forma de los datos.

Error en backtesting	
Modelo	**MSE**
Minimización de MSE (grid_search)	39200
Minimización de AIC (autoarima)	44400

Cuadro 6: comparación del MSE de los dos modelos óptimos planteados. Se han evaluado en datos futuros (*backtesting*).

con menor AIC encontrado es:

$$(p, d, q) = (0, 1, 1) \text{ y } (P, D, Q, m) = (0, 0, 1, 12).$$

Comparación de los modelos óptimos propuestos

Finalmente, se va a comparar ambos modelos según el MSE y utilizando *backtesting*. En nuestro caso, el conjunto de test que reservamos al inicio incluía los datos desde 2016 hasta abril de 2023, así que se evaluará sobre ese rango, como puede verse en la Figura 22. Concretamente vemos como el modelo obtenido mediante minimización de MSE (*grid_search*) se ajusta, en general, mejor que el obtenido mediante minimización de AIC (*autoarima*). Esto, además, queda denotado numéricamente en el Cuadro 6.

En definitiva, el modelo elegido es un modelo SARIMA de parámetros:

$$(p, d, q) = (2, 1, 0) \text{ y } (P, D, Q, m) = (0, 0, 0, 0),$$

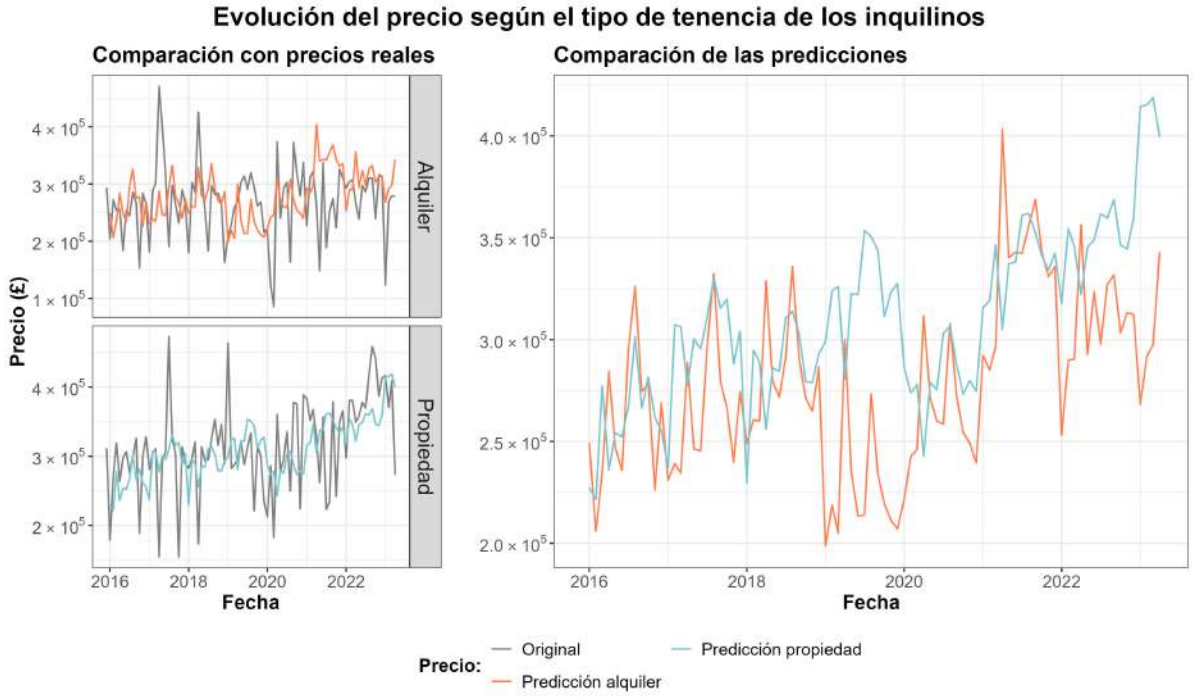

Figura 23: evolución del precio en función del tipo de tenencia de los inquilinos. A la izquierda tenemos la comparación con el precio real para esa categoría (representado en gris). A la derecha se comparan las predicciones obtenidas.

donde la ausencia de término estacional lo hace equivalente a un modelo ARIMA. Además, el error es de 4 órdenes de magnitud que, como se vio en la Figura 16, es un orden de magnitud inferior a la media de los precios mensuales, o lo que es lo mismo, se tiene un error de estimación de aproximadamente el 10 %.

4.2.3. Predicción de la evolución de categorías específicas

A partir de los resultados anteriores, se analizan por separado los distintos tipos de vivienda (alquiler, unifamiliares, de segunda mano...).

El procedimiento comienza con el entrenamiento de un modelo SARIMA con los parámetros definidos en el apartado anterior, pero con los datos filtrados. Por ejemplo, si el análisis se centra en las viviendas unifamiliares, se entrena el modelo con los registros de este tipo de vivienda entre 1995 y diciembre de 2015. Posteriormente, se evalúa sobre el conjunto de test, que, de nuevo, solo incluye los casos de viviendas unifamiliares. De esta forma, además de generar las predicciones para categorías específicas, se mide el rendimiento del modelo sobre estos casos.

Siguiendo esta idea, se ha representado a la izquierda de las Figuras 23, 24 y 25 una comparación entre la evolución de precio real de cada categoría junto a su predicción. A la derecha se tienen las predicciones de todas las categorías de cada variable, pero esto se comentará en la sección de resultados.

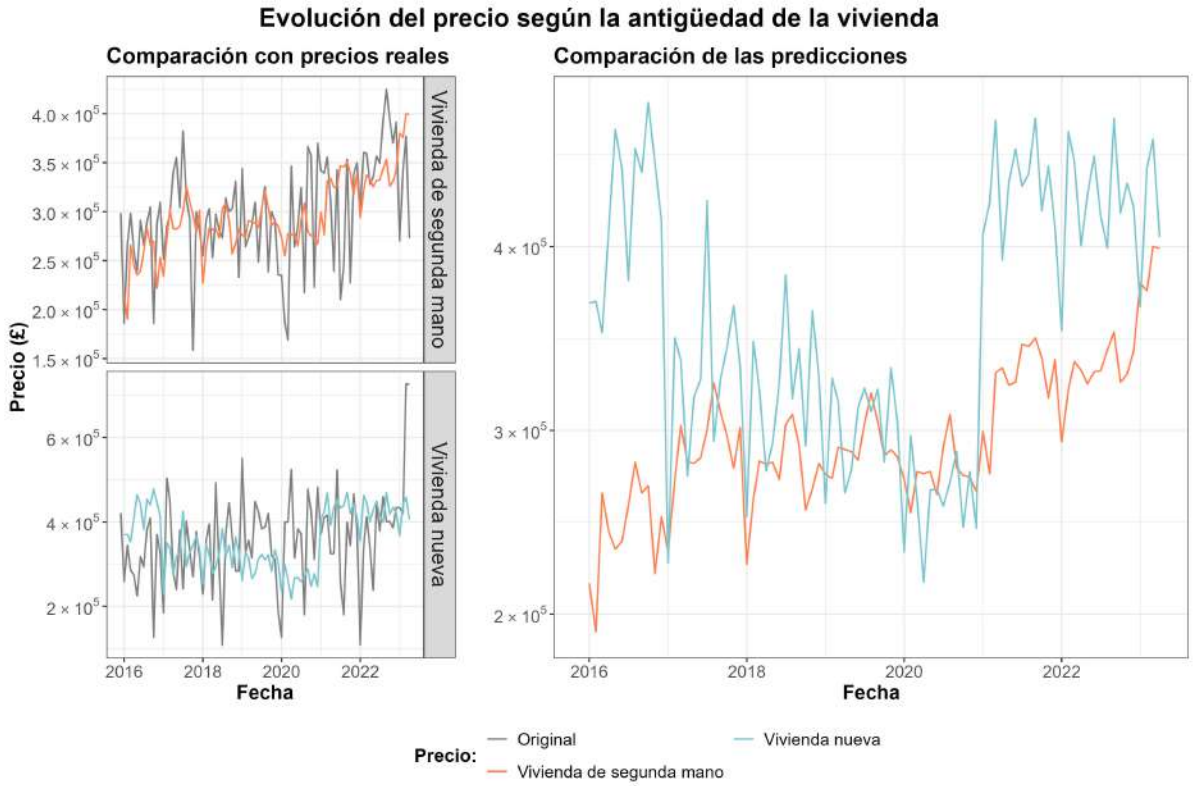

Figura 24: evolución del precio en función de la antigüedad de la vivienda. A la izquierda tenemos la comparación con el precio real para esa categoría (representado en gris). A la derecha se comparan las predicciones obtenidas.

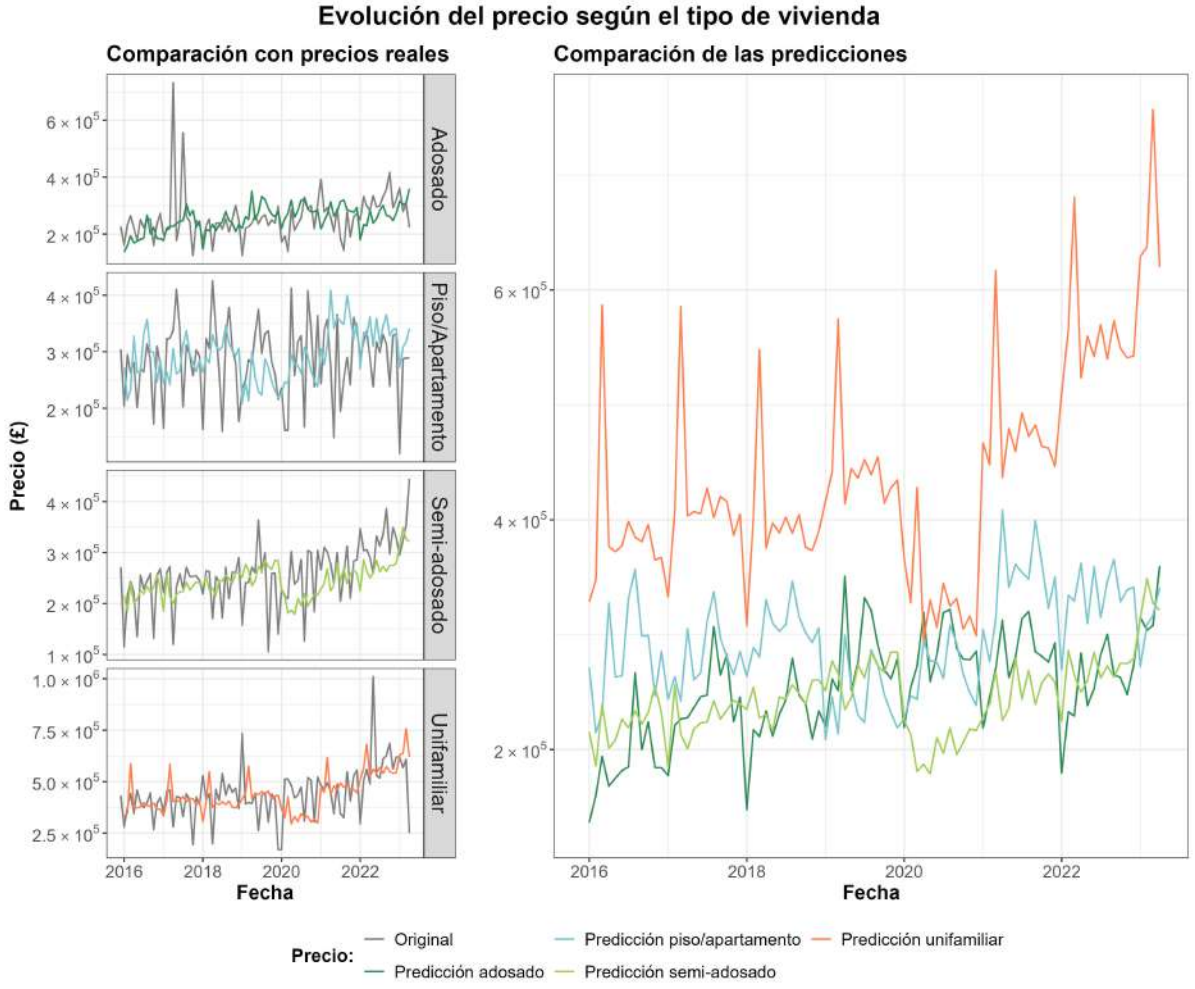

Figura 25: evolución del precio en función del tipo de vivienda. A la izquierda tenemos la comparación con el precio real para esa categoría (representado en gris). A la derecha se comparan las predicciones obtenidas.

Error en backtesting de categorías concretas	
Categoría	**MSE**
Contrato de alquiler	56200
Contrato de propiedad	47900
Vivienda de segunda mano	41700
Vivienda nueva	97300
Vivienda adosada	57600
Piso/Apartamento	59100
Vivienda semi-adosada	43300
Vivienda unifamiliar	93200

Cuadro 7: medidas del error (MSE) de estimación de cada categoría por separado. Se aprecia que las debilidades principales del modelo son las viviendas nuevas y las unifamiliares.

Volviendo a la comparación con el precio real, visualmente se aprecia que no todas se ajustan de manera precisa a los resultados reales. Se pueden considerar estos casos como debilidades del modelo. Además, en el Cuadro 7 se indican numéricamente las medidas del error, por lo que es sencillo determinar que los casos más problemáticos para el modelo son los de aquellas viviendas que son nuevas y aquellas que son unifamiliares. En la gráfica se identifica este problema por el alto rango de precios que toman estos tipos de viviendas.

5. Resultados principales

5.1. Modelo de regresión

El modelo de regresión desarrollado presenta un gran rendimiento en la predicción de precios medios de distritos. Estas estimaciones pueden realizarse con distintas precisiones temporales (precios medios por distritos diarios, mensuales o anuales), pero se considera que la más adecuada es la mensual, ya que la diaria tiene un rendimiento bastante menor y la anual no presenta una mejora significativa a cambio de la pérdida de información por pasar a esta escala.

Además, podemos usar el mismo modelo para predecir los precios de viviendas concretas, aunque, como se muestra en el Cuadro 4, el rendimiento en este contexto es considerablemente inferior. Se piensa que la razón de los resultados tan dispares radica en que, para los distritos, se disponía de información adicional mediante los datos censales, mientras que para las viviendas individuales solo se contaba con características básicas. Sería interesante explorar la posibilidad de incluir información más detallada, como los metros cuadrados de la vivienda y el número de habitaciones, ya que esto podría mejorar significativamente el rendimiento del modelo en la predicción de viviendas específicas.

En resumen, este modelo no solo permite un análisis actual de los precios medios por región, sino que también ofrece una predicción de cómo evolucionarán en los próximos años, lo que representa una ventaja considerable a nivel de inversión en el sector inmobiliario.

5.2. Series temporales: predicción de la tendencia del precio medio de las transacciones en Reino Unido

Para complementar las predicciones de precios realizadas con el modelo anterior, se han estimado las tendencias del precio medio de las transacciones en Reino Unido. Se ha encontrado que con un modelo autorregresivo puede representarse con un error de aproximadamente el 10 % del valor estimado, lo que permite realizar predicciones relativamente precisas.

Después, se han entrenado modelo con los parámetros en datos más específicos, como viviendas con contrato de alquiler, de segunda mano o unifamiliares. De esta forma, se ha identificado que las categorías peor representadas por el modelo son las viviendas nuevas y las unifamiliares, ambas duplicando el error que presentan las otras categorías, como se muestra en el Cuadro 7.

Además, se han aprovechado estas predicciones para hacer una comparación entre la evolución de los precios de las distintas categorías, lo que ayudará a elegir en qué sector invertir. Esto queda representado en la parte derecha de las Figuras 23, 24 y 25, de las que se extrae la siguiente información:

- En todas las categorías se observa un crecimiento en el precio, aunque este difiere en términos de velocidad y estabilidad a lo largo del tiempo. En principio, las categorías más interesantes serán aquellas con un crecimiento rápido o que presenten inestabilidades, lo que permitiría invertir en momentos de baja del mercado y vender en momentos de alza. No obstante, la estabilidad en los precios ayuda en la estimación del retorno de inversión de la construcción de viviendas.
- En la Figura 23 se observa que la evolución de los precios de los contratos de propiedad tiene un crecimiento más estable que la de contratos de alquiler. Usando la información disponible, se podrían hacer contratos de alquiler que se renueven en los momentos de alza del mercado y evitar momentos en los que el precio baja drásticamente.
- En la Figura 24 lo más llamativo es el repentino crecimiento de las viviendas nuevas. Esto podría deberse al gran error de estimación que presenta esta categoría concreta, aunque cabe destacar que ocurre muy cerca de la pandemia de COVID-19 y del Brexit. Por otro lado, se observa un crecimiento relativamente rápido del precio de las transacciones de viviendas de segunda mano.
- En la Figura 25, se observa que las propiedades unifamiliares también presentan un incremento similar al de las viviendas nuevas (de la misma magnitud, aunque parece menor por la escala del eje vertical). De las otras tres opciones, las viviendas semi-adosadas son las que tienen un crecimiento más marcado, aunque las adosadas presentan una estabilidad muy interesante para el sector de la construcción.

En resumen, si se busca invertir en viviendas para venderlas, las más interesantes son las semi-adosadas y las de segunda mano. Si el objetivo es construir la vivienda para venderla, la opción de viviendas adosadas son las que mayor estabilidad de precio presentan. Finalmente, si se tienen

viviendas de alquiler, se observa que los precios de los contratos bajan a final de año y verano, por lo que sería óptimo esperar a meses intermedios para renovarlos[12].

6. Recomendaciones de uso

Los resultados obtenidos indican que el modelo propuesto es capaz de realizar predicciones adecuadas, por lo que el siguiente paso lógico es añadir al conjunto de entrenamiento datos actuales.

De igual forma, es necesario mantener actualizada la información del censo utilizada para garantizar una predicción lo más certera posible. Además, en lugar de imputar la información faltante de los datos del censo por el valor medio, podrían utilizarse los datos censales de años anteriores. De esta forma, distritos que hayan cambiado de nombre o se unan a otros quedarán mejor representados, favoreciendo al modelo de regresión

Otro factor a tener en cuenta es que, en caso de contar con datos más detallados sobre las viviendas, estas características podrían añadirse al modelo de regresión para mejorar su rendimiento en la predicción de precios de viviendas concretas.

También resultaría interesante incorporar estos resultados en una interfaz de consulta o una API. Esto permitiría una visualización de los resultados por distrito más dinámica.

Por último, aunque nos hemos restringido al dominio geográfico que planteaban los datos (Inglaterra y Gales), es relativamente directo añadir nuevas regiones, aunque hay que tener en cuenta que algunas variables como la distancia de un pueblo o ciudad a Londres puede dejaría de tener sentido.

7. Conclusión

Este proyecto demuestra como, mediante la ciencia de datos, es posible extraer información de conjuntos de datos inicialmente inaccesibles, y de construir modelos para predecir con precisión tanto el precio de las viviendas (por región o viviendas específicas) como las tendencias del mercado inmobiliario en Reino Unido. A través de la utilización de técnicas avanzadas de regresión y series temporales, se ha logrado estimar con un alto grado de precisión los precios medios por distrito, lo que ofrece un recurso valioso para la toma de decisiones en este sector.

El análisis ha destacado la importancia de factores geográficos y socio-demográficos, como la distancia a Londres o la densidad de población, así como la relevancia de contar con datos detallados y actualizados. La combinación de distintas fuentes de datos, como los precios históricos de Kaggle, los censos nacionales y la información geográfica, ha permitido crear un modelo adaptable a diferentes niveles de precisión geográfica. No obstante, se identificaron ciertas limitaciones en la predicción de precios de viviendas individuales, en gran parte debido a la falta de

[12] Además, como se mostró al inicio del documento en la Figura 5, el principal tipo de viviendas con contratos de alquiler son pisos y apartamentos.

información granular sobre las características específicas de las propiedades.

Para futuras mejoras, sería interesante incorporar datos adicionales, como el tamaño de las viviendas o el número de habitaciones, lo que podría aumentar la precisión de las predicciones a nivel individual. Asimismo, el uso de técnicas de imputación más avanzadas podrían mejorar aún más los resultados obtenidos.

En resumen, los modelos expuestos en este proyecto nos proporcionan una base sólida sobre la que seguir desarrollando en el análisis predictivo de precios inmobiliarios. Las predicciones obtenidas pueden ser usadas no solo para mejorar la eficiencia en la compra-venta de propiedades, sino también para identificar oportunidades de inversión con altos retornos en zonas geográficas clave.

Apéndice

Código utilizado

El código utilizado puede encontrarse en este repositorio de *GitHub*: `https://github.com/DomingoPC/tfm-codigo-adjunto`.

A modo de resumen: se ha trabajado con *Python* y se han hecho los gráficos con *R*. Los notebooks de *Python* incluyen un análisis general de los datos, la construcción de modelos de regresión con GPU y el análisis de series temporales.

Referencias

[1] W. O. Gibin, "Uk property price data 1995-2023-04," Disponible en https://www.kaggle.com/datasets/willianoliveiragibin/uk-property-price-data-1995-2023-04/data (19/10/2024).

[2] O. for National Statistics (UK), "Official census and labour market statistics," Disponible en https://www.nomisweb.co.uk/query/select/getdatasetbygeog.asp (19/10/2024).

[3] G. maps and data, "Mapa de reino unido," Disponible en https://gadm.org/download_country.html (19/10/2024).

[4] OpenCage, "Api de posición geográfica por nombre (versión de prueba)," Disponible en https://opencagedata.com/ (19/10/2024).

[5] Nominatim, "Api de posición geográfica por nombre (open source)," Disponible en https://nominatim.org/release-docs/develop/api/Overview/ (19/10/2024).

[6] G. de Reino Unido, "Descripción de las variables de los datos de kaggle," Disponible en https://www.gov.uk/guidance/about-the-price-paid-data (19/10/2024).

[7] "Google maps platform," Disponible en https://developers.google.com/maps/documentation/geocoding/overview?hl=es-419 (25/10/2024).

[8] K. Peek, "Getting distance between two points based on latitude/longitude," Disponible en https://stackoverflow.com/questions/19412462/getting-distance-between-two-points-based-on-latitude-longitude (22/10/2024).

[9] J. A. Rodrigo and J. E. Ortiz, "Skforecast: time series forecasting with python, machine learning and scikit-learn," Disponible en https://cienciadedatos.net/documentos/py27-time-series-forecasting-python-scikitlearn (22/10/2024).

[10] S. J. Rey, D. Arribas-Bel, and L. J. Wolf, "Geographic data science with python," Disponible en https://geographicdata.science/book/notebooks/11_regression.html#bringing-space-into-the-regression-framework (22/10/2024).

[11] cuml, "Parámetros linear regression," Disponible en https://docs.rapids.ai/api/cuml/stable/api/#linear-regression (02/11/2024).

[12] ——, “Parámetros elastic-net,” Disponible en https://docs.rapids.ai/api/cuml/stable/api/#elasticnet-regression (02/11/2024).

[13] ——, “Parámetros random forest,” Disponible en https://docs.rapids.ai/api/cuml/stable/api/#random-forest (02/11/2024).

[14] XGBoost, “Parámetros xgboost,” Disponible en https://xgboost.readthedocs.io/en/latest/parameter.html (22/10/2024).

[15] Wolfram, “Simple random walk simulation,” Disponible en https://demonstrations.wolfram.com/SimulatingTheSimpleRandomWalk/ (05/11/2024).