

UNIVERSIDAD DE BARCELONA

TRABAJO DE FIN DE MÁSTER

MÁSTER EN BIG DATA & DATA SCIENCE

**Efectos geoespaciales en la
modelización del precio de la
vivienda en la ciudad de Madrid**

Autores:

Andrea AZÁBAL
LAMOSO,
Fabio SANTAMARÍA
IGLESIAS

Profesor:

Miguel Ángel DE LA
LLAVE MONTIEL

20 de septiembre de 2021



UNIVERSITAT DE
BARCELONA



Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

Abstract

El objetivo de este proyecto consiste en realizar una estimación fiable del precio de la vivienda en la ciudad de Madrid conocidos los atributos de cada inmueble. Este tipo de estudios ha sido ampliamente realizado en el ámbito de la econometría, sin embargo, muy poca literatura recoge la influencia de los efectos espaciales en el poder predictivo de las distintas modelizaciones al incorporar información geográfica. En este trabajo, y con el propósito de estudiar más en profundidad estos efectos, se proponen no solo modelos de regresión lineal múltiples, sino también desarrollos que introducen la posible autocorrelación espacial tanto en la variable dependiente como en los residuos del sistema. De este modo, se busca alcanzar una especificación óptima, cuyas predicciones puedan ser aplicadas en el mercado inmobiliario.

Se tiene que los modelos más prometedores son el de retardo espacial (SAR) y el de error espacial (SEM), consiguiendo ambos romper las dependencias espaciales entre vecinos próximos, lo que lleva a concluir que los efectos espaciales son erróneamente subestimados y su incorporación aporta mejoras significativas. Las modelizaciones implementadas suponen un resultado esperanzador debido a su gran potencial, tanto a la hora de explicar la variabilidad del precio de la vivienda, así como herramienta de tasación.

Palabras clave: vivienda, precio, Madrid, econometría espacial, modelo de regresión lineal múltiple, modelo de retardo espacial, modelo de error espacial, modelo geográficamente ponderado, *Gradient Boost*, *SatScan*, *web scraping*.

Índice general

1. Objetivos.	7
1.1. Principales objetivos	7
1.2. Antecedentes	8
1.3. Contexto actual	8
1.4. <i>Stakeholders</i>	9
1.5. DAFO	10
2. Desarrollo.	13
2.1. Gestión del equipo de trabajo	13
2.2. Fuentes de información y <i>benchmark</i>	13
2.3. <i>Timing</i> del proyecto	15
2.3.1. Fases del desarrollo	15
2.3.2. Cronograma de hitos temporales	15
2.3.3. Análisis económico y <i>payback</i>	16
2.4. Hipótesis y planteamientos realizados	17
2.4.1. Regresión lineal múltiple (RLM)	17
2.4.2. Modelos de retardo espacial (SAR)	18
2.4.3. Modelos de error espacial (SEM)	18
2.4.4. Modelos geográficamente ponderados (GWR)	19
2.4.5. <i>Gradient Boosting</i> (GB)	19
2.5. Desarrollo y programación utilizada	20
2.5.1. Base de datos	20
2.5.2. Preparación del <i>dataset</i> Entrenamiento y validación	21
Información geográfica	22
2.5.3. Evaluación del modelo	25
2.6. Comparativas entre modelos	26
2.6.1. RLM	26
Evaluación de los modelos	27
2.6.2. SAR	29
2.6.3. SEM	30
2.6.4. GWR	31
2.6.5. <i>Gradient Boosting</i>	32
2.6.6. Comparativa final	35
3. Conclusiones	37
Conclusions	37
3.1. Soluciones planteadas y objetivos conseguidos	37
3.2. Aplicación real	37
3.3. Reflexión final: problemas y soluciones	38
3.3.1. Extracción de la base de datos	38
3.3.2. Obtención de información geoespacial	38

3.3.3. Problemas con la modelización	39
Evaluación con <i>SatScan</i>	39
Overfitting con <i>Gradient Boosting</i>	39
A. Variables de la base de datos	41
B. Análisis univariante	43
B.1. Variables originales	43
B.1.1. Variables numéricas continuas	43
B.1.2. Variables numéricas discretas	45
B.1.3. Variables dicotómicas	47
B.2. Transformaciones	48
C. RLM	51
C.1. Modelo lineal básico	52
C.2. Modelo lineal espacial	55
C.3. <i>Multiadaptive regression splines</i>	58
D. SAR	61
E. SEM	65
F. GWR	69
G. GB	73
Bibliografía	75

Siglas

DAFO Debilidades, Amenazas, Fortalezas, Oportunidades.

GB Gradient Boost.

GWR Geographically Weighted Regression.

MARS Multiadaptative regression splines.

RGPD Reglamento General de Protección de Datos.

RLM Regresión lineal múltiple.

SAR Spatial Autoregressive Model.

SaTScan Spatial Scan Statistics.

SEM Spatial Error Model.

Glosario

Aprendizaje automático Algoritmos de computación que mejoran de manera automatizada a través de la experiencia y el uso de datos.

Autocorrelación espacial Mide el grado en el que una variable geográfica está correlacionada con ella misma en dos puntos o zonas diferentes del área de estudio.

Benchmark Comparar productos, servicios y procesos de trabajo que pertenezcan a organizaciones que evidencien las buenas prácticas sobre el área de interés.

Cronograma Herramienta gráfica que presenta un detalle de las actividades que se deben desarrollar en los tiempos establecidos.

Econometría Ciencia que utiliza herramientas matemáticas y estadísticas para estimar las relaciones económicas.

Heterocedasticidad En estadística, se dice que un modelo presenta heterocedasticidad cuando la varianza de los errores no es constante en todas las observaciones realizadas.

Hipótesis nula Suposición que se utiliza para negar o afirmar un suceso en relación a un parámetro de una población o muestra.

Idealista Compañía española fundada el 4 de octubre del 2000 que ofrece a través de Internet servicios inmobiliarios en España, Italia y Portugal.

p-value Probabilidad de que un valor estadístico calculado sea posible dada una hipótesis nula cierta.

Significación estadística Un resultado es estadísticamente significativo cuando es improbable que haya sucedido debido al azar.

Tasa de aprendizaje Parámetro de un algoritmo de optimización que determina el salto en cada iteración en el avance hacia el mínimo de la función de coste.

Valor atípico u *outlier* Observación que es numéricamente distante del resto de los datos.

Web scraping Proceso de recopilar información de forma automática de la *web*.

Árbol de decisión Dado un conjunto de datos, se fabrican diagramas de construcciones lógicas que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema en el ámbito del aprendizaje automático.

Capítulo 1

Objetivos.

1.1. Principales objetivos

El principal objetivo de este trabajo es realizar una correcta modelización del precio del m^2 de la vivienda en la ciudad de Madrid (España). En una localidad tan diversa y sumamente poblada, el importe de la vivienda es muy susceptible tanto a la localización del inmueble como a los servicios que lo rodean (transporte público, hospitales, colegios, etc.), lo cual la convierte en una ciudad idónea para realizar un análisis de este tipo.

Para ello, en primer lugar, se evidenciará la necesidad de considerar la interacción espacial entre observaciones como otro elemento más en el estudio. Con este propósito, se plantea un análisis de los residuos de distintos modelos de regresión tanto desde el punto de vista clásico como desde un enfoque espacial, valorando aquellos métodos que consigan romper la heterocedasticidad y dependencia espacial que previsiblemente estén presentes en el precio de la vivienda. La finalidad es, por tanto, obtener un algoritmo robusto que consiga deshacerse de las fuertes dependencias espaciales y que, a diferencia de los modelos más simples, arroje predicciones acertadas.

Como punto de partida de nuestro análisis, contamos con una base de datos extraída del portal inmobiliario [Idealista](#), en la cual se recogen un total de 6000 observaciones de todo tipo de viviendas localizadas en la ciudad de Madrid. La información proporcionada incluye las principales características de cada vivienda, las cuales pueden clasificarse en variables internas o externas. Las primeras indican las características intrínsecas de la vivienda tales como la superficie en metros cuadrados o el número de habitaciones, mientras que las segundas comprenden la geo-localización almacenada en coordenadas geográficas. Para el estudio de la dependencia espacial, es esencial contar con la ubicación exacta de cada observación, ya que será un parámetro fundamental a la hora de poner de manifiesto los efectos de proximidad. Tras la extracción de las variables, se hará una mejora de nuestro conjunto de datos, limpiando posibles duplicados o valores erróneos en la muestra y tratando de identificar posibles actualizaciones en las características de un mismo inmueble.

Asimismo, durante todo el proceso de extracción, almacenado, limpieza y tratamiento de datos se respetará el Reglamento General de Protección de Datos (RGPD) y no se incluirá información personal en la base de datos ni en los resultados obtenidos en el estudio.

1.2. Antecedentes

Es posible, al trabajar con datos de corte transversal, encontrar los denominados **efectos espaciales** que se manifiestan a través de distintas dependencias entre observaciones con cierta proximidad geográfica. Estos efectos han sido ampliamente ignorados a lo largo de la historia por el hecho de que no pueden ser tratados por la econometría estándar, la cual se fundamenta en el análisis e interpretación de sistemas económicos con el fin de predecir variables tales como, por ejemplo, el precio de bienes y servicios.

Debido a la necesidad de resolver los problemas de origen geoespacial que la econometría estándar no puede solucionar, nació la "econometría espacial", término acuñado por Paelinck y Klaassen[1] y que hace referencia a las técnicas que tratan las consecuencias causadas por efectos espaciales en el análisis estadístico de modelos econométricos tradicionales[2].

En las últimas décadas, la importancia y relevancia de este tipo de análisis ha ido en auge, debido, en parte, a las cada vez más accesibles y extensas bases de datos geo-referenciados, así como al incremento de la capacidad de computación de modelos cada vez más complejos. La definición de econometría espacial ha ido refinándose y ampliándose con el paso de las décadas[3], siendo los estudios metodológicos más relevantes aquellos desarrollados por Cliff y Ord[4], Anselin[5] y el ganador del Premio Nóbel de Economía Paul Krugman[6].

Centrándose, en particular, en el ámbito del mercado inmobiliario, se encuentran diversos trabajos de modelizado del precio de la vivienda, siendo el más notable el realizado por Dubin (1998)[7], en el cual se realiza por primera vez un análisis teórico en este marco, introduciendo el concepto de modelos autorregresivos y explorando las dificultades involucradas en la estimación de dichos modelos cuando estos presentan autocorrelaciones espaciales en los términos de error. Otros trabajos similares son los elaborados por Pace[8] o Basu y Thibodeau[9].

1.3. Contexto actual

El mercado inmobiliario en España ha sufrido grandes altibajos durante las últimas décadas. Tras el mayor "parón" inmobiliario de la historia de España, producido en el año 2008, hubo un cambio de ciclo en el que los posibles compradores descendieron significativamente y además se tornaron más selectivos[10] [11].

La situación fue mejorando durante la década de los 2010, aunque a partir de 2020 se observa de nuevo una ralentización tanto en la subida del precio de la vivienda como en el volumen de compraventas. Este hecho parece indicar que pueda repetirse una situación similar a la de la crisis de 2008, en la cual el comprador sea reticente a tomar una decisión arriesgada y prefiera informarse adecuadamente.

Es, por tanto, el momento idóneo para proporcionar herramientas de análisis al comprador que le ayuden a tomar una decisión informada y acertada a la hora de adquirir una vivienda. Una herramienta de modelizado del

precio del metro cuadrado como la nuestra, cuyas predicciones sean robustas ante efectos espaciales y, por consiguiente, aporten mayor fiabilidad, es justamente lo que el demandante de vivienda necesita. A su vez, también se trata de un valioso y potente recurso para el sector empresarial, ya que aporta beneficios tales como una correcta valoración o tasación de inmuebles, que además puede descomponerse por características y determinar la aportación de cada una de ellas al precio total de cada vivienda.

Así, en España en general y en la ciudad de Madrid en particular, son pocos los trabajos actuales que modelizan el precio del metro cuadrado teniendo en cuenta las particularidades geo-espaciales. La ciudad de Madrid es, sin duda, una de las que más variabilidad presenta en el precio de la vivienda entre distritos o barrios, lo cual la convierte en una elección interesante para este tipo de análisis.

Trabajos previos han realizado estudios similares en España[12], e incluso existen algunos recientes que se centran en la ciudad de Madrid[13][14]. Sin embargo, en ninguno de ellos se toma el mismo enfoque ni se utilizan modelos predictivos aplicando técnicas de aprendizaje automático o *machine learning*, como es nuestra intención.

1.4. Stakeholders

Según la definición proporcionada en su primer uso[15], un *stakeholder* o actor clave es un "miembro de cualesquier grupos cuyo apoyo la empresa o –en nuestro caso– proyecto necesita para sobrevivir". Posteriormente, el concepto se fue extendiendo para incluir no solo las partes indispensables, sino también todos los grupos interesados en la iniciativa.

Un correcto análisis de actores clave durante la planificación de un proyecto ayuda a determinar las necesidades y requerimientos básicos del mismo, así como a identificar posibles riesgos o conflictos.

En el ámbito que nos compete, se puede realizar la distinción entre actores clave **internos** (forman parte del proyecto), y **externos**:

- Actores clave externos:

Por una parte, los principales interesados en un correcto modelizado del precio de la vivienda son los demandantes de bienes inmuebles o potenciales compradores. Conocer sus inquietudes nos ayudará a determinar los factores clave a la hora de realizar un análisis del mercado inmobiliario.

De manera análoga, los oferentes de propiedades o tenedores satisfacen la demanda del mercado inmobiliario y han de ser considerados en igual medida.

Debido a la complejidad del sector inmobiliario en una ciudad como Madrid, también existen otros actores clave a tener en cuenta como fondos de inversión locales o grandes compañías e instituciones. Sin embargo, analizar en detalle las disimilitudes entre estos y los actores clave mencionados previamente está fuera del alcance de este proyecto y se considerarán como un mismo grupo de interesados.

- Actores clave internos:

Los trabajadores de nuestro proyecto son aquellos que se encargan de la planificación y realización de todos los hitos y etapas con el fin de proporcionar un producto atractivo para los actores clave externos. Distinguimos así entre dos roles en nuestro proyecto: el rol funcional y el rol técnico.

Se profundizará más en las tareas asignadas a cada uno de ellos en el apartado 2.1 "Gestión del equipo de trabajo".

1.5. DAFO

El análisis DAFO es una técnica de planificación estratégica utilizada para identificar las fortalezas, debilidades, oportunidades o amenazas de un proyecto[16]. Aplicado al nuestro, identificamos los siguientes puntos relevantes:

- Fortalezas:

Al contrario de los modelos econométricos normalmente utilizados, en nuestro caso vamos a tener en cuenta los efectos espaciales en las observaciones, lo cual supone una innovación a la hora de predecir el precio del metro cuadrado, especialmente en el territorio nacional. Bajo este análisis, se espera obtener resultados mejorados con respecto a otros competidores del sector inmobiliario.

- Debilidades:

Nuestro modelo puede presentar un gran sesgo en la variable dependiente debido a que el precio que recuperamos en nuestro conjunto de datos no tiene por qué coincidir con el precio real de la vivienda. También podemos tener problemas en la limpieza del conjunto de datos, por ejemplo si existen viviendas duplicadas que no somos capaces de detectar por el hecho de poseer pequeñas discrepancias entre las observaciones.

Por otra parte, no tenemos la certeza de que nuestros modelos vayan a arrojar buenas predicciones si no conseguimos deshacernos de los efectos espaciales a la hora de realizar los distintos análisis, por lo que se trata de un riesgo que tenemos que asumir.

- Oportunidades:

La mayor ventaja que ofrece este desarrollo es su versatilidad. Por una parte, puede ser aplicado fácilmente a distintas escalas siempre que se disponga de un *dataset* adecuado. Así, por ejemplo, podría reutilizarse para modelizar el precio de la vivienda en distintas ciudades o incluso entre distintas naciones. Por tanto, la capacidad de expandirse a nivel nacional o internacional es muy grande.

De manera similar, un modelo econométrico espacial puede ser aplicado a distintos ámbitos más allá del mercado inmobiliario. Así, siempre que se consideren las variables adecuadas y se disponga del *dataset*, se puede realizar este tipo de modelizado y tener en cuenta factores que no suelen considerarse habitualmente. Por ejemplo, otras posibles aplicaciones podrían ser estudios epidemiológicos[17], del poder adquisitivo o cualquier otra variable económica que pueda presentar una dependencia espacial[18].

Por último, como ya hemos comentado en el apartado 1.3 "Contexto actual", no se han realizado demasiados análisis exhaustivos de este tipo a nivel regional y/o nacional, por lo que la ausencia de competidores supone una gran ventaja a la hora de afrontar un proyecto de esta índole.

■ Amenazas

Existe una considerable dificultad a la hora de obtener un conjunto de datos para el mercado de la vivienda en Madrid debido a que la información se encuentra protegida. Este hecho provoca que el conjunto de datos con el que vamos a trabajar sea difícil de conseguir y, además, reducido.

Para tratar de solventar este problema, se ha desarrollado un código con la finalidad de consultar el portal inmobiliario Idealista, de forma que podamos recuperar la información publicada por parte de sus usuarios relativa a las viviendas anunciadas en compraventa. El conjunto de datos obtenido a través de este programa puede presentar valores inconsistentes, requiriendo una limpieza de los mismos.

Así, tenemos una gran dependencia en la información proporcionada por nuestra fuente de datos. A pesar de tratarse del mayor portal inmobiliario *online* en España, podemos encontrar un gran problema a la hora de obtener unos resultados de calidad ya que las conclusiones que obtengamos en nuestro trabajo pueden no ser extrapolables a muestras más grandes debido a las limitaciones mencionadas.

FORTALEZAS	DEBILIDADES
1. Innovación en la metodología 2. Predicciones mejoradas	1. Datos sesgados 2. Incertidumbre en los resultados
OPORTUNIDADES	AMENAZAS
1. Versatilidad 2. Aplicación a otros ámbitos y escalas 3. Pocos competidores	1. Dificultad en la obtención de datos 2. Muestra reducida

FIGURA 1.1: Matriz de análisis DAFO

Capítulo 2

Desarrollo.

2.1. Gestión del equipo de trabajo

En el desarrollo del proyecto pueden definirse dos roles claros: el rol técnico y el rol funcional. Las funciones de ambos son complementarias y cada uno de ellos es desempeñado por un integrante del equipo. A lo largo de la gestión y realización del proyecto, ambos roles deberán estar en constante comunicación, alertando de riesgos o demoras en el seguimiento de la planificación con el objetivo de progresar con fluidez en los hitos pre establecidos.

Las tareas específicas de cada rol son las listadas a continuación:

- **Rol técnico**

- Preparación, extracción y limpieza del conjunto de datos.
- Implementación de los algoritmos para los modelos predictivos.
- Resolución de problemas en la codificación.
- Recopilación, exposición e interpretación de los resultados.

- **Rol funcional**

- Desarrollo de los objetivos principales, antecedentes y contexto actual.
- Determinación de los actores clave, análisis DAFO, *benchmark* y análisis de costes y beneficios.
- Organización y planificación del proyecto, incluyendo la gestión del equipo y el establecimiento de hitos temporales.
- Planteamiento de hipótesis y marco teórico de los modelos predictivos.
- Exposición de conclusiones.

2.2. Fuentes de información y *benchmark*

En la exposición de los antecedentes y el contexto actual, se describieron trabajos previos que han buscado modelizar el precio de la vivienda en diferentes territorios y utilizando un amplio abanico de técnicas de análisis.

Para la realización de este proyecto, es útil fijarse en las conclusiones alcanzadas por estudios similares como punto de partida para nuestro modelizado.

Así, valiéndonos del marco teórico proporcionado por la bibliografía, el cual asienta la base para las hipótesis que se plantearán en las siguientes secciones, buscamos alcanzar un objetivo más ambicioso que los estudios realizados hasta la fecha.

Lo innovador de nuestra propuesta radica tanto en el amplio número de métodos de modelización –tanto interpretables como no interpretables– que se aplican sobre el conjunto de datos, así como las mejoras en los resultados respecto a las obtenidas por otros estudios similares. Al tenerse en cuenta los efectos de las correlaciones espaciales en el *dataset*, se va a proceder a analizar los residuos de los distintos modelos con el fin de dilucidar qué técnicas logran romper la dependencia espacial y la heterocedasticidad.

En la literatura encontramos todo tipo de variables explicativas en los desarrollos, tales como la población, paro, renta, salarios, ahorro o tipo de interés hipotecario, así como costes de construcción o incluso impuestos y servicios públicos[19][20][21][22][23][24][25]. Sin embargo, son escasos los trabajos que tienen un alcance similar al que aspiramos. Si nos fijamos en aquellos que buscan analizar el problema de la dependencia espacial, vemos que la extensión de su investigación se limita a muy pocos algoritmos de predicción, y, en particular, no se realiza una comparación detallada entre modelos predictivos en un marco nacional[26].

En lo referente a los resultados obtenidos, existe un amplio abanico de especificaciones –unas mejores que otras–, pero generalmente suele llegarse a ajustes relativamente pobres. Es decir, no son capaces de explicar una gran parte de la fluctuación del precio de la vivienda.

En particular, encontramos modelos de tipo OLS, SAR y GWR, pero los resultados y residuos nos son analizados tan en profundidad, pues no se calculan los estadísticos más relevantes como, por ejemplo, el *I de Moran*[27]. En otros estudios se han realizado análisis más profundos, pero no se ha conseguido llegar a coeficientes de determinación globales[28] y locales[29] óptimos ($\bar{R}^2 < 0,7$).

Como veremos en nuestras conclusiones, vamos a ser capaces de obtener mejores estimaciones ($\bar{R}^2 > 0,75$) en todos los modelos gracias a la adición de variables espaciales y la ponderación de las observaciones en función de sus vecindades. Además, complementaremos la evaluación de los resultados con un análisis *SatScan*, que no es habitual en la bibliografía.

Por último, la fuente de datos que manejamos proporciona la ventaja de permitirnos trabajar directamente con los precios del mercado inmobiliario más actualizados en cada momento, posicionándonos a la vanguardia en lo referente a estimaciones del precio de la vivienda.

2.3. *Timing del proyecto*

En esta sección se busca establecer una serie de criterios que regirán el avance del proyecto, en primer lugar desde un punto de vista general, identificando las particularidades de cada fase, y de manera más detallada, estableciendo un cronograma que marcará los principales hitos temporales.

Asimismo, se estudiarán las inversiones realizadas o gastos repercutidos para llevar a cabo la iniciativa, así como el *payback* o beneficio con el objetivo de determinar el plazo de retorno del proyecto.

2.3.1. Fases del desarrollo

El proyecto se llevará a cabo en tres fases: diseño, realización y lanzamiento.

- **Diseño**

En esta primera fase, se plantean los objetivos principales del proyecto y se planifica la manera óptima de alcanzarlos teniendo en cuenta los recursos disponibles, así como considerando posibles imprevistos que pudieran surgir. De esta manera, se establecen unos hitos temporales que ayudarán a controlar el progreso del trabajo llevado a cabo y permitirán identificar riesgos que impidan satisfacer las metas prefigadas.

- **Realización**

La segunda fase consiste en efectuar todas las tareas necesarias para alcanzar los objetivos del proyecto. Durante esta fase, se concretarán las hipótesis y algoritmos a implementar y se aplicarán a los datos previamente almacenados.

- **Lanzamiento**

Durante la fase final del proyecto, se analizarán los resultados obtenidos con el objetivo de exponer unas conclusiones relevantes que transmitan de manera clara y precisa las ideas principales extraídas tras la aplicación de los distintos modelos de predicción. Se determinará si el resultado es satisfactorio o, de lo contrario, se plantearán las dificultades encontradas y se proveerán alternativas para sortearlas en el futuro.

Por último, de ser plausible, se discutirán las aplicaciones reales del proyecto en el mercado inmobiliario nacional.

2.3.2. Cronograma de hitos temporales

En la figura 2.1 se representa un cronograma en el que se han sintetizado los hitos temporales más relevantes de cada una de las etapas expuestas previamente. Así, pueden observarse, por ejemplo, las fechas de comienzo y finalización de cada fase, así como un desglose de las principales tareas a llevar a cabo en el desarrollo del proyecto.

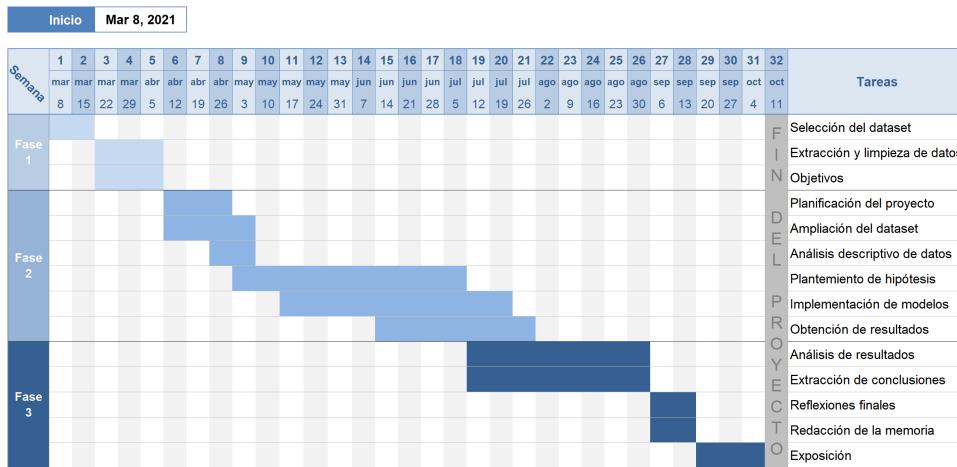


FIGURA 2.1: Cronograma del proyecto

2.3.3. Análisis económico y *payback*

En la planificación del proyecto es necesario realizar una estimación económica, teniendo en cuenta gastos, beneficios y plazos de recuperación de la inversión. Como veremos a continuación, podemos dividir los gastos económicos en iniciales, fijos y variables.

Al comienzo del proyecto, necesitamos crear un proceso de minado de datos desde nuestra fuente, así como otro que los transforme en variables útiles. Además, debemos desarrollar y evaluar los modelos con los cuales queremos realizar nuestra predicción.

Una vez contemos con modelizaciones fiables, la parte fundamental del gasto se centrará en la actualización de nuestra base de datos (con la finalidad de recoger óptimamente las tendencias que sigue el mercado), y en la supervisión de los modelos.

Por último, los gastos variables podrían surgir de necesidades específicas, como la posibilidad de que la fuente de la cual extraemos la información cambie la estructura en la que presenta los datos o introduzca nuevas medidas anti-minado. Esto requeriría revisar y modificar los procesos de extracción, almacenamiento y transformación implementados.

Por otra parte, nuestro principal *target* son empresas y particulares, es decir, tanto demandantes como oferentes de vivienda que pudiesen estar interesados en un peritaje inicial del valor de un inmueble. Actualmente ya existen empresas que se dedican a realizar esta actividad. No obstante, nosotros nos valdremos de modelos innovadores que nos permitirán competir con ellas de manera que consigamos abrirnos hueco en el mercado y generar beneficios.

En relación a los períodos de ejecución, se plantea desarrollar todo el proyecto en un plazo de 6 meses, a partir del cual podremos empezar a obtener flujos de caja que amorticen nuestra inversión inicial. El *payback* depende fuertemente de nuestra habilidad a la hora de colocar nuestros modelos a la vanguardia del mercado. En una estrategia conservadora, estimamos un plazo de 12 meses para recuperar la inversión inicial.

Adicionalmente, es conveniente resaltar que el desarrollo técnico de este proyecto es muy versátil. Por un lado, podría ser aplicado al mercado de la vivienda a diferentes escalas y zonas geográficas del planeta. Por otro lado, se podría elaborar un modelo econométrico espacial aplicado a distintos ámbitos más allá del sector inmobiliario. Así, de cara a afrontar nuevos proyectos de esta índole, nos enfrentaríamos a un coste inicial más reducido y un intervalo de desarrollo y puesta en marcha más corto.

2.4. Hipótesis y planteamientos realizados

Con este trabajo perseguimos resaltar la relevancia del factor espacial en el precio de la vivienda, concretamente en la ciudad de Madrid. Fijando este objetivo, vamos a proceder a plantear las hipótesis preliminares para los algoritmos de predicción a implementar, siendo estos tanto interpretables como no interpretables.

Asimismo, los distintos modelos predictivos serán juzgados tanto en base a sus respectivas bondades de ajuste como a través del análisis de sus residuos con la finalidad de determinar su idoneidad.

2.4.1. Regresión lineal múltiple (RLM)

Nuestro punto de partida será una regresión lineal múltiple, cuya forma funcional viene dada por:

$$Y = \sum_{i=1}^k X_i \beta_i + \epsilon \quad (2.1)$$

donde Y es la variable dependiente de interés (el precio de la vivienda), X_i son las variables explicativas del modelo, β_i es el coeficiente de regresión que mide la influencia de cada variable X_i sobre Y y ϵ es el error aleatorio.

Las principales hipótesis de este tipo de regresión son:

- **Linealidad** en la relación entre X_i e Y .
- **Independencia** entre las observaciones, entre las variables explicativas y entre los residuos del modelo.
- **Normalidad** en la distribución de los residuos.
- **Homocedasticidad** en los residuos.

En cuanto se viola una de las hipótesis, el modelo deja de ser óptimo y no podemos garantizar la fiabilidad de sus predicciones. Como veremos en el siguiente apartado, nuestro conjunto de datos está muy afectado por la dependencia espacial, lo cual se traduce en dependencias entre observaciones y heterocedasticidad en los residuos. Para hacer frente a este inconveniente, tomamos dos planteamientos alternativos:

- Adición de variables espaciales con la intención de vencer la dependencia espacial.

- Adición de no linealidades mediante modelos *Multiadaptive regression splines* (MARS).

Aun así, no esperamos lograr romper completamente los efectos espaciales, por lo que recurriremos a modelos más robustos en los que se añade un término de dependencia espacial, bien en la variable dependiente (modelos de retardo espacial), bien en los residuos (modelos de error espacial).

2.4.2. Modelos de retardo espacial (SAR)

Este tipo de modelos incluyen la correlación espacial en la variable dependiente y permiten a las observaciones en una determinada zona depender de observaciones en áreas vecinas. El modelo de retardo espacial básico se define como[30]:

$$Y = X\beta + \rho WY + \epsilon \quad (2.2)$$

siendo W la matriz de pesos espaciales, ϵ los errores independientes y ρ el nivel de relación autorregresiva espacial entre la variable dependiente y sus observaciones vecinas. Es decir, ρ es el impacto "boca a boca", lo cual quiere decir que las observaciones están impactadas por lo que sucede a su alrededor.

Resolviendo el sistema se obtiene:

$$Y = (I - \rho W)^{-1}(X\beta + \epsilon) \rightarrow E[Y] = (I - \rho W)^{-1}(X\beta) \quad (2.3)$$

De esta forma, esperamos obtener un $\rho \neq 0$ muy significativo, de manera que los residuos del sistema puedan considerarse independientes y, por tanto, estemos ante una mejor especificación del modelo.

2.4.3. Modelos de error espacial (SEM)

Como ya hemos argumentado, este tipo de modelos explican la dependencia espacial en el término de error o residual, es decir, el error lleva implícita una estructura espacial.

Se define como:

$$Y = X\beta + e \quad (2.4)$$

$$e = \lambda We + \epsilon \quad (2.5)$$

donde W es la matriz de pesos espaciales, ϵ el término aleatorio de error y λ es el parámetro autorregresivo.

Resolviendo el sistema:

$$Y = X\beta + (I - \lambda W)^{-1}\epsilon \quad (2.6)$$

En esta ocasión esperamos vencer por completo la heterocedasticidad de los residuos y, al igual que en el modelo SAR, lograr una muy buena especificación del sistema.

2.4.4. Modelos geográficamente ponderados (GWR)

Hasta ahora hemos definido modelos de regresión global general, en los cuales se tienen valores únicos de los parámetros β_i para todas las observaciones del conjunto de datos.

Con este tipo de modelizado, sin embargo, en lugar de tener un coeficiente global para cada variable, los coeficientes pueden variar en función del espacio. La idea fundamental es la medición de la relación entre la variable respuesta y sus variables explicativas independientes a través de la combinación de las diferentes áreas geográficas.

El modelo se define como[31]:

$$Y_s = \beta_{s1}X_1 + \dots + \beta_{s1}X_p + \epsilon \quad (2.7)$$

siendo s cada zona geográfica. Es decir, en el modelo ponderado geográficamente se tienen diferentes estimadores para cada una de las variables dependiendo de la localización.

Resolviendo el sistema:

$$\beta_s = (X^t W_s X)^{-1} X^t W_s Y \quad (2.8)$$

Así, conseguimos reducir la dependencia espacial de los residuos del modelo, aunque no vamos a romper la heterocedasticidad de los mismos como veremos más adelante.

2.4.5. Gradient Boosting (GB)

El método de *Gradient Boosting* es una técnica de aprendizaje automático o *Machine Learning* que genera un modelo predictivo a partir de un conjunto de algoritmos de predicción débiles, típicamente árboles de decisión[32][33].

Al combinar *weak learners* de forma iterativa, el objetivo es que el algoritmo F aprenda a predecir valores $\hat{y} = F(x)$ minimizando el error cuadrático medio $\frac{1}{n} \sum_i (\hat{y}_i - y_i)^2$, donde i es el índice sobre el subconjunto de entrenamiento. De esta manera, en cada iteración el árbol de decisión se centra en disminuir los errores arrojados en la predicción previa.

- \hat{y}_i es la predicción del modelo.
- y_i es el valor observado.
- n es el número de observaciones.

La predicción final se obtendrá a partir de la suma de todas las predicciones de los árboles de decisión implementados.

Al contrario de los modelos propuestos hasta ahora, el método de GB se trata de una técnica no interpretable, que además requiere de un gran esfuerzo en la parametrización o *fine-tunning*, de manera que no se caiga en un sobreajuste al conjunto de datos.

2.5. Desarrollo y programación utilizada

En este apartado buscamos plantear los pasos previos a la modelización, así como exponer los análisis y métodos de evaluación que se llevarán a cabo para cada una de las implementaciones propuestas.

2.5.1. Base de datos

La extracción de la información del portal inmobiliario *Idealista* se ha llevado a cabo mediante un método de *web scraping* en el que se ha barrido cada uno de los 21 distritos de la ciudad de Madrid, de manera que se ha obtenido un total de 5935 observaciones divididas según se indica en la tabla 2.1.

Distrito	n	Distrito	n	Distrito	n
Arganzuela	301	Fuencarral	256	Salamanca	270
Barajas	257	Hortaleza	276	San Blas	263
Carabanchel	293	Latina	293	Tetuán	292
Centro	336	Moncloa	286	Usera	313
Chamartín	267	Moratalaz	240	Vicálvaro	310
Chamberí	256	Puente de Vallecas	312	Villa de Vallecas	264
Ciudad Lineal	257	Retiro	262	Villaverde	331

CUADRO 2.1: Número de viviendas por cada distrito de la base de datos.

El detalle de las variables explicativas recuperadas, así como la información acerca de su variabilidad, pueden consultarse en los apéndices A y B.

Asimismo, en la gráfica 2.2 se puede conocer la distribución del precio de mercado por metro cuadrado de los inmuebles en los diferentes distritos. Como se puede observar, hay una diferencia del 380 % entre el barrio más caro, el de Salamanca, y el más barato, Villaverde. La variabilidad puede observarse en el mapa de la figura 2.3.

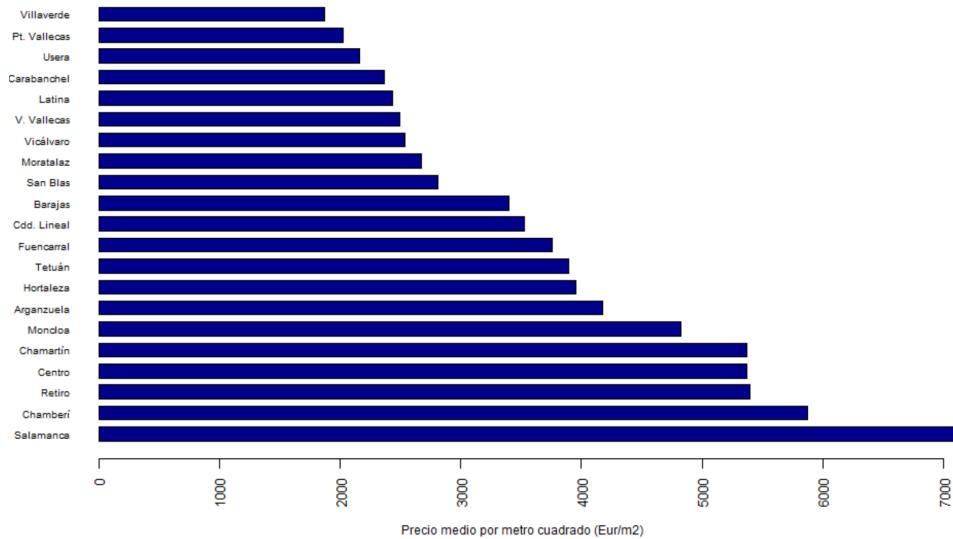


FIGURA 2.2: Precio medio del metro cuadrado por distrito.

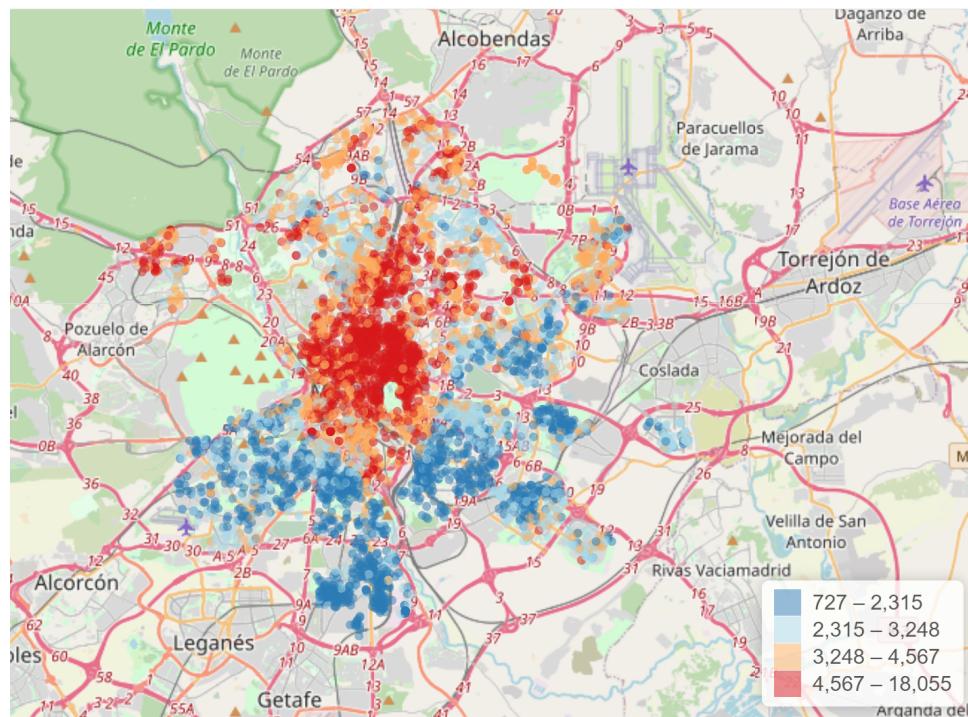


FIGURA 2.3: Distribución del precio del metro cuadrado.

2.5.2. Preparación del dataset

El desarrollo se ha realizado en el lenguaje de programación R, mediante los paquetes *stats*, *earth*, *spatialreg*, *spdep*, *sp*, *spgwr*, *GWmodel*, *fBasics*, *lmtest*, *gbm*.

Entrenamiento y validación

Para la implementación de las modelizaciones expuestas, en primer lugar se divide el *set* de datos en dos subconjuntos: entrenamiento (70 %) y *testing*

(30 %). Gracias a este paso previo, podremos validar el modelo y evaluar su poder de predicción.

Información geográfica

Por otra parte, vamos a valernos del proyecto colaborativo [OpenStreetMap](#) para descargar información geográfica relevante (colegios, hospitales, etc.). La situación de los puntos de interés será incluida en nuestro conjunto de datos, permitiéndonos ponderar cada observación en relación a su proximidad a dichas localizaciones.

A continuación se incluyen las visualizaciones de los datos descargados, así como las variables calculadas a partir de los mismos:

- **Hospitales**

La información se ha descargado mediante una búsqueda con *key = 'amenity'* y *value = "hospital"*:

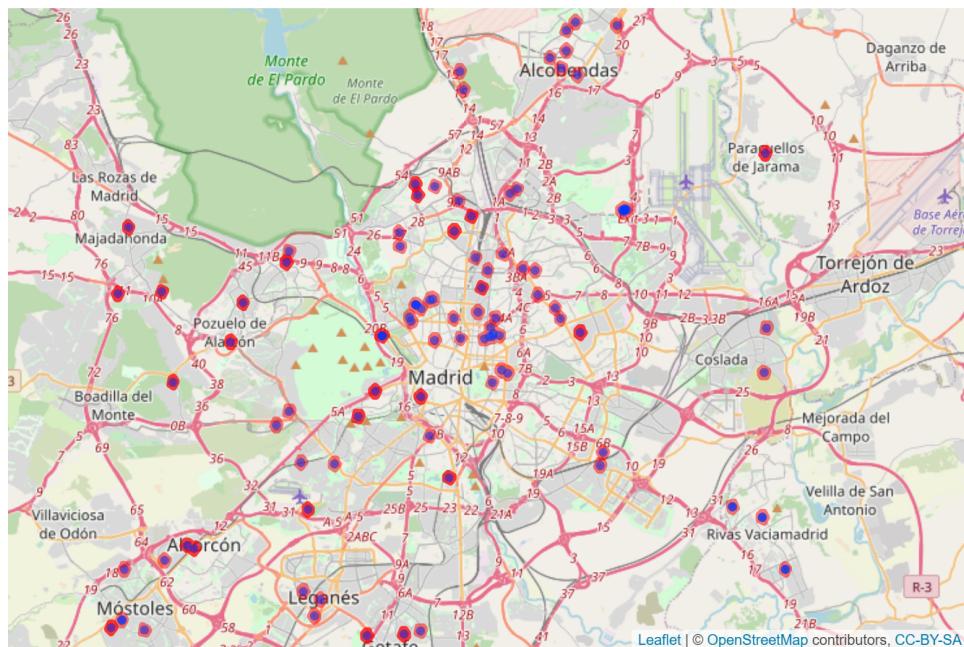


FIGURA 2.4: Mapa de polígonos con los principales hospitales de Madrid.

A partir de esta información se ha calculado la densidad de hospitales en un radio de 1km para cada vivienda.

- **Centros comerciales**

La información se ha descargado mediante una búsqueda con *key = 'shop'* y *value = "mall"*:

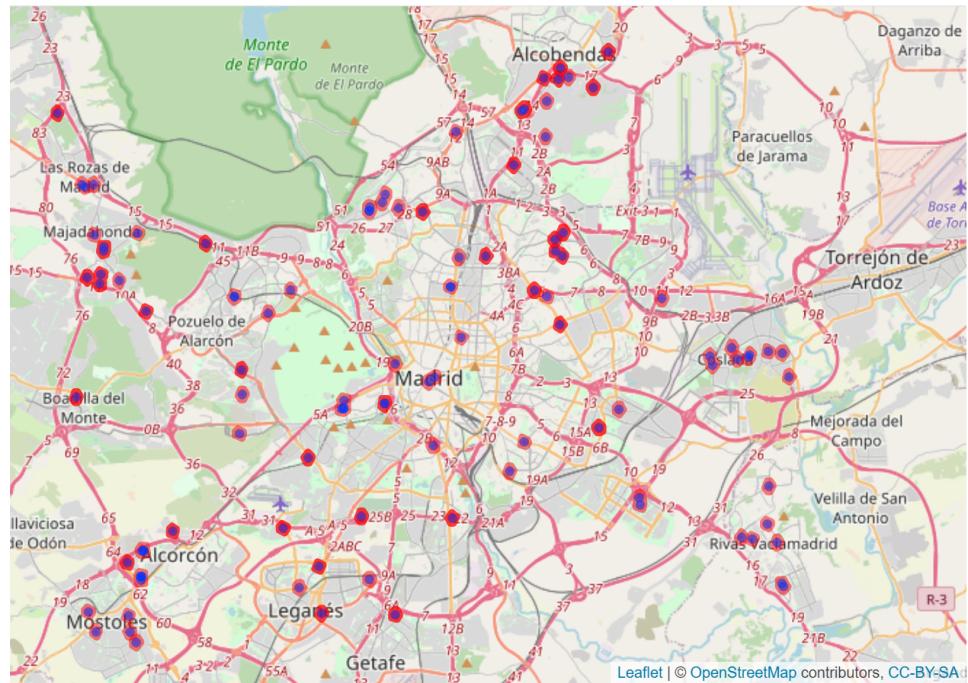


FIGURA 2.5: Mapa de polígonos con los principales centros comerciales de Madrid.

A partir de esta información se ha calculado la densidad de centros comerciales en un radio de 1km para cada vivienda.

- Transporte público

La información se ha descargado mediante una búsqueda con *key = 'public_transport'* y *value = "station"*:

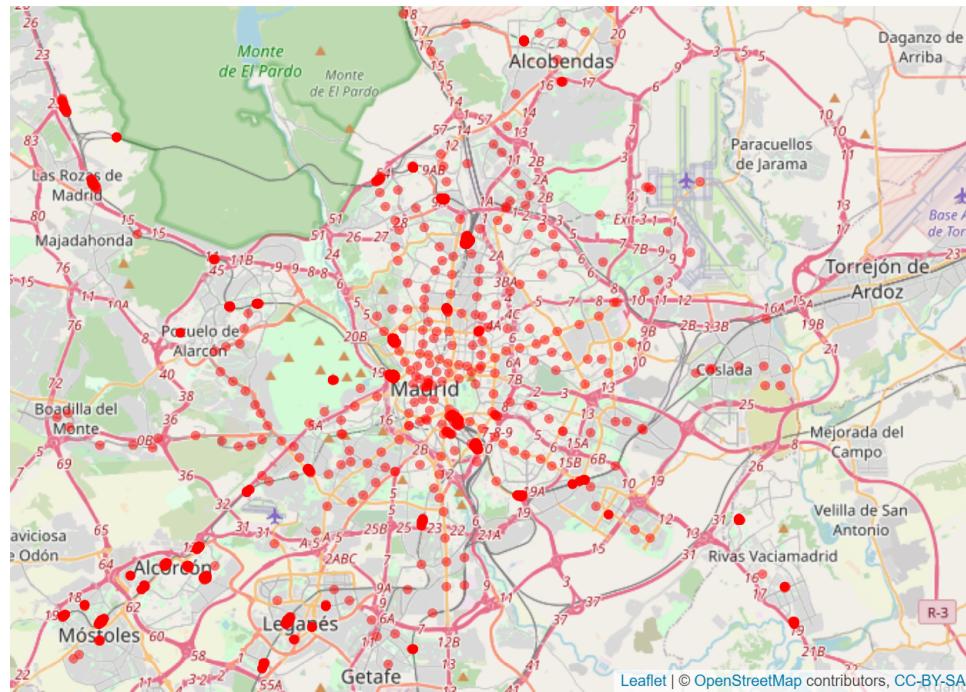


FIGURA 2.6: Mapa de estaciones de metro y cercanías RENFE de Madrid.

A partir de esta información se ha calculado la distancia más cercana a una estación de metro o de cercanías RENFE para cada vivienda.

■ Colegios

La información se ha descargado mediante una búsqueda con `key = 'amenity'` y `value = "school"`:

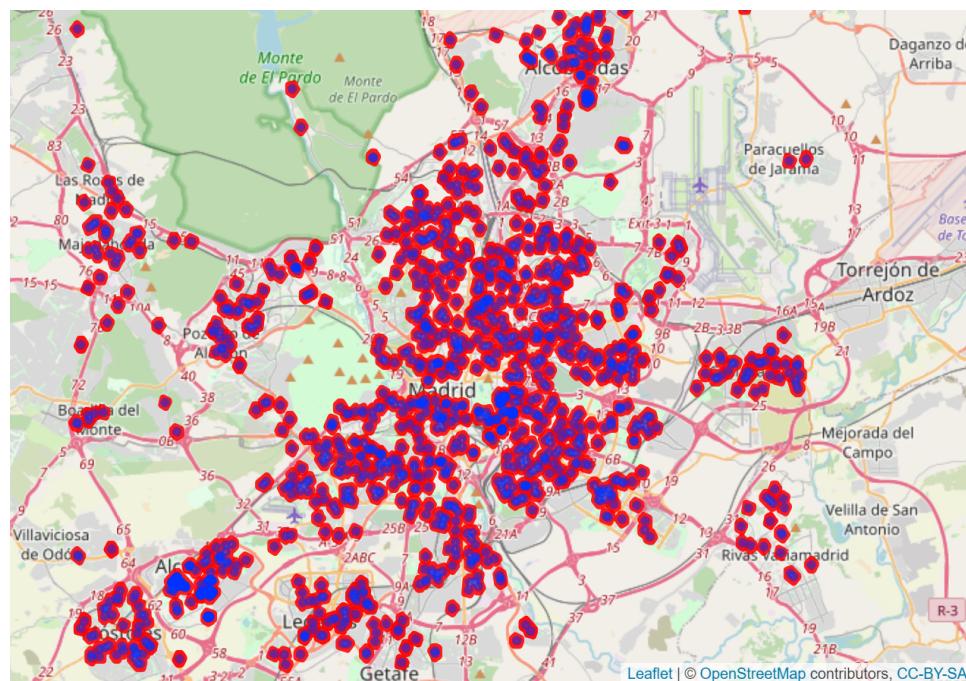


FIGURA 2.7: Mapa de polígonos con los colegios de Madrid.

A partir de esta información se ha calculado la densidad de colegios en un radio de 1km para cada vivienda.

Así, para la determinación de ubicaciones más próximas, el número de observaciones en el vecindario establecido para cada ubicación puede expresarse mediante una matriz de ponderaciones o pesos espaciales W . Para los modelos autorregresivos SAR y SEM, la matriz se ha restringido a los 10 vecinos más cercanos.

2.5.3. Evaluación del modelo

A la hora de evaluar cada modelo y determinar su bondad de ajuste, vamos a apoyarnos fundamentalmente en cuatro validaciones:

- **Coeficiente de determinación ajustado o *pseudo- \bar{R}^2* .** Su valor indica la proporción de variabilidad en la variable endógena explicada por el modelo en relación a la variabilidad total, ajustándose al número de grados de libertad[34]:

$$\bar{R}^2 = 1 - \frac{n - 1}{n - k - 1} \frac{SS_{\text{res}}}{SS_{\text{tot}}} \quad (2.9)$$

siendo n el tamaño de la base de datos, k el número de variables explicativas, SS_{res} la suma de residuos al cuadrado y SS_{tot} la suma total de cuadrados.

- ***I* de Moran.** Este indicador proporciona una medida de la autocorrelación espacial, comparando el valor en una determinada área i en relación al resto de áreas $j \neq i$ [35]. Su forma viene dada por

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad (2.10)$$

siendo N el número de áreas consideradas, w_{ij} las componentes de la matriz de pesos espaciales y y_i el valor de la variable Y en el área i .

- **Test de Jarque-Bera.** Se trata de una prueba de bondad de ajuste para comprobar si una muestra de datos tiene la asimetría y curtosis de una distribución normal[36]. Su forma es

$$I = \frac{n}{6} \left(S^2 + \frac{1}{4}(K - 3)^2 \right) \quad (2.11)$$

donde n es el número de observaciones, S la asimetría de la muestra y K la curtosis.

- **Spatial Scan Statistics.** Detecta y evalúa *clusters* en el espacio, permitiendo diferenciar si estos ocurren de forma aleatoria o siguen una distribución de probabilidad determinada. Para ello, se analiza gradualmente en intervalos espaciales si la variable en cuestión toma valores diferentes a los esperados[37].

En nuestro caso, usaremos la herramienta *SatScan*, un *software* especializado mediante el cual analizaremos los residuos de cada modelo,

esperando que su distribución sea normal en cada región del espacio. Definimos estos intervalos espaciales usando círculos que contengan al 10 % de la población y cuyo centro esté localizado en cada una de nuestras observaciones. El programa *SatScan* se valdrá de simulaciones Montecarlo, obteniendo un p-value para cada región que no cumpla la distribución esperada.

2.6. Comparativas entre modelos

Una vez definidos el marco teórico y el método de evaluación para cada modelo, se recogen los resultados obtenidos para cada uno de ellos.

2.6.1. RLM

En este apartado se incluyen los detalles de las distintas modelizaciones RLM, entre las cuales se encuentran:

- Modelo lineal básico o *baseline*.

En este primer desarrollo se utiliza la función 'lm' con el objetivo de implementar un modelo de regresión lineal múltiple en el que no se han incluido variables espaciales.

- Modelo lineal espacial.

Se ha añadido complejidad al modelo lineal básico descargando información geoespacial. El detalle completo de los atributos añadidos al *dataset* se tiene en el apartado [2.5.2](#).

- *Multiadaptive regression splines*.

Con esta técnica, se sigue conservando la información geoespacial y además se permite la presencia de no linealidades, dividiendo las variables que presentan este tipo de comportamiento de manera que se posibilite que una misma característica tenga contribuciones diferentes al modelo. En concreto, los atributos seleccionados han sido **habitaciones** (3), **baños** (2) y **dist_centro** (4.4km).

La siguiente tabla resume la comparación entre las tres modelizaciones, mostrándose siempre los resultados obtenidos al aplicar el algoritmo sobre un subconjunto de prueba compuesto por el 30 % de la base de datos original:

Modelo	\bar{R}^2	I de Moran (p-value)	Jarque-Bera (p-value)
RLM (<i>baseline</i>)	0,768676	0,218662 ($< 2,2 \cdot 10^{-16}$)	179,9 ($< 2,2 \cdot 10^{-16}$)
RLM espacial	0,779565	0,184652 ($< 2,2 \cdot 10^{-16}$)	175,9 ($< 2,2 \cdot 10^{-16}$)
MARS	0,789346	0,161736 ($< 2,2 \cdot 10^{-16}$)	206,9 ($< 2,2 \cdot 10^{-16}$)

CUADRO 2.2: Principales resultados para modelos RLM.

Evaluación de los modelos

La sección 2.4.1 establece las hipótesis básicas de una regresión lineal. Sin embargo, los resultados de los modelos predictivos entran en contradicción con algunas de dichas hipótesis fundamentales, como se expondrá a continuación:

- Normalidad en la distribución de los residuos.

La distribución de los residuos puede advertirse en las gráficas C.4, C.9 y C.14, en las cuales han sido representados en función de la variable dependiente estimada para los tres modelos lineales. En dichas gráficas, puede observarse cómo los residuos se concentran en torno al valor cero, pero la nube de puntos no termina de asemejarse a un ruido blanco como cabría esperar de verse distribuidos normalmente.

Asimismo, las figuras C.3, C.8 y C.13 muestran los gráficos cuantil-cuantil o *Q-Q plots*. En ellos ha sido reflejado mediante una recta el comportamiento de una distribución normal. Como puede observarse, la distribución de los residuos se asemeja a una normal para los valores centrales, mientras que se desvía en los extremos. Esto implica que la cantidad de observaciones situadas en el primer cuantil (así como en el último) difiere con respecto al supuesto teórico de una distribución normal.

El *test* de Jarque-Bera aplicado, además, concluye que, efectivamente, los residuos se desvían de una normal en los tres modelos de regresión lineal implementados. Este hecho podría implicar problemas a la hora de realizar estimaciones, puesto que los errores no son aleatorios sino que tienen una estructura subyacente que el modelo no es capaz de captar. Sin embargo, no tiene por qué darse necesariamente esta casuística ya que, como se verá a continuación a la hora de calcular la bondad de ajuste, las predicciones arrojadas sobre nuevos conjuntos de datos son aceptables.

- Independencia entre residuos.

El hecho de que exista una autocorrelación espacial entre los residuos, tal y como indica el resultado obtenido en el *test I de Moran* (figuras C.2, C.7 y C.12), viola la hipótesis de independencia entre los mismos. Como consecuencia, los modelos lineales generados son poco fiables, puesto que los parámetros de regresión pueden estar sesgados, traduciéndose en una sobreestimación o subestimación de su poder predictivo.

En concreto, esto significa que los precios de los inmuebles están interconectados, es decir, los pisos caros están cerca de los caros y viceversa. Ignorar este hecho puede acarrear graves consecuencias a la hora de generalizar el modelo y realizar nuevas predicciones, especialmente en diferentes zonas geográficas.

- Coeficiente de determinación \bar{R}^2 .

Debido a los argumentos enumerados, este tipo de modelos no son los más idóneos a la hora de hacer frente a un *dataset* fuertemente correlacionado espacialmente, como es nuestro caso. Para asegurar hasta

qué punto son buenos predictores estos modelos RLM, se ha trabajado sobre un subconjunto de prueba y se han recogido los resultados en el cuadro 2.2.

En este, se observa que el simple hecho de añadir información geoespacial a la parametrización está consiguiendo mejorar ligeramente el coeficiente de determinación ajustado del modelo lineal básico, lo cual implica una mejora en las predicciones.

Sin embargo, la comparativa entre las tres especificaciones arroja un claro vencedor: el modelo MARS. Era de esperar puesto que, mediante la adición de no linealidades, hemos dado flexibilidad a los estimadores, hecho que permite alcanzar una mayor precisión en las predicciones, traduciéndose en un valor superior del coeficiente de determinación ajustado. De esta forma, incluso teniendo en cuenta la penalización aportada por la inclusión de variables adicionales en el sistema de ecuaciones, se consigue aumentar en un 2 % la precisión del algoritmo con respecto al *baseline*.

- *Spatial Scan Statistics*

Como cabría esperar de la fuerte dependencia espacial observada entre los residuos, en el análisis con *SatScan* se observan *clusters* donde los residuos no poseen la distribución normal esperada (figuras C.5, C.10 y C.15). Estas regiones del espacio son especialmente problemáticas para nuestros modelos, siendo sus predicciones menos confiables.

Modelo	<i>Clusters</i> (p-value < 0,1)	Observaciones <i>test</i>
RLM (<i>baseline</i>)	3	9.32 %
RLM espacial	2	2.64 %
MARS	2	2.58 %

CUADRO 2.3: Resultados del análisis *SatScan* para los modelos de regresión lineal.

En la tabla 2.3 se recogen los resultados obtenidos para los tres modelos lineales. Así, en el modelo *baseline* se observan tres regiones con un p-value <0,1, representando el 9.32 % del subconjunto de prueba. Por su parte, el hecho de añadir variables espaciales a la parametrización en el modelo espacial reduce el número de *clusters* a dos más pequeños en los que tan solo cae el 2.64 % de las instancias de prueba. Un resultado similar se obtiene para el MARS, en el cual también hay dos regiones del espacio problemáticas que suponen el 2.58 % del total del subconjunto de prueba.

Estos resultados están en consonancia con los obtenidos en los otros indicadores, donde vimos una mejora tanto para el modelo espacial como para el de tipo MARS.

En resumen, los modelos RLM proporcionan un buen punto de partida en la modelización del precio de la vivienda, pero es necesario ir un paso más

allá para deshacerse de los efectos espaciales y mejorar la bondad de ajuste. Con el objetivo de alcanzar una mejor especificación, es necesario atacar explícitamente la autocorrelación espacial presente en el *dataset*, lo cual permitirá incrementar el porcentaje de varianza explicada para la variable dependiente de los modelos predictivos. En concreto, se plantea la implementación de modelos SAR, SEM y GWR.

2.6.2. SAR

Bajo el marco teórico expuesto en la sección 2.4.2, se ha desarrollado un modelo de retardo espacial que tendrá como objetivo incorporar la autocorrelación espacial en la variable dependiente.

El cuadro 2.4 áuna los principales resultados de aplicar el modelo, una vez entrenado, sobre el subconjunto de prueba.

ρ	\bar{R}^2	<i>I</i> de Moran (p-value)	Jarque-Bera (p-value)
$0,521 \pm 0,017$	0,8129067	-0,01319369 (0,8971)	365,1 ($< 2,2 \cdot 10^{-16}$)

CUADRO 2.4: Resultados sobre el subconjunto de prueba del modelo SAR.

Si nos fijamos, un valor del parámetro ρ significativo implica un alto nivel de relación autorregresiva espacial entre la variable dependiente y sus observaciones vecinas. Es decir, se están incorporando satisfactoriamente los efectos espaciales al modelo.

Este factor se ve recalado mediante el *test I de Moran* (figura D.2), para el cual observamos un valor compatible con la hipótesis nula según la cual no existe dependencia espacial en los residuos. Por tanto, se concluye que los residuos no están autocorrelados espacialmente y **se ha conseguido romper la dependencia espacial**.

Cabe destacar, como era de esperar, un aumento en el coeficiente de determinación \bar{R}^2 con respecto a las modelizaciones previas de más de un 2.5 %. Este hecho es debido principalmente a que, al tratarse de un modelo con un sesgo menor, su capacidad predictiva se ha visto acentuada a la hora de estimar el valor de nuevas viviendas no incluidas en el *dataset* original de entrenamiento, independientemente de su zona geográfica.

En cuanto a la normalidad en la distribución de los residuos, no se aprecia una gran diferencia si se compara con las conclusiones extraídas en el apartado anterior, siendo tanto las gráficas como el resultado del *test de Jarque-Bera* indicativos de una desviación respecto a la normal (ver figuras D.3 y D.4).

En lo referente al análisis estadístico espacial realizado con *SatScan* (ver figura D.5), se observan dos *clusters* que representan el 4.15 % de las instancias del subconjunto de prueba. A diferencia de lo ocurrido con los modelos RLM, en este caso encontramos una región grande en el centro del mapa. Este resultado pone de manifiesto que, aunque el *test I de Moran* nos indica que se ha conseguido romper la dependencia espacial, el centro de la

ciudad contiene una región donde los residuos no presentan la esperada distribución normal.

Así, las posibles causas de este efecto pueden ser las expuestas a continuación:

1. Omisión de variables relevantes. Pese a que nuestro modelo incluye la variable **dist_centro** (representando la distancia al centro de Madrid), así como los diferentes distritos en los que se halla cada piso, puede que la dimensión espacial no se esté teniendo en cuenta adecuadamente o haya más atributos relevantes en esta zona.
2. Mala especificación del modelo. Es posible que nuestro modelo no sea el adecuado para resolver este tipo de problema y, aunque se incluyan más variables explicativas, no se aprecie ninguna mejora significativa.
3. Problemas con la linealidad del modelo. Puede ocurrir que las variables presenten no linealidades precisamente en las regiones identificadas con *SatScan*.

2.6.3. SEM

El siguiente modelo que incorpora efectos espaciales es el modelo de error espacial, el cual se diferencia del modelo de retardo espacial en que la estructura espacial va implícita en los residuos (ver sección 2.4.3).

De nuevo, se recogen en el cuadro 2.5 los principales resultados de aplicar el modelo, una vez entrenado, sobre el subconjunto de prueba.

λ	\bar{R}^2	I de Moran (p-value)	Jarque-Bera (p-value)
$0,611 \pm 0,019$	0,8117324	$-0,01212374 (0,8763)$	$197,2 (< 2,2 \cdot 10^{-16})$

CUADRO 2.5: Resultados sobre el subconjunto de prueba para el modelo SEM.

Un valor significativo del parámetro autorregresivo $\lambda \neq 0$ confirma la presencia de autocorrelación espacial en los residuos del modelo. Asimismo, el resultado del *test I de Moran* (figura E.2) es similar al obtenido para el modelo de retardo espacial, por lo que de nuevo puede concluirse que **se ha conseguido vencer la dependencia espacial**.

Al conseguir mejorar significativamente la especificación del modelo, podemos decir que se trata de un predictor menos sesgado mediante el cual se ha conseguido incrementar notablemente la capacidad predictiva con respecto al modelo *baseline*, en concreto en un 4 %. Cabe esperar, por tanto, mayor fiabilidad en las estimaciones efectuadas mediante esta implementación para el precio de la vivienda.

El resto de resultados alcanzados son muy similares a los del modelo de retardo espacial (SAR), viéndose una vez más cómo los residuos se alejan de una distribución normal (figura E.3 y E.4). Por su parte, en el análisis estadístico espacial realizado con *SatScan* (figura E.5), se observa un *cluster* que constituye el 3.03 % de las instancias del subconjunto de prueba y el cual

está localizado en el centro del mapa. Al igual que ocurre con el modelo SAR, el resultado nos indica que, pese a que el *test I de Moran* concluye que se consigue romper la dependencia espacial, el centro contiene una región donde los residuos no presentan la esperada distribución normal. Las posibles causas de este fenómeno ya fueron discutidas en el apartado anterior.

2.6.4. GWR

El modelo GWR usa las coordenadas de cada instancia y divide el mapa en zonas geográficas en las cuales aplica una regresión lineal múltiple (ver sección 2.4.4). Para definir las zonas geográficas, se ha buscado la mejor separación a través del método *gwr.sel*, el cual devuelve un *bandwidth* idóneo de 1.56 km.

A continuación, en la tabla 2.6, se exponen los resultados obtenidos sobre el subconjunto de prueba:

\bar{R}^2	<i>I</i> de Moran (p-value)	Jarque-Bera (p-value)
0,7738334	0,1679498 ($< 2,2 \cdot 10^{-16}$)	162,5539 ($< 2,2 \cdot 10^{-16}$)

CUADRO 2.6: Resultados sobre el subconjunto de prueba para el modelo GWR.

El resultado del *test I de Moran* revela que existe una autocorrelación entre los residuos del modelo. Este hecho era de esperar pues, al contrario de lo especificado en los modelos SAR y SEM, en este caso no se ha definido ningún parámetro que introduzca las posibles relaciones espaciales entre observaciones del conjunto de datos. Por tanto, de manera análoga a las conclusiones alcanzadas para los modelos de regresión lineal (RLM), el resultado del *test* es indicativo de la necesidad de tener en cuenta la influencia de observaciones vecinas a la hora de estimar el precio de cada inmueble y el hecho de no considerar este factor puede conllevar predicciones sesgadas o erróneas.

Por su parte, no se observan novedades en el resultado del *test* de normalidad, siendo este consistente con las conclusiones extraídas hasta ahora en las modelizaciones previas (figuras F.2 y F.3).

Por último, el análisis *SatScan* nos devuelve dos *clusters* significativos, con un p-value <0.1 , los cuales representan el 6.51 % de la base de datos de prueba (ver F.4). En estas regiones, el algoritmo falla más de lo esperado a la hora de estimar el valor de los inmuebles. Tal y como vimos en el modelo SAR, es posible que esto pueda deberse a varios motivos como una omisión de variables relevantes –cuya solución consistiría en incluir nuevas variables que recojan mejor las características particulares de la zona–, una mala especificación del modelo debido a las dependencias espaciales entre vecinos o incluso problemas con la linealidad. Sin embargo, en los modelos RLM solo se vio una ligera mejora a la hora de tratar de romper la linealidad de las variables, por lo que las dos primeras son las causas que creemos más probables. Por esta razón, es oportuno tratar de trabajar con modelos que sean más flexibles a la hora de tratar los datos, como es *Gradient Boost*.

2.6.5. Gradient Boosting

Gradient Boosting (GB) es un método no interpretable de aprendizaje automático (ver sección 2.4.5), el cual requiere de la especificación de hiperparámetros de entrada para optimizar su comportamiento. Este proceso se conoce como *fine-tuning* y es fundamental si se busca implementar la mejor especificación del algoritmo. Con esta finalidad, se ha generado y evaluado un conjunto de modelos, cada uno de ellos con distintos valores para el *learning rate* (α) y la profundidad de los árboles de decisión.

En el cuadro 2.7 se muestran los principales resultados de la búsqueda, sobre los que se seleccionará aquel con mayor poder predictivo, es decir, aquel con un valor superior para el coeficiente de determinación ajustado \bar{R}^2 sobre el subconjunto de prueba.

α	SSE	Árboles	Tiempo (s)	\bar{R}^2 (test)	Profundidad
0.001	0.271	10000	70.78	0.6586	1
0.001	0.245	10000	132.58	0.7285	2
0.001	0.233	10000	191.84	0.7577	3
0.005	0.227	10000	70.33	0.7710	1
0.010	0.220	10000	69.77	0.7844	1
0.050	0.214	9938	71.75	0.7934	1
0.100	0.213	7722	71.30	0.7943	1
0.005	0.211	10000	131.93	0.8040	2
0.005	0.207	10000	187.27	0.8143	3
0.010	0.205	9997	132.90	0.8160	2
0.100	0.204	2455	131.90	0.8199	2
0.100	0.205	1423	187.67	0.8200	3
0.050	0.203	3880	131.40	0.8205	2
0.050	0.203	2285	188.80	0.8211	3
0.010	0.203	9910	188.06	0.8214	3

CUADRO 2.7: *Grid search* para el modelo de *Gradient Boost*.

Así, el modelo óptimo que hallamos cuenta con un *learning rate* de $\alpha = 0,01$ y una profundidad de 3 niveles en cada árbol. Además, para garantizar que nuestro modelo no presenta *overfitting*, se representa el coeficiente de determinación ajustado \bar{R}^2 para los subconjuntos de entrenamiento y prueba en función del número de árboles de decisión.

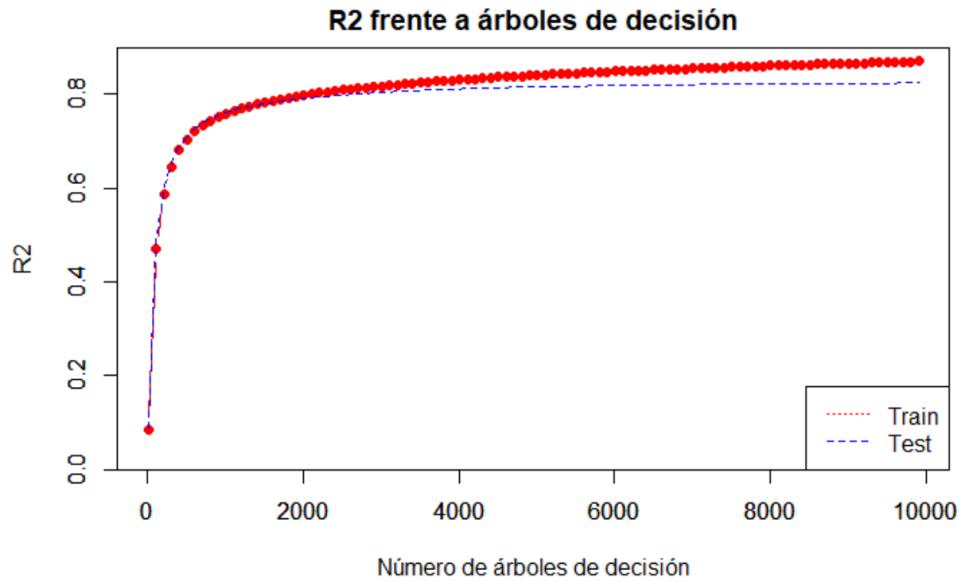


FIGURA 2.8: Evolución del coeficiente de determinación ajustado \bar{R}^2 en función del número de árboles de decisión.

Comprobamos, por tanto, en la figura 2.8 que el algoritmo va mejorando el resultado sobre el conjunto de entrenamiento conforme aumenta el número de árboles. No obstante, para el conjunto de prueba esta mejora se detiene, alcanzando un máximo en 9910 árboles. A partir de dicho punto, no tiene sentido seguir incrementando la complejidad del algoritmo pues, aunque un aumento siempre mejoraría los resultados sobre el conjunto de entrenamiento, no se estaría progresando en el aprendizaje sino que más bien se estarían memorizando las instancias de entrenamiento sin mejorar las predicciones sobre nuevos *datasets*.

Una vez conocida la mejor especificación del modelo, se procede a recopilar los principales resultados del mismo (cuadro 2.8).

\bar{R}^2	I de Moran (p-value)	Jarque-Bera (p-value)
0,8217865	0,07633408 ($6,722 \cdot 10^{-15}$)	494,6932 ($< 2,2 \cdot 10^{-16}$)

CUADRO 2.8: Resultados sobre el subconjunto de prueba para el modelo GB.

Así, este desarrollo destaca por su coeficiente de determinación ajustado $\bar{R}^2 = 0,82$ sobre el conjunto de *test*, el valor más alto de los obtenidos hasta ahora. El resto de resultados no son reseñables, pues no difieren mucho de los ya argumentados en las modelizaciones previas. En concreto, los residuos no presentan una distribución normal, tal y como indica el *test* de Jarque-Bera (figuras G.2 y G.3), y el *test I de Moran* nos muestra que están correlacionados espacialmente (ver figura G.1), como era de esperar.

Cabe destacar que los supuestos de normalidad e independencia de residuos son condiciones necesarias para validar los modelos interpretables con los que hemos tratado en este trabajo. No obstante, *Gradient Boost* es

un modelo no paramétrico para el cual estas suposiciones iniciales no son requeridas para alcanzar una modelización robusta. En efecto, al no tener que estimar parámetros, no tenemos que inferir su distribución a partir de las suposiciones de independencia, homocedasticidad y normalidad de los residuos.

Sin embargo, por este mismo motivo no pueden deducirse efectos marginales sobre las variables explicativas. En su lugar, podemos contar el número de veces que una variable aparece en los árboles de decisión que componen el algoritmo, lo cual nos ofrece una intuición sobre lo relevante que es el papel que juega cada atributo a la hora de estimar el precio de un inmueble (imagen 2.9). En este caso, se ve cómo la variable más importante es **dist_centro**, la cual explica entre un 35 % y un 40 % de la variabilidad del precio. Le sigue **ascensor**, con casi un 15 %, mientras que el resto de variables no supera el 10 %.

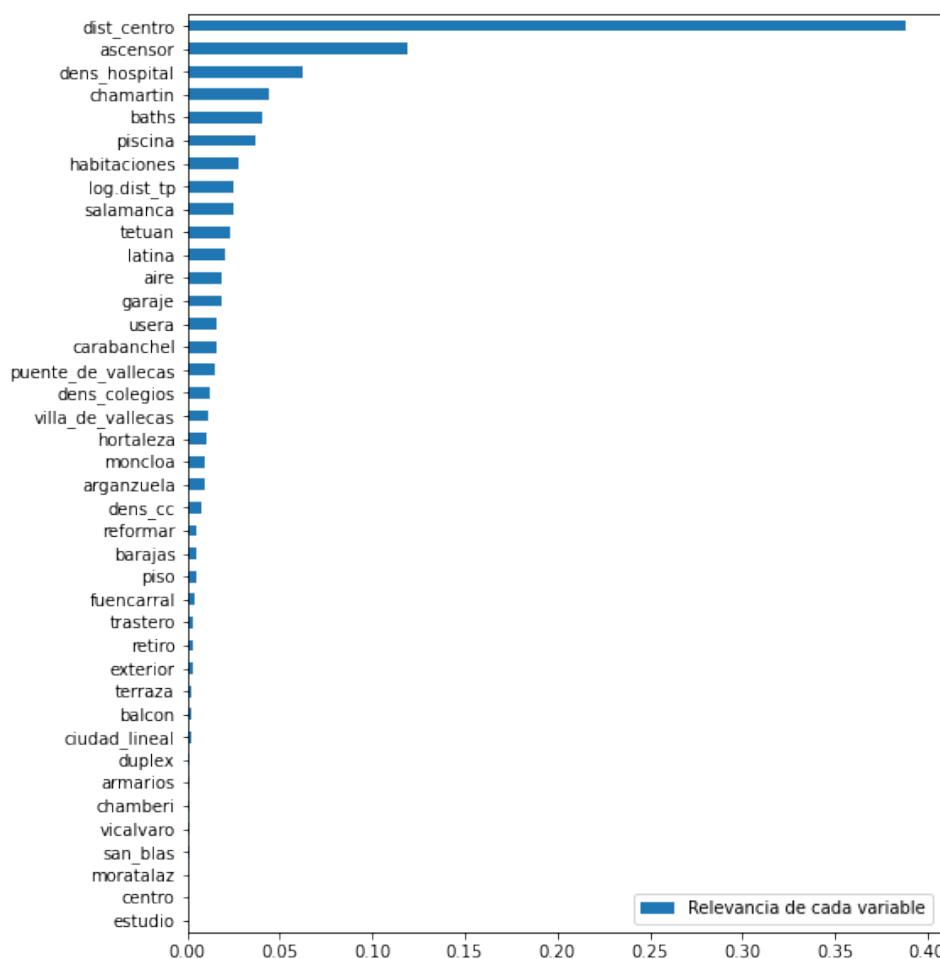


FIGURA 2.9: Relevancia de las variables del modelo GB.

Por último, se realiza un análisis de los efectos espaciales con *SatScan*, tratando de detectar *clusters* en zonas en las que el modelo no funcione del todo bien. Para ello, se examinan los residuos, tal y como se ha explicado en la teoría (ver sección 2.5.3), observándose un *cluster* situado en el sur de

la ciudad y otro más pequeño al este (ver G.4). En conjunto, ambos representan el 9.55 % de las instancias de prueba. Como ya se ha argumentado anteriormente, este resultado puede deberse a que están omitiéndose variables relevantes en dichos *clusters* o a que no se están teniendo en cuenta los efectos espaciales, como la dependencia entre vecinos próximos.

2.6.6. Comparativa final

En la tabla 2.9 se muestra la comparativa final para los distintos algoritmos implementados.

Modelos	\bar{R}^2	<i>I</i> de Moran (p-value)	Jarque-Bera (p-value)
RLM (<i>baseline</i>)	0,769	0,219 ($< 2,2 \cdot 10^{-16}$)	179,898 ($< 2,2 \cdot 10^{-16}$)
RLM espacial	0,780	0,185 ($< 2,2 \cdot 10^{-16}$)	175,875 ($< 2,2 \cdot 10^{-16}$)
MARS	0,789	0,162 ($< 2,2 \cdot 10^{-16}$)	206,945 ($< 2,2 \cdot 10^{-16}$)
SAR	0,813	-0,0132 (0,8971)	365,081 ($< 2,2 \cdot 10^{-16}$)
SEM	0,812	-0,0121 (0,8763)	197,243 ($< 2,2 \cdot 10^{-16}$)
GWR	0,774	0,168 ($< 2,2 \cdot 10^{-16}$)	162,554 ($< 2,2 \cdot 10^{-16}$)
GB	0,822	0,0763 ($6,722 \cdot 10^{-15}$)	494,693 ($< 2,2 \cdot 10^{-16}$)

CUADRO 2.9: Comparativa de resultados sobre el subconjunto de prueba.

Como puede observarse, todos los modelos poseen un coeficiente de determinación ajustado superior al $\bar{R}^2 > 75\%$ sobre una base de datos de prueba, por lo que todos ellos proporcionan una explicación lo suficientemente buena de la variabilidad de la variable dependiente.

En lo referente al poder predictivo de cada uno de ellos, es necesario diferenciar el modelo de *Machine Learning*, *Gradient Boost*, por tratarse de un algoritmo no interpretable. Así, si bien es el que mejores predicciones arroja de todos los modelos implementados, no tenemos conocimiento de cómo está contribuyendo cada variable, es decir, no se conoce el efecto marginal de cada atributo al precio final de la vivienda. En este sentido, se trata de una "caja negra" que puede no ser del todo conveniente si lo que se busca es entender cómo se ve afectado el valor de un inmueble en función de sus características.

De entre las modelizaciones restantes, las cuales sí son interpretables, aquellas con mayor capacidad predictiva son, como era de esperar tras haber estudiado sus residuos, el modelo de retardo espacial y el modelo de error espacial. Estos algoritmos sí que permiten conocer cómo afecta la variación de una variable independiente al resultado final, por lo que son una elección acertada en las situaciones en las que se requiera considerar este tipo de impacto sobre el precio de la vivienda.

Por otra parte, es útil considerar la información recopilada en el cuadro 2.10, en el cual se muestran los resultados del análisis estadístico espacial realizado mediante la herramienta *SatScan*. Conociéndose el número de *clusters* con p-value inferior al 10 %, así como el porcentaje de población que estos representan y su distribución en el espacio, podremos discernir las zonas

problemáticas para cada modelo. De esta forma, por ejemplo, para los modelos interpretables SAR y SEM se tiene que el centro de Madrid es probablemente más propenso a arrojar errores en la predicción, si bien el porcentaje total de la población que representan es bastante bajo (< 5 %). En contraposición, el modelo no interpretable GB es menos fiable al sur y este de la ciudad, comprendiendo en este caso más del doble de observaciones con respecto al SAR y SEM (casi un 10 %).

Modelo	<i>Clusters</i> (p-value < 0,1)	Observaciones <i>test</i>
RLM (<i>baseline</i>)	3	9.32 %
RLM espacial	2	2.64 %
MARS	2	2.58 %
SAR	2	4.15 %
SEM	1	3.03 %
GWR	2	6.51 %
GB	2	9.55 %

CUADRO 2.10: Comparativa de resultados del análisis *SatS-can.*

Capítulo 3

Conclusiones

3.1. Soluciones planteadas y objetivos conseguidos

El principal objetivo de este trabajo ha consistido en implementar una correcta modelización del precio del metro cuadrado en la ciudad de Madrid, para lo cual se ha buscado romper tanto la heterocedasticidad como la dependencia espacial con el objetivo de arrojar predicciones robustas sobre futuras valoraciones de nuevos inmuebles.

Con esta premisa en mente, y remitiéndonos al cuadro 2.9, podemos descartar los modelos RLM, GWR y GB ya que no cumplen con la meta que nos hemos propuesto. En efecto, ninguna de ellas es capaz de vencer los efectos espaciales, por lo que sus estimaciones serán en general sesgadas y poco fiables.

Por su parte, como se ha argumentado previamente, **los modelos SAR y SEM sí consiguen deshacerse de los efectos espaciales**, tal y como se ha concluido tras un análisis de sus residuos. Son, por tanto, los modelos que consiguen alcanzar la finalidad establecida.

En concreto, a la hora de discernir cuál de los dos modelos es más apropiado, puede argumentarse que el poder de predicción del modelo SAR es ligeramente superior (0.5 %), mientras que el modelo SEM parece eliminar más exitosamente la autocorrelación espacial. Por tanto, a priori no existen grandes disimilitudes entre ambos y la elección deberá realizarse basándose en las discrepancias entre resultados –de haberlas– al aplicarse sobre diferentes bases de datos.

3.2. Aplicación real

Una vez hemos alcanzado el objetivo fundamental del proyecto, es decir, tras definir un modelo predictivo capaz de estimar exitosamente el precio de un inmueble en la ciudad de Madrid conocidos sus atributos, procedemos a estudiar la aplicabilidad de nuestro desarrollo en el mercado inmobiliario.

En concreto, conocido el margen de error de la implementación, puede emplearse como una herramienta fidedigna de tasación bajo unos márgenes preestablecidos y evaluarse de antemano el riesgo a la hora de confiar en la valoración del bien inmueble. Por otra parte, la robustez del modelo ante efectos espaciales plantea la posibilidad de hacerlo extrapolable a otras

regiones del ámbito nacional, siendo necesaria una validación previa sobre una muestra del parque inmobiliario en cada caso. Este hecho se vio reflejado en las fortalezas identificadas durante la etapa de planificación, y es una de las ventajas fundamentales a la hora de apostar por un proyecto de este carácter o envergadura.

Por último, si bien los hechos expuestos hasta ahora parecen indicar que existe una buena oportunidad de inversión, con una gran rentabilidad incluso a corto plazo, los encargados finales de determinar la validez y aplicabilidad reales del proyecto serán los *stakeholders* (enunciados en la sección 1.4), pues serán estos los principales interesados y quienes propiciarán que se lleve a cabo la idea.

3.3. Reflexión final: problemas y soluciones

Durante el desarrollo del proyecto se han encontrado problemas y dificultades de diversa índole, los cuales se recopilan a continuación.

3.3.1. Extracción de la base de datos

El portal inmobiliario Idealista, principal fuente de datos, establece una serie de medidas anti-minado en su página *web* que han dificultado la extracción de inmuebles en venta en la ciudad de Madrid.

Así, con la finalidad de construir un proceso automatizado que consiguiera eludir la protección de datos, se ha desarrollado un código recursivo de *webscraping* que, junto con un servicio VPN que posibilitaba rotar la IP de acceso al servidor de Idealista, nos ha permitido descargar el código fuente de los inmuebles que cumplían las características deseadas para posteriormente extraer la información relevante para la modelización.

Asimismo, para afrontar el inconveniente de almacenar información duplicada, se ha incorporado un filtro al algoritmo de *webscraping* que posibilita la discriminación de diferentes copias de un mismo inmueble. De manera análoga, en una posterior etapa de procesamiento de los datos almacenados, se ha procedido a descartar aquellas observaciones que presentaban información claramente errónea que pudiera influir negativamente en la implementación de los algoritmos de predicción.

3.3.2. Obtención de información geoespacial

Para la adquisición de información geográfica relevante, nos hemos valido del proyecto colaborativo [OpenStreetMap](#). De esta manera, se ha ponderado cada observación de la base de datos en relación a su proximidad a la localización de los puntos de interés más relevantes.

El principal obstáculo en lo referente al procesamiento de esta información y su incorporación a la modelización de los algoritmos predictivos ha consistido en su complejidad a la hora de ser codificada en el lenguaje R. Así, la extracción de coordenadas y el cálculo de distancias espaciales no ha sido una tarea trivial debido a la necesidad de crear funciones complejas en R

que precisaban tratar con nuevos tipos de objetos para los cuales la utilización de sus atributos (polígonos, puntos, etc.) ha supuesto una gran curva de aprendizaje.

Una vez conocida la estructura y el funcionamiento de la información geoespacial, se han calculado los atributos relevantes, bien distancias, bien densidades de puntos, para cada vivienda del *dataset* sin mayores dificultades.

3.3.3. Problemas con la modelización

Evaluación con *SatScan*

Para realizar el análisis estadístico espacial con *SatScan*, se ha implementado una llamada *API* a la herramienta desde el IDE *RStudio*. Para ello, se han tenido que usar funciones en R con la finalidad de correr el análisis, devolver los resultados y representarlos gráficamente.

Así, el principal problema ha consistido en el desarrollo de funciones propias complejas y el entendimiento de las llamadas al programa *SatScan*. Además, la implementación requiere de la definición de un número considerable de parámetros –como, por ejemplo, la especificación de las fechas de recopilación de la información almacenada en la base de datos– que a priori no parecen tener relación con el desarrollo planteado pero que son obligatorias aunque el objetivo sea un análisis puramente espacial (como es nuestro caso). De esta forma, la necesidad de usar una herramienta tan avanzada como *SatScan*, adaptándola a nuestros modelos, ha supuesto un gran reto. La cantidad de tiempo y esfuerzo invertidos, así como el conocimiento adquirido, han sido considerables.

Después de haber sido capaces de manejar correctamente la herramienta mediante las funciones definidas, la implementación ha resultado directa y se han obtenido los principales resultados sin mayor inconveniente.

Overfitting con *Gradient Boosting*

El método de *Machine Learning* seleccionado, *Gradient Boosting*, es muy sensible a cambios aplicados en la parametrización. Por este motivo, es necesario llevar a cabo un proceso de *fine tuning* para la selección de los atributos idóneos en la modelización. De no llevarse a cabo este proceso, es probable caer tanto en *underfitting*, cuando el modelo es demasiado simple y no es capaz de explicar la variabilidad del precio de la vivienda, como en *overfitting*, es decir, un sobreajuste a los datos de entrenamiento que ocasiona una pérdida de poder de generalización sobre nuevos subconjuntos de datos.

En nuestro caso en particular, hemos observado cómo una elección de un *learning rate* demasiado alto llevaba a *overfitting*. Lo mismo sucedía al seleccionar un número elevado de árboles de decisión.

La solución ha consistido en la ejecución de una *grid search*, en la cual se han buscado los valores óptimos para los parámetros *learning rate*, número de árboles de decisión y profundidad de dichos árboles de decisión. La selección se ha realizado mediante la comparativa entre los distintos coeficientes de determinación sobre el subconjunto de prueba, hasta finalmente alcanzar el resultado especificado en la sección 2.6.5.

Apéndice A

Variables de la base de datos

A continuación se presenta el diccionario de datos, el cual proporciona el significado de cada variable y, entre paréntesis, su tipo:

- precio (numérica continua): precio expresado en euros.
- metros (numérica continua): número de metros cuadrados.
- habitaciones (numérica discreta): número de habitaciones.
- baths (numérica discreta): número de baños.
- terraza (binaria): indicador de terraza.
- ascensor (binaria): indicador de ascensor.
- aire (binaria): indicador de aire acondicionado.
- garaje (binaria): indicador de garaje.
- piscina (binaria): indicador de piscina.
- trastero (binaria): indicador de trastero.
- armarios (binaria): indicador de armarios.
- reformar (binaria): indicador de vivienda para reformar.
- buen.estado (binaria): indicador de vivienda en buen estado.
- exterior (binaria): indicador de exterior.
- balcon (binaria): indicador de balcón.
- duplex (binaria): indicador de dúplex.
- estudio (binaria): indicador de estudio.
- piso (binaria): indicador de piso.
- atico (binaria): indicador de ático.
- dens_cc (numérica discreta): número de centros comerciales en un radio de 1km
- dens_hospital (numérica discreta): número de hospitales en un radio de 1km
- dist_tp (numérica continua): distancia a la estación de metro o cercanías más cercana.
- dens_colegios (numérica discreta): número de colegios en un radio de 1km.
- dist_centro (numérica continua): distancia al centro de la ciudad.

- distrito (categórica): barrio en el que se sitúa la vivienda.

Apéndice B

Análisis univariante

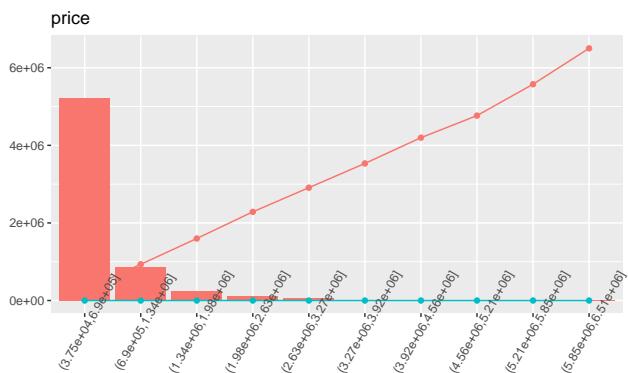
En este apéndice se recoge el análisis univariante de las variables posteriormente utilizadas en los distintos modelizados.

B.1. Variables originales

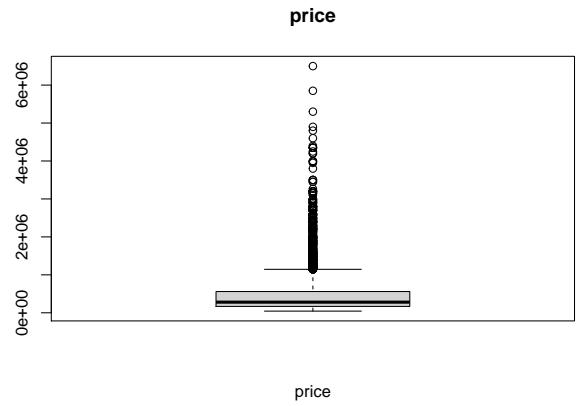
En primer lugar, se representan las variables extraídas de la fuente, sin haberles aplicado ninguna transformación.

B.1.1. Variables numéricas continuas

Se incluye un histograma y un *boxplot* para cada una de las variables numéricas continuas almacenadas en la base de datos.

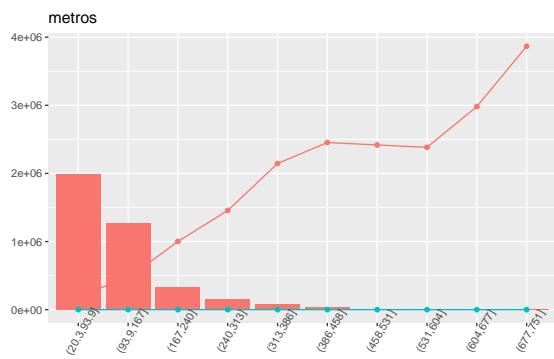


(A) Histograma

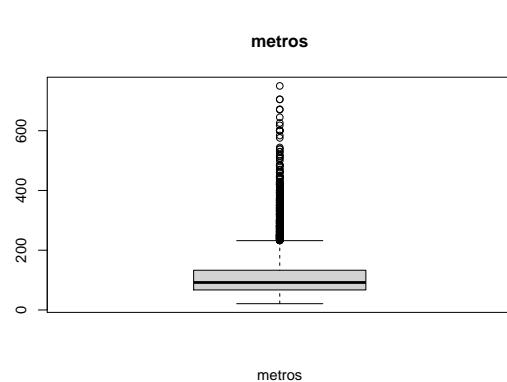


(B) Boxplot

FIGURA B.1: Precio de la vivienda.

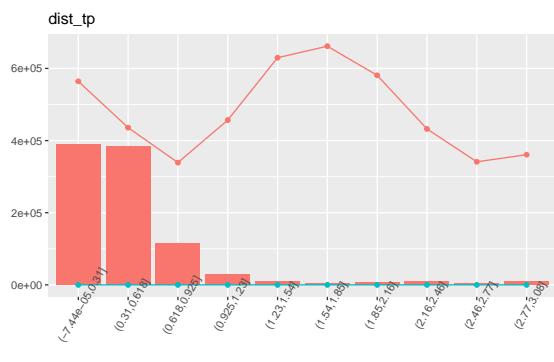


(A) Histograma

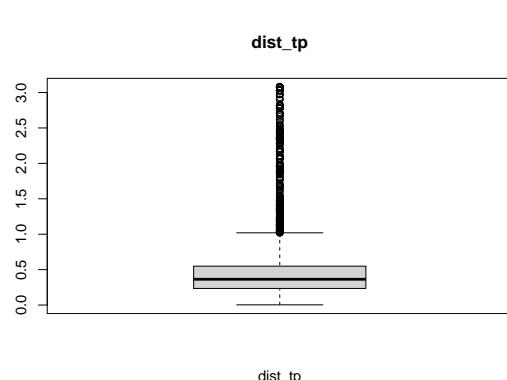


(B) Boxplot

FIGURA B.2: Metros cuadrados de los inmuebles.

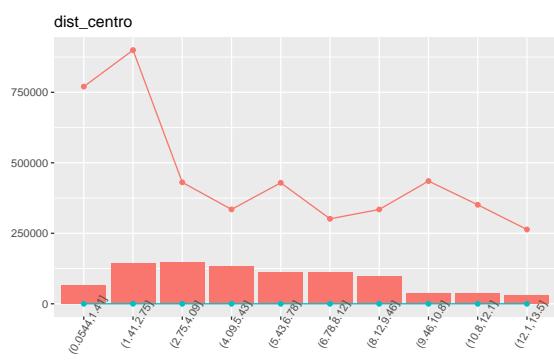


(A) Histograma

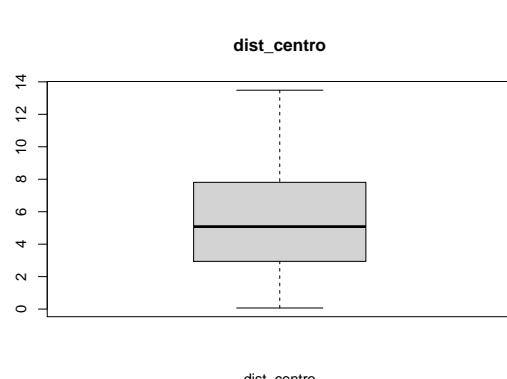


(B) Boxplot

FIGURA B.3: Distancia a metro o cercanías más cercana.



(A) Histograma

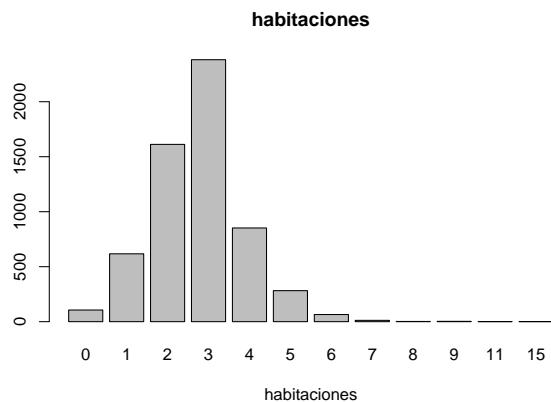


(B) Boxplot

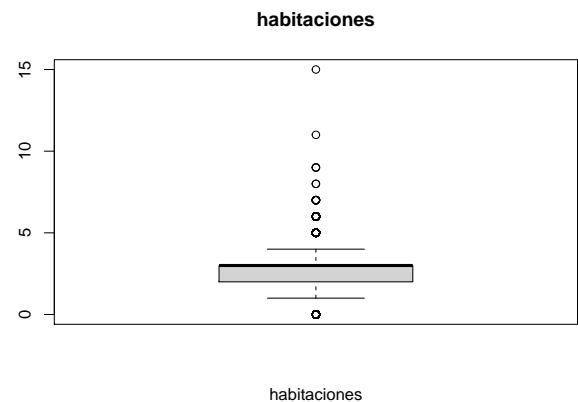
FIGURA B.4: Distancia al centro de la ciudad.

B.1.2. Variables numéricicas discretas

Se muestran, para cada variable numérica discreta, un histograma y un *boxplot*.

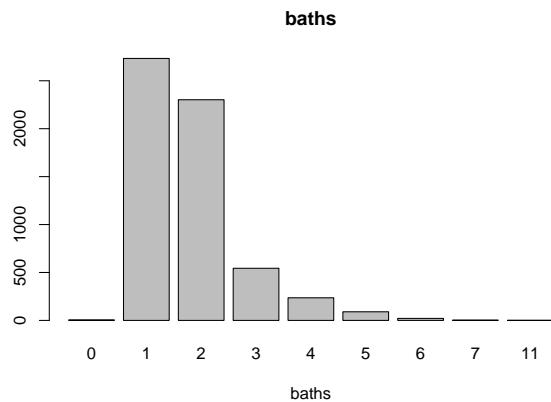


(A) Histograma



(B) Boxplot

FIGURA B.5: Número de habitaciones.

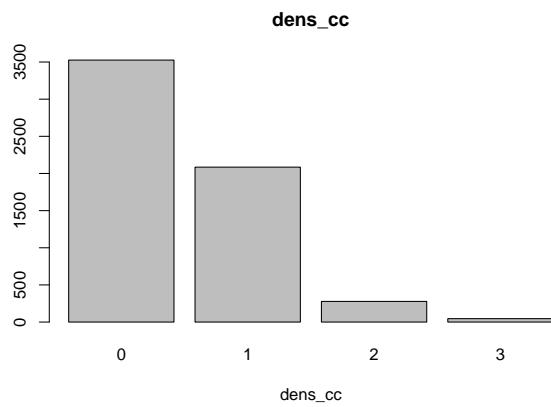


(A) Histograma

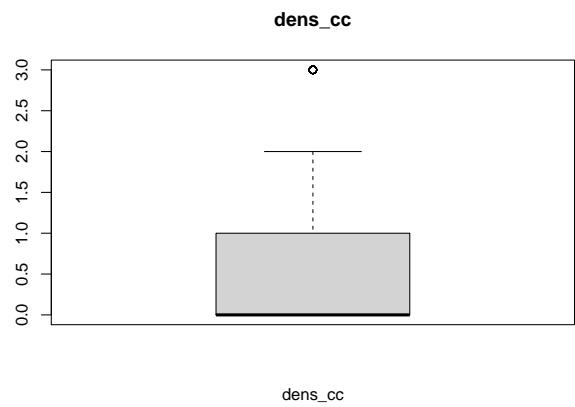


(B) Boxplot

FIGURA B.6: Número de baños y/o aseos.

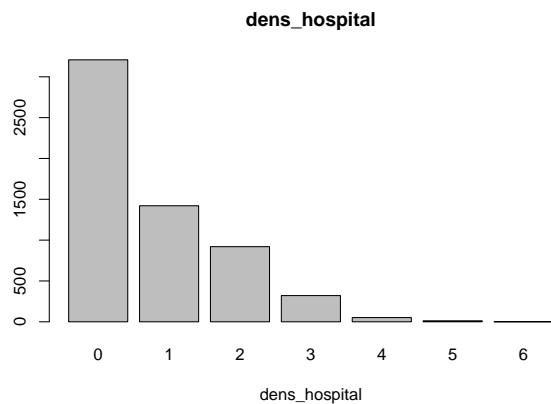


(A) Histograma

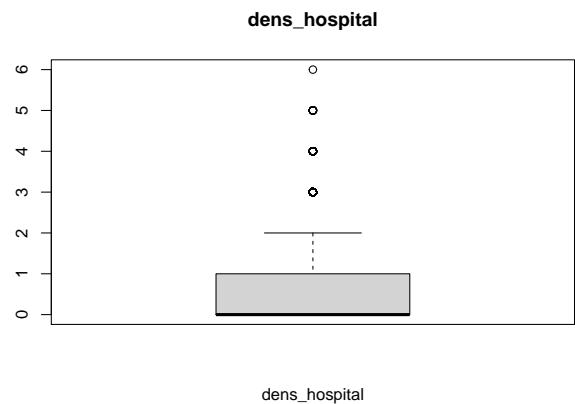


(B) Boxplot

FIGURA B.7: Densidad de centros comerciales.



(A) Histograma



(B) Boxplot

FIGURA B.8: Densidad de hospitales.

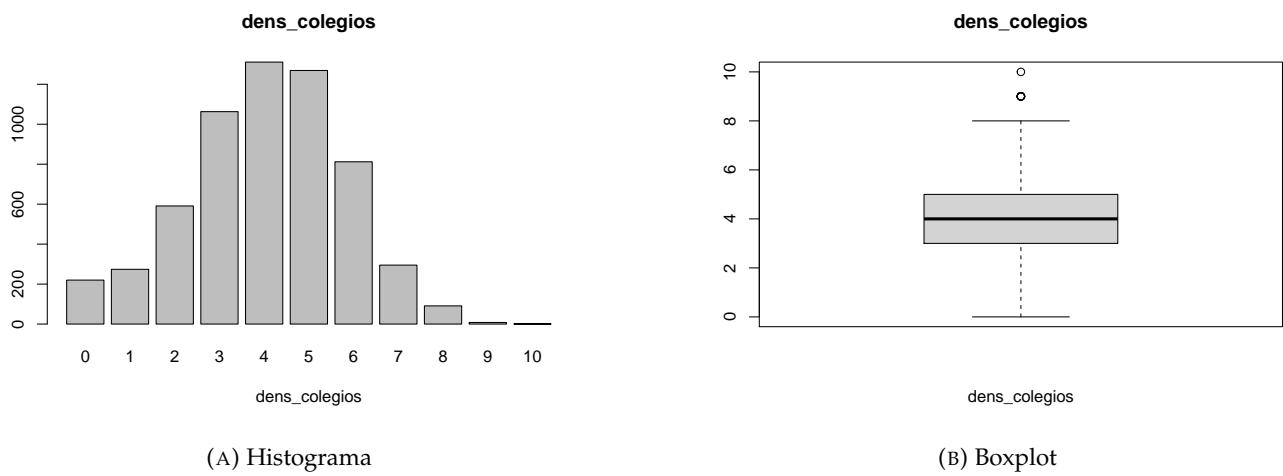


FIGURA B.9: Densidad de colegios.

B.1.3. Variables dicotómicas

Se incluye el desglose por categorías de las variables binarias.

Variable	0	1
terraza	3503	2432
ascensor	1592	4343
aire	2726	3209
garaje	3914	2021
piscina	4595	1340
trastero	3810	2125
armarios	2424	3511
buen.estado	1367	4568
reformar	4966	969
exterior	710	5225
balcon	5199	736
atico	5682	253
duplex	5742	193
estudio	5840	95
piso	540	5395

B.2. Transformaciones

Las transformaciones aplicadas a las variables de la base de datos se enumeran a continuación:

- $\log.pm = \log\left(\frac{\text{precio} (\text{€})}{m^2}\right)$

Logaritmo natural del precio por metro cuadrado.

- $\log.dist_tp = \log(\text{dist_tp})$

Logaritmo natural de la distancia (expresada en kilómetros) a la boca de metro o cercanías más cercana.

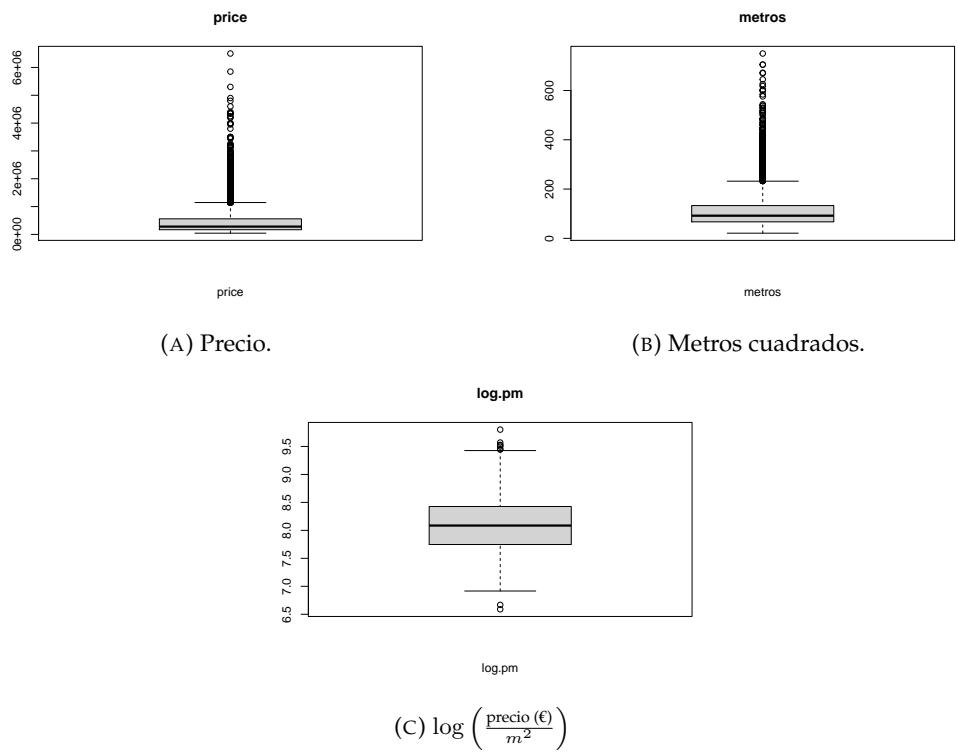


FIGURA B.10: Aplicando el logaritmo natural a la relación entre el precio expresado en euros (A) y los metros cuadrados (B), se consigue una distribución con menos *outliers* (C).

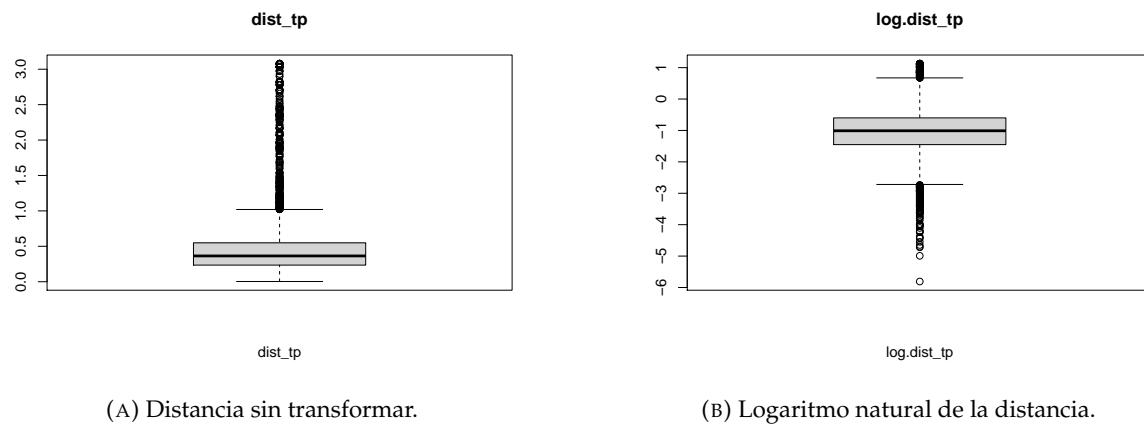


FIGURA B.11: Transformación de la distancia a la estación de metro o cercanías más cercana.

Apéndice C

RLM

C.1. Modelo lineal básico

Residuals:

	Min	1Q	Median	3Q	Max
	-1.28463	-0.13444	0.00133	0.13702	1.04017

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.351244	0.027420	304.566	< 2e-16 ***
habitaciones	-0.067523	0.004568	-14.780	< 2e-16 ***
baths	0.060102	0.006199	9.696	< 2e-16 ***
terraza	-0.019288	0.007651	-2.521	0.011744 *
ascensor	0.130612	0.009622	13.574	< 2e-16 ***
aire	0.047017	0.007842	5.995	2.20e-09 ***
garaje	0.060041	0.010165	5.907	3.77e-09 ***
piscina	0.087429	0.011247	7.774	9.56e-15 ***
trastero	0.021850	0.008832	2.474	0.013399 *
armarios	-0.015642	0.007958	-1.966	0.049399 *
reformar	-0.082663	0.010433	-7.923	2.96e-15 ***
exterior	0.022109	0.011595	1.907	0.056628 .
balcon	0.039990	0.011138	3.591	0.000334 ***
duplex	-0.178455	0.026521	-6.729	1.95e-11 ***
estudio	-0.176368	0.033526	-5.261	1.51e-07 ***
piso	-0.138176	0.018439	-7.494	8.15e-14 ***
distrитobarajas	-0.252243	0.023552	-10.710	< 2e-16 ***
distrитocarabanchel	-0.518984	0.022545	-23.020	< 2e-16 ***
distrитocentro	0.240854	0.022218	10.840	< 2e-16 ***
distrитochamartin	0.212257	0.023561	9.009	< 2e-16 ***
distrитochamberi	0.333923	0.024294	13.745	< 2e-16 ***
distrитociudad-lineal	-0.195384	0.023208	-8.419	< 2e-16 ***
distrитofuencarral	-0.173381	0.023872	-7.263	4.50e-13 ***
distrитohortaleza	-0.157680	0.023290	-6.770	1.47e-11 ***
distrитolatina	-0.443847	0.022858	-19.418	< 2e-16 ***
distrитomoncloa	0.052301	0.022781	2.296	0.021736 *
distrитomoratalaz	-0.348086	0.023556	-14.777	< 2e-16 ***
distrитopuente-de-vallecas	-0.629362	0.022682	-27.748	< 2e-16 ***
distrитoretiro	0.218994	0.023603	9.278	< 2e-16 ***
distrитosalamanca	0.480287	0.023855	20.133	< 2e-16 ***
distrитosan-blas	-0.429868	0.023244	-18.494	< 2e-16 ***
distrитotetuan	-0.041660	0.023136	-1.801	0.071834 .
distrитousera	-0.585896	0.022274	-26.304	< 2e-16 ***
distrитovicalvaro	-0.503282	0.022911	-21.966	< 2e-16 ***
distrитovilla-de-vallecas	-0.574140	0.023259	-24.685	< 2e-16 ***
distrитovillaverde	-0.741351	0.022437	-33.042	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2272 on 4118 degrees of freedom
 Multiple R-squared: 0.7614, Adjusted R-squared: 0.7594
 F-statistic: 375.6 on 35 and 4118 DF, p-value: < 2.2e-16

FIGURA C.1: Parámetros del modelo lineal básico.

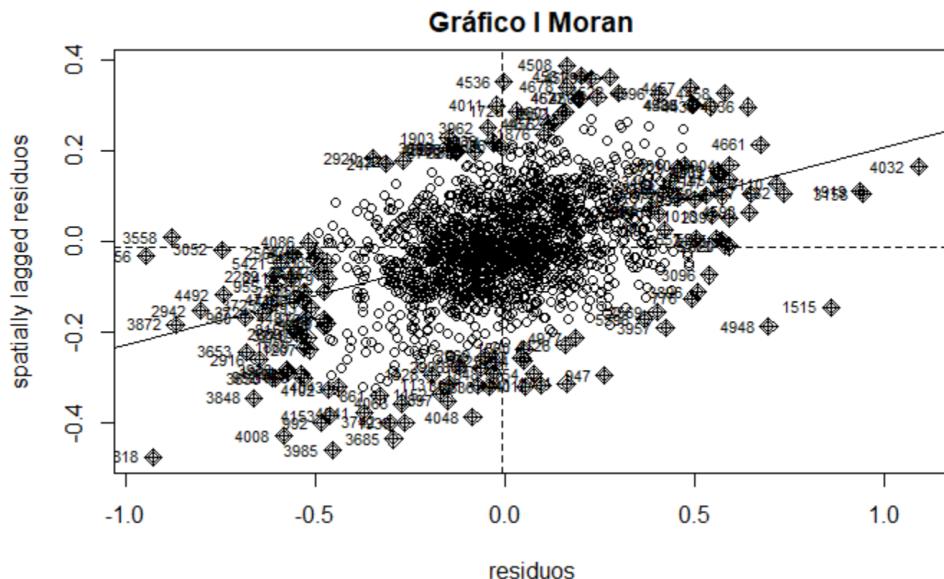


FIGURA C.2: Gráfico *I de Moran* para el modelo lineal básico.

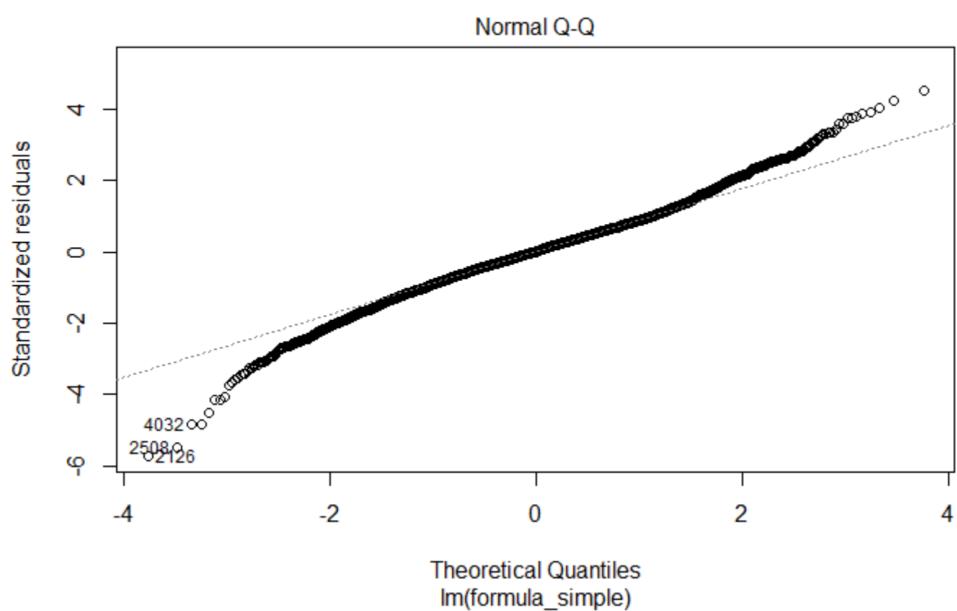


FIGURA C.3: *Q-Q plot* para el modelo lineal básico.

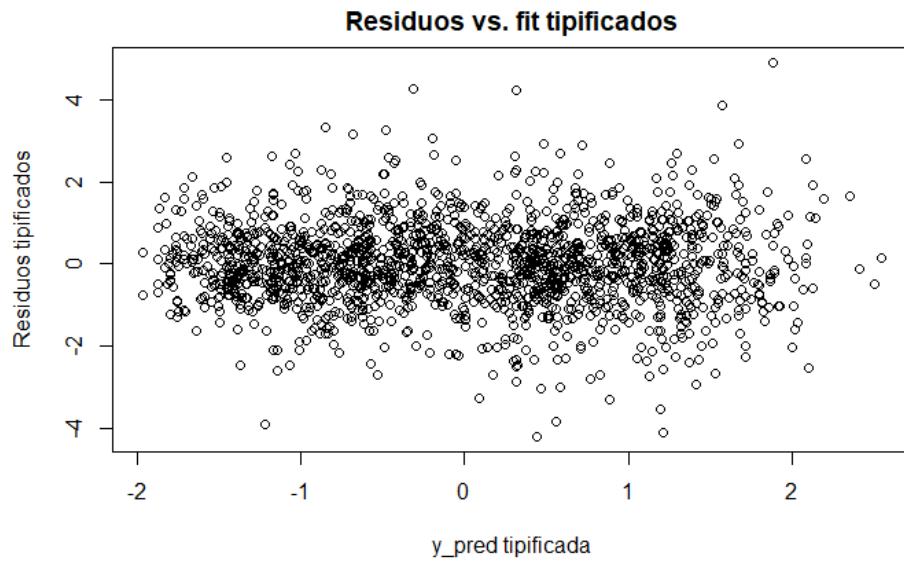


FIGURA C.4: Gráfico de residuos para el modelo lineal básico.

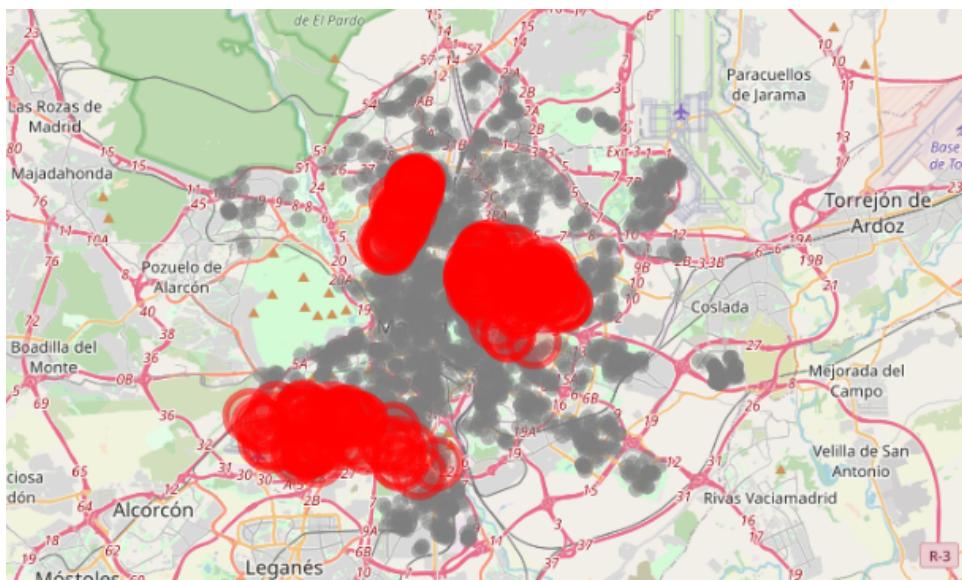


FIGURA C.5: Mapa SatScan para el modelo lineal básico.
Clusters con p-value <0.1.

C.2. Modelo lineal espacial

Residuals:

	Min	1Q	Median	3Q	Max
	-1.31180	-0.12760	0.00379	0.13615	0.98006

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.454682	0.030854	274.024	< 2e-16 ***
habitaciones	-0.069138	0.004427	-15.617	< 2e-16 ***
baths	0.054344	0.006031	9.010	< 2e-16 ***
terraza	-0.012320	0.007433	-1.657	0.097497 .
ascensor	0.127592	0.009323	13.686	< 2e-16 ***
aire	0.042521	0.007600	5.595	2.35e-08 ***
garaje	0.077306	0.009938	7.779	9.20e-15 ***
piscina	0.118554	0.011188	10.597	< 2e-16 ***
trastero	0.031339	0.008576	3.654	0.000261 ***
armarios	-0.019721	0.007712	-2.557	0.010589 *
reformar	-0.089518	0.010114	-8.851	< 2e-16 ***
exterior	0.022849	0.011232	2.034	0.041996 *
balcon	0.033359	0.010802	3.088	0.002027 **
duplex	-0.171649	0.025697	-6.680	2.71e-11 ***
estudio	-0.195602	0.032492	-6.020	1.90e-09 ***
piso	-0.131748	0.017871	-7.372	2.02e-13 ***
distrítobarajas	0.116553	0.035407	3.292	0.001004 **
distrítocarabanchel	-0.388395	0.023526	-16.509	< 2e-16 ***
distrítocentro	0.210549	0.021704	9.701	< 2e-16 ***
distrítochamartin	0.301881	0.026642	11.331	< 2e-16 ***
distrítochamberí	0.353695	0.025442	13.902	< 2e-16 ***
distrítociudad-lineal	-0.051670	0.026144	-1.976	0.048175 *
distrítofuencarral	0.044620	0.030367	1.469	0.141820
distrítohortaleza	0.061059	0.029779	2.050	0.040390 *
distrítolatina	-0.357933	0.023973	-14.931	< 2e-16 ***
distrítomoncloa	0.159269	0.023713	6.717	2.11e-11 ***
distrítomoratalaz	-0.174662	0.025509	-6.847	8.66e-12 ***
distrítopuente-de-vallecas	-0.495705	0.024277	-20.418	< 2e-16 ***
distrítoretiro	0.266448	0.023348	11.412	< 2e-16 ***
distrítosalamanca	0.500784	0.024943	20.077	< 2e-16 ***
distrítosan-blas	-0.189404	0.028558	-6.632	3.73e-11 ***
distrítotetuan	0.086357	0.025176	3.430	0.000609 ***
distrítousera	-0.491404	0.022844	-21.511	< 2e-16 ***
distrítovicalvaro	-0.218729	0.032158	-6.802	1.18e-11 ***
distrítovilla-de-vallecas	-0.303711	0.030450	-9.974	< 2e-16 ***
distrítovillaverde	-0.509186	0.027404	-18.581	< 2e-16 ***
dens_cc	0.042421	0.006345	6.686	2.60e-11 ***
dens_hospital	0.014123	0.004850	2.912	0.003610 **
log.dist_tp	0.001918	0.005420	0.354	0.723472
dens_colegios	-0.013983	0.002628	-5.321	1.09e-07 ***
dist_centro	-0.039477	0.003154	-12.518	< 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.22 on 4113 degrees of freedom
 Multiple R-squared: 0.7766, Adjusted R-squared: 0.7745
 F-statistic: 357.5 on 40 and 4113 DF, p-value: < 2.2e-16

FIGURA C.6: Parámetros del modelo lineal espacial.

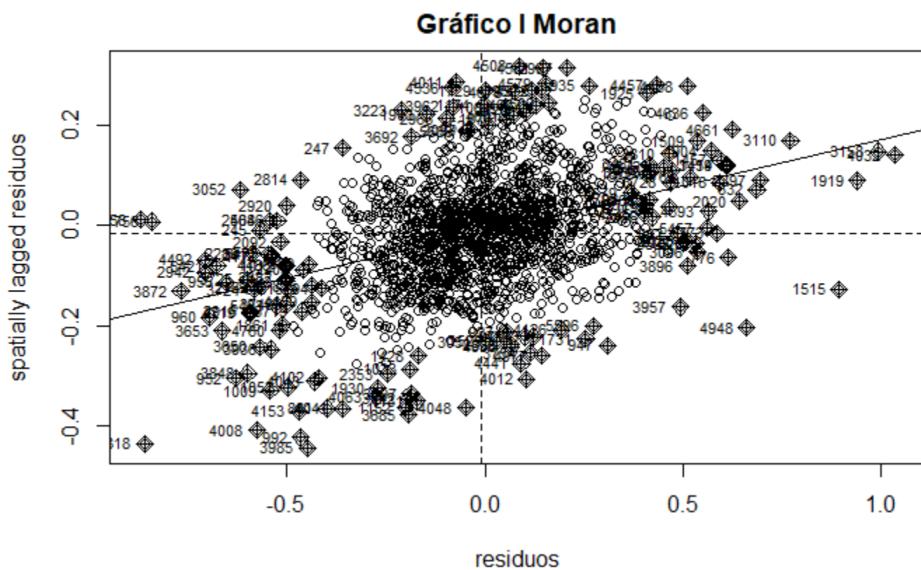


FIGURA C.7: Gráfico *I de Moran* para el modelo lineal espacial.

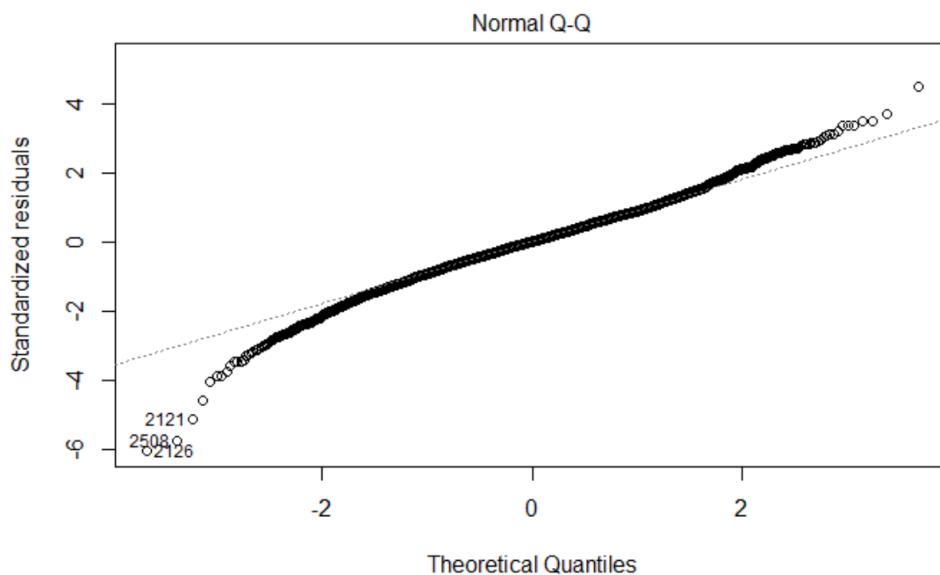


FIGURA C.8: *Q-Q plot* para el modelo lineal espacial.

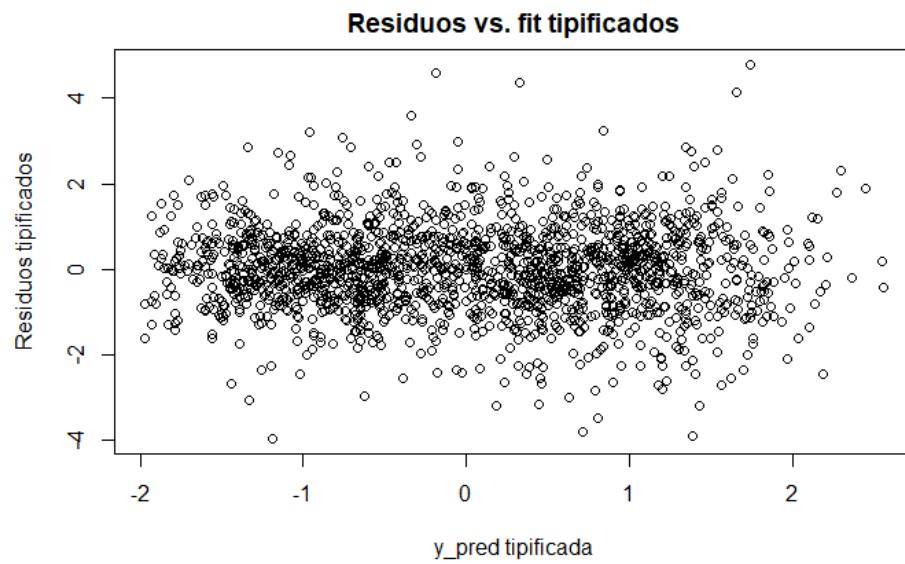


FIGURA C.9: Gráfico de residuos para el modelo lineal espacial.

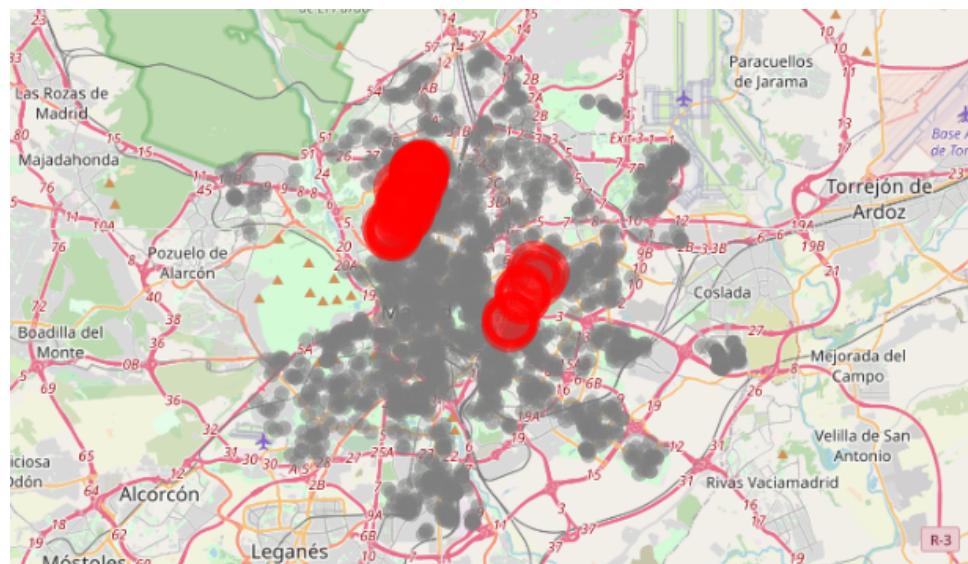


FIGURA C.10: Mapa *SatScan* para el modelo lineal espacial.
Clusters con $p\text{-value} < 0.1$.

C.3. Multiadaptive regression splines

Residuals:

	Min	1Q	Median	3Q	Max
	-1.25263	-0.12525	0.00143	0.12850	0.87280

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.864677	0.035009	224.649	< 2e-16 ***
habitaciones_hasta_3	0.069307	0.005873	11.801	< 2e-16 ***
habitaciones_desde_3	-0.058744	0.007192	-8.168	4.12e-16 ***
baths_hasta_2	-0.027356	0.009600	-2.850	0.004398 **
baths_desde_2	0.056888	0.007472	7.613	3.30e-14 ***
terraza	-0.011608	0.007230	-1.605	0.108462
ascensor	0.126175	0.009254	13.635	< 2e-16 ***
aire	0.048478	0.007414	6.539	6.95e-11 ***
garaje	0.079922	0.009765	8.184	3.62e-16 ***
piscina	0.121410	0.011998	10.119	< 2e-16 ***
zonas.verdes	-0.006768	0.010733	-0.631	0.528328
trastero	0.027157	0.008362	3.248	0.001173 **
armarios	-0.013001	0.007606	-1.709	0.087474 .
reformar	-0.090901	0.009862	-9.217	< 2e-16 ***
exterior	0.029000	0.010934	2.652	0.008028 **
balcon	0.026825	0.010515	2.551	0.010775 *
duplex	-0.161633	0.025011	-6.463	1.15e-10 ***
estudio	-0.232986	0.032617	-7.143	1.07e-12 ***
piso	-0.133665	0.017391	-7.686	1.89e-14 ***
distrítobarajas	0.079836	0.034530	2.312	0.020824 *
distrítocarabanchel	-0.242039	0.024835	-9.746	< 2e-16 ***
distrítocentro	0.073667	0.022982	3.205	0.001359 **
distrítochamartín	0.459596	0.027964	16.435	< 2e-16 ***
distrítochamberí	0.344085	0.024777	13.887	< 2e-16 ***
distrítociudad-lineal	0.114030	0.027754	4.109	4.06e-05 ***
distrítofuencarral	0.115092	0.029929	3.845	0.000122 ***
distrítohortaleza	0.134924	0.029386	4.591	4.53e-06 ***
distrítolatina	-0.227672	0.024896	-9.145	< 2e-16 ***
distrítomoncloa	0.211130	0.023402	9.022	< 2e-16 ***
distrítomoratalaz	0.016351	0.027722	0.590	0.555330
distrítopuente-de-vallecas	-0.301251	0.026831	-11.228	< 2e-16 ***
distrítoretiro	0.299435	0.022801	13.133	< 2e-16 ***
distrítosalamanca	0.530319	0.024360	21.770	< 2e-16 ***
distrítosan-blas	-0.098159	0.028451	-3.450	0.000566 ***
distrítotetuán	0.261283	0.027212	9.602	< 2e-16 ***
distrítousera	-0.324525	0.024765	-13.104	< 2e-16 ***
distrítovicalvaro	-0.188366	0.031341	-6.010	2.01e-09 ***
distrítovilla-de-vallecas	-0.253169	0.029843	-8.483	< 2e-16 ***
distrítovillaverde	-0.395899	0.027744	-14.270	< 2e-16 ***
dens_cc	0.039509	0.006184	6.389	1.85e-10 ***
dens_hospital	0.022262	0.004749	4.687	2.86e-06 ***
log.dist_tp	-0.004851	0.005288	-0.917	0.359005
dens_colegios	-0.003265	0.002646	-1.234	0.217346
centro_hasta_4.394	0.144841	0.007586	19.093	< 2e-16 ***
centro_desde_4.394	0.003008	0.004109	0.732	0.464232

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2139 on 4109 degrees of freedom
 Multiple R-squared: 0.7892, Adjusted R-squared: 0.7869
 F-statistic: 349.6 on 44 and 4109 DF, p-value: < 2.2e-16

FIGURA C.11: Parámetros del modelo con no linealidades.

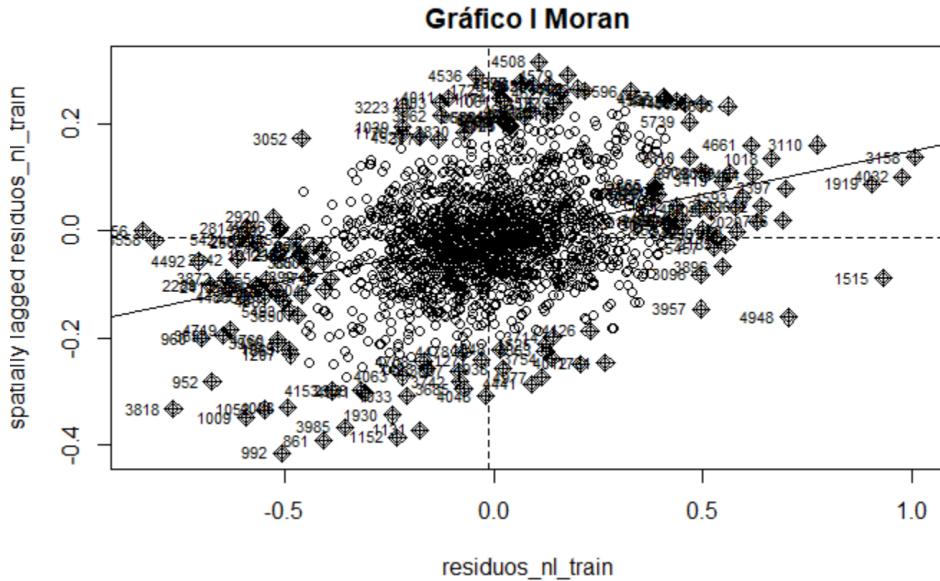


FIGURA C.12: Gráfico *I de Moran* para el modelo con no linealidades.

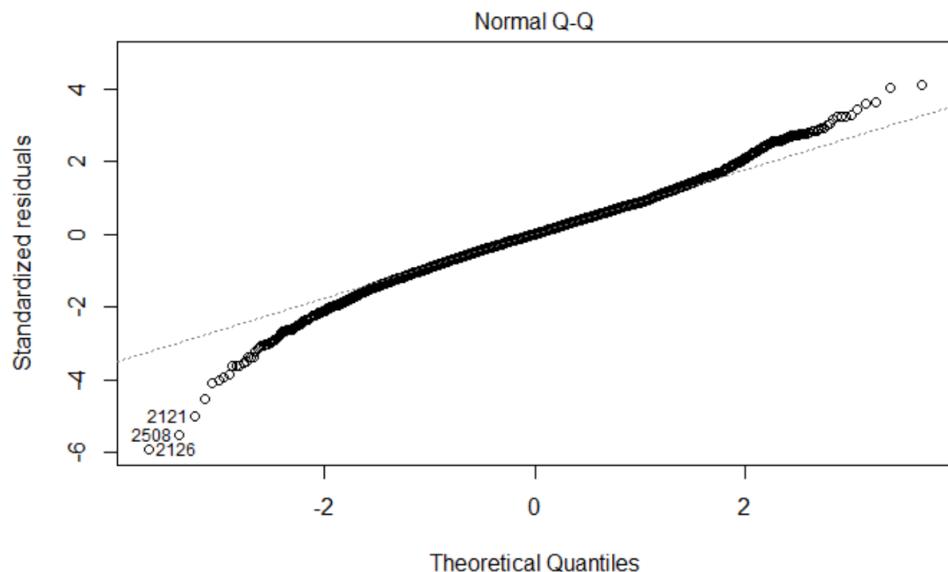


FIGURA C.13: *Q-Q plot* para el modelo con no linealidades.

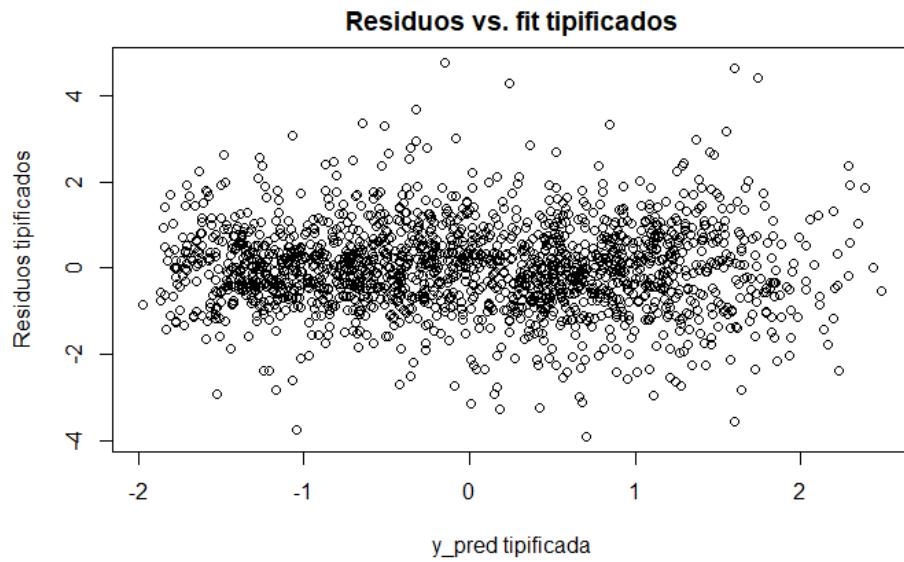


FIGURA C.14: Gráfico de residuos para el modelo con no linealidades.

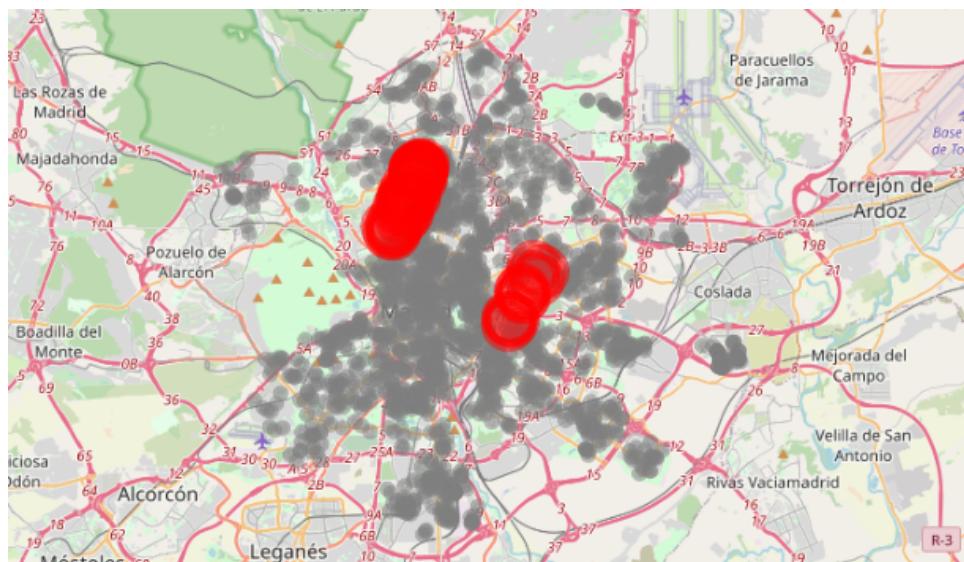


FIGURA C.15: Mapa SatScan para el modelo con no linealidades. *Clusters con p-value <0.1*.

Apéndice D

SAR

```

Residuals:
    Min      1Q   Median      3Q     Max
-1.3308877 -0.1134820  0.0018913  0.1183338  0.8296725

Type: lag
Coefficients: (asymptotic standard errors)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.8893806 0.1398051 27.8200 < 2.2e-16
habitaciones -0.0559590 0.0039411 -14.1989 < 2.2e-16
baths         0.0328888 0.0053506  6.1467 7.910e-10
terraza        -0.0065705 0.0066075 -0.9944 0.3200235
ascensor       0.0993678 0.0083789 11.8593 < 2.2e-16
aire           0.0366261 0.0067573  5.4203 5.951e-08
garaje          0.0539027 0.0088338  6.1019 1.048e-09
piscina         0.0915634 0.0101269  9.0416 < 2.2e-16
trastero        0.0240307 0.0076626  3.1361 0.0017122
armarios        -0.0119619 0.0068239 -1.7529 0.0796120
reformar        -0.0873927 0.0090038 -9.7062 < 2.2e-16
exterior         0.0327866 0.0100491  3.2626 0.0011038
balcon          0.0254299 0.0096242  2.6423 0.0082349
duplex          -0.1644304 0.0225147 -7.3033 2.809e-13
estudio          -0.1763958 0.0301049 -5.8594 4.646e-09
piso             -0.1362733 0.0155959 -8.7378 < 2.2e-16
dens_cc          0.0223012 0.0057504  3.8782 0.0001052
dens_hospital   0.0022541 0.0043797  0.5147 0.6067874
log_dist_tp     -0.0036398 0.0047463 -0.7669 0.4431630
dens_colegios   -0.0025553 0.0023438 -1.0902 0.2756096
dist_centro      -0.0237921 0.0028877 -8.2390 2.220e-16
arganzuela      0.2131380 0.0261810  8.1410 4.441e-16
barajas          0.2940766 0.0245048 12.0008 < 2.2e-16
carabanchel     0.0350349 0.0209342  1.6736 0.0942149
centro           0.2994511 0.0292932 10.2226 < 2.2e-16
chamartin        0.3639192 0.0280808 12.9597 < 2.2e-16
chamberi         0.3702881 0.0305659 12.1144 < 2.2e-16
ciudad_lineal   0.1978959 0.0231729  8.5400 < 2.2e-16
fuencarral       0.2332124 0.0237743  9.8094 < 2.2e-16
hortaleza        0.2268209 0.0231186  9.8112 < 2.2e-16
latina           0.0750348 0.0211014  3.5559 0.0003767
moncloa          0.2843827 0.0256519 11.0862 < 2.2e-16
moratalaz       0.1598701 0.0222865  7.1734 7.316e-13
puente_de_vallecas -0.0224672 0.0200586 -1.1201 0.2626819
retiro           0.3532170 0.0283972 12.4385 < 2.2e-16
salamanca        0.4651952 0.0318225 14.6184 < 2.2e-16
san_blas          0.1413026 0.0208683  6.7712 1.278e-11
tetuan           0.2568948 0.0241523 10.6364 < 2.2e-16
usera            -0.0047997 0.0208607 -0.2301 0.8180253
vicalvaro         0.1358375 0.0208480  6.5156 7.239e-11
villa_de_vallecas  0.0878420 0.0203990  4.3062 1.661e-05

Rho: 0.52152, LR test value: 732.21, p-value: < 2.22e-16
Asymptotic standard error: 0.017438
z-value: 29.908, p-value: < 2.22e-16
wald statistic: 894.47, p-value: < 2.22e-16

Log likelihood: 813.9156 for lag model
ML residual variance (sigma squared): 0.038438, (sigma: 0.19606)
Number of observations: 4154
Number of parameters estimated: 43
AIC: -1541.8, (AIC for lm: -811.62)
LM test for residual autocorrelation
test value: 0.3213, p-value: 0.57082

```

FIGURA D.1: Parámetros del modelo de retardo espacial.

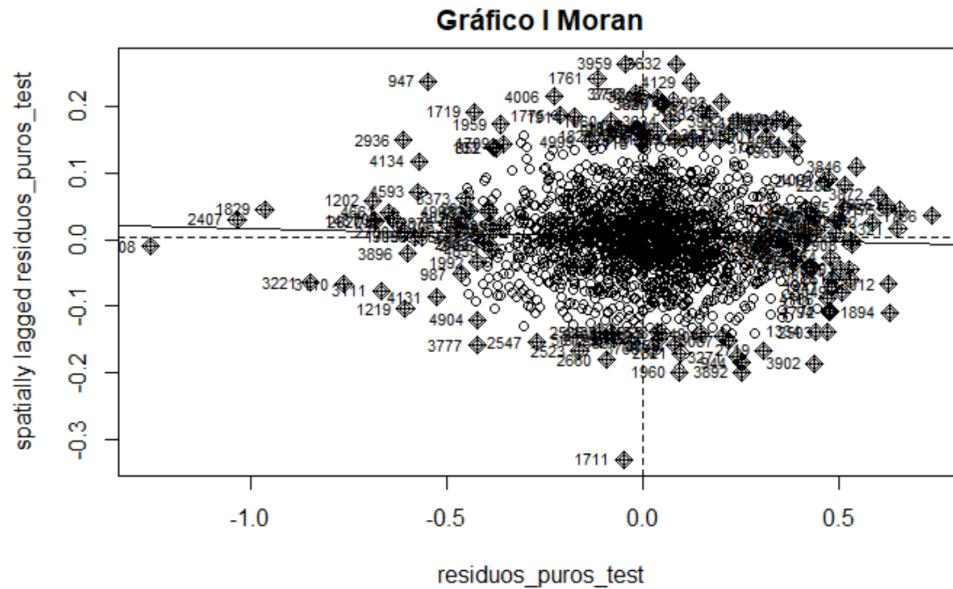


FIGURA D.2: Gráfico *I de Moran* para el modelo de retardo espacial.

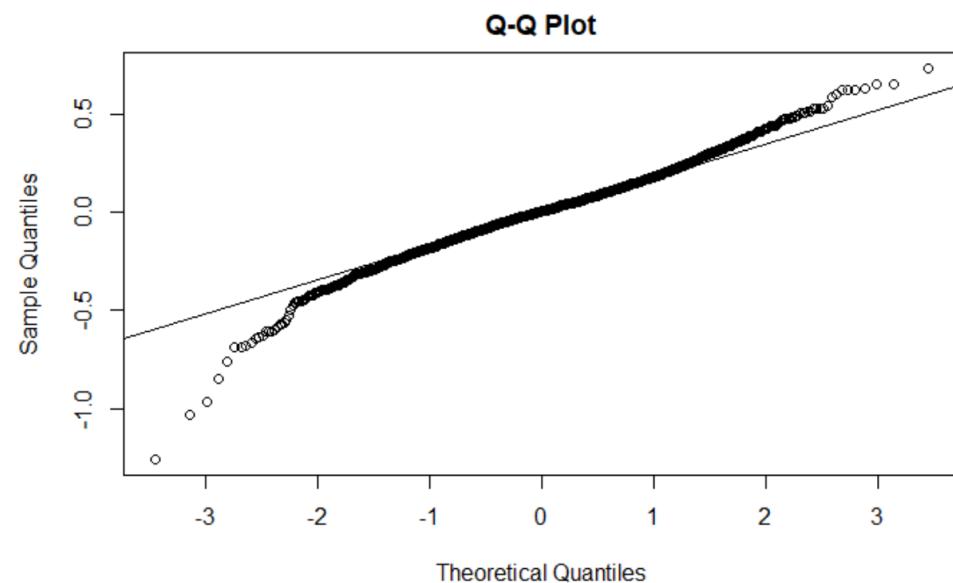


FIGURA D.3: Gráfico *QQ plot* para el modelo de error espacial.

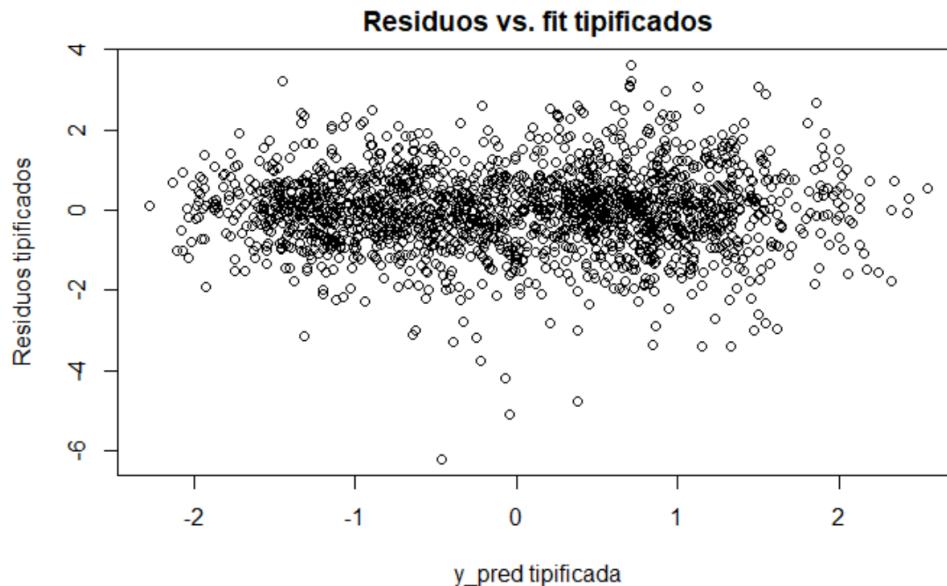


FIGURA D.4: Gráfico de residuos para el modelo de retardo espacial.

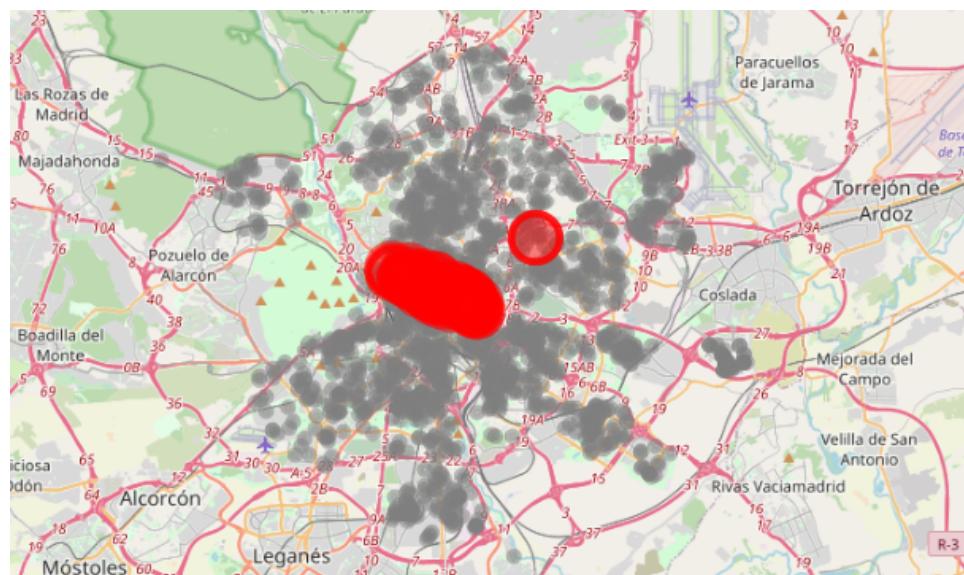


FIGURA D.5: Mapa SatScan para el modelo de retardo espacial. Clusters con p-value <0.1.

Apéndice E

SEM

```

Residuals:
      Min        1Q     Median      3Q      Max
-1.2817819 -0.1174442  0.0073733  0.1198824  0.9062334

Type: error
Coefficients: (asymptotic standard errors)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 7.9675199 0.0608459 130.9458 < 2.2e-16
habitaciones -0.0594414 0.0040256 -14.7660 < 2.2e-16
baths         0.0227806 0.0055813  4.0816 4.473e-05
terraza       -0.0101833 0.0067279 -1.5136 0.1301265
ascensor       0.1139527 0.0088470 12.8803 < 2.2e-16
aire           0.0391704 0.0068717  5.7003 1.196e-08
garaje          0.0799636 0.0091683  8.7217 < 2.2e-16
piscina         0.0967280 0.0111442  8.6796 < 2.2e-16
trastero        0.0241300 0.0078694  3.0663 0.0021671
armarios        -0.0096764 0.0069730 -1.3877 0.1652300
reformar        -0.1029102 0.0090423 -11.3810 < 2.2e-16
exterior         0.0261648 0.0100274  2.6093 0.0090718
balcon          0.0218590 0.0096245  2.2712 0.0231359
duplex          -0.1459692 0.0231318 -6.3103 2.784e-10
estudio          -0.1981760 0.0290542 -6.8209 9.047e-12
piso             -0.1317709 0.0160394 -8.2155 2.220e-16
dens_cc          0.0265696 0.0099218  2.6779 0.0074087
dens_hospital   0.0234791 0.0076183  3.0819 0.0020565
log.dist_tp      -0.0015672 0.0069996 -0.2239 0.8228362
dens_colegios   -0.0094127 0.0037885 -2.4846 0.0129708
dist_centro      -0.0411878 0.0054813 -7.5142 5.729e-14
arganzuela       0.5187164 0.0535890  9.6795 < 2.2e-16
barajas          0.6344644 0.0532774 11.9087 < 2.2e-16
carabanchel     0.1258545 0.0495977  2.5375 0.0111645
centro           0.7090361 0.0544149 13.0302 < 2.2e-16
chamartin        0.8180061 0.0495977 16.4928 < 2.2e-16
chamberi         0.8148408 0.0543574 14.9904 < 2.2e-16
ciudad_lineal   0.4976418 0.0479545 10.3774 < 2.2e-16
fuencarral       0.5738665 0.0501959 11.4325 < 2.2e-16
hortaleza        0.5694439 0.0486217 11.7117 < 2.2e-16
latina           0.1524045 0.0500524  3.0449 0.0023276
moncloa          0.6830975 0.0482698 14.1517 < 2.2e-16
moratalaz        0.3386704 0.0509158  6.6516 2.900e-11
puente_de_vallecas 0.0376582 0.0474467  0.7937 0.4273727
retiro           0.7406831 0.0538121 13.7642 < 2.2e-16
salamanca        0.9554165 0.0530990 17.9931 < 2.2e-16
san_blas          0.3598003 0.0479149  7.5092 5.951e-14
tetuan           0.6196996 0.0494958 12.5202 < 2.2e-16
usera             0.0011292 0.0490076  0.0230 0.9816172
vicalvaro         0.3080728 0.0501301  6.1455 7.973e-10
villa_de_vallecas 0.1776465 0.0494488  3.5925 0.0003275

Lambda: 0.61055, LR test value: 630.78, p-value: < 2.22e-16
Asymptotic standard error: 0.019319
z-value: 31.604, p-value: < 2.22e-16
Wald statistic: 998.81, p-value: < 2.22e-16

Log likelihood: 730.95 for error model
ML residual variance (sigma squared): 0.039451, (sigma: 0.19862)
Number of observations: 4154
Number of parameters estimated: 43
AIC: -1375.9, (AIC for lm: -747.12)

```

FIGURA E.1: Parámetros del modelo de error espacial.

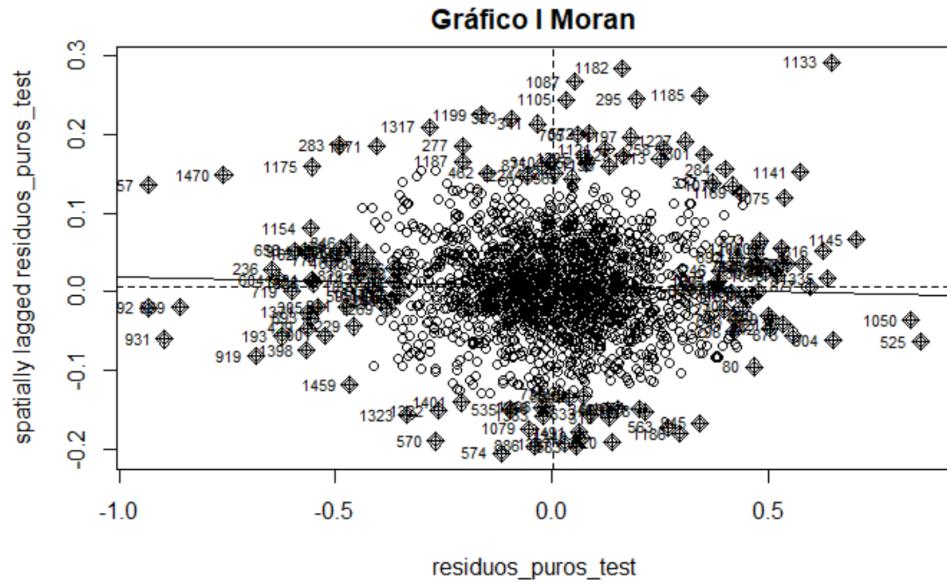


FIGURA E.2: Gráfico *I de Moran* para el modelo de error espacial.

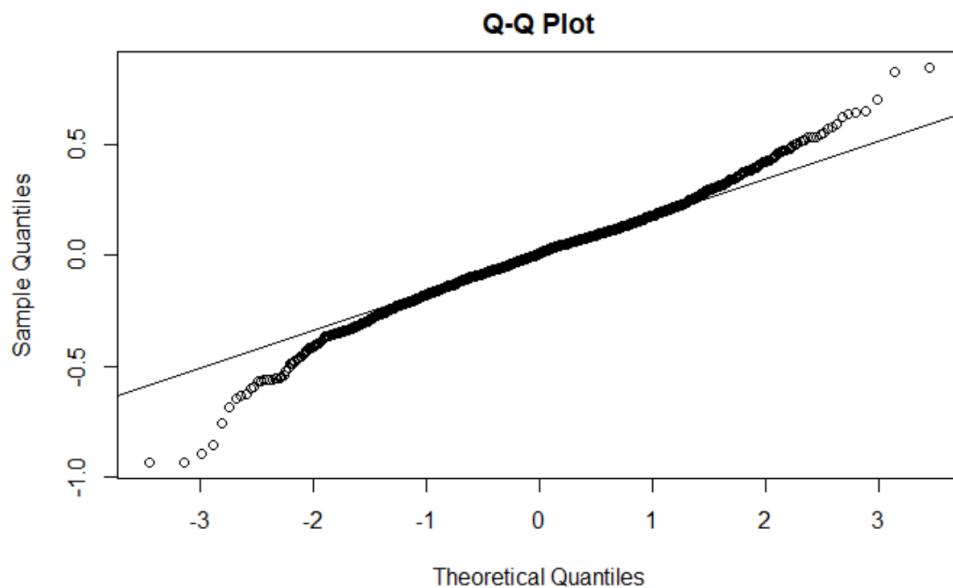


FIGURA E.3: Gráfico *QQ plot* para el modelo de error espacial.

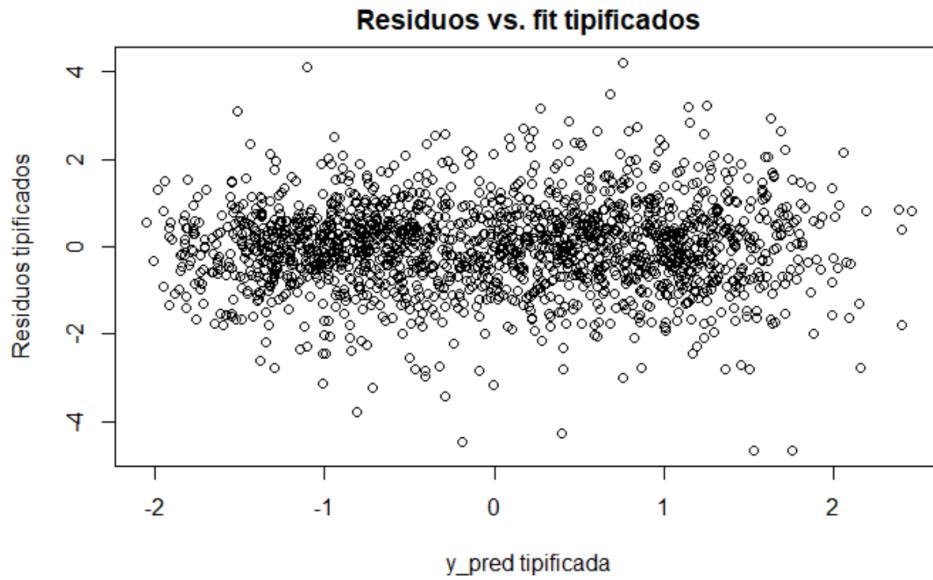


FIGURA E.4: Gráfico de residuos para el modelo de error espacial.

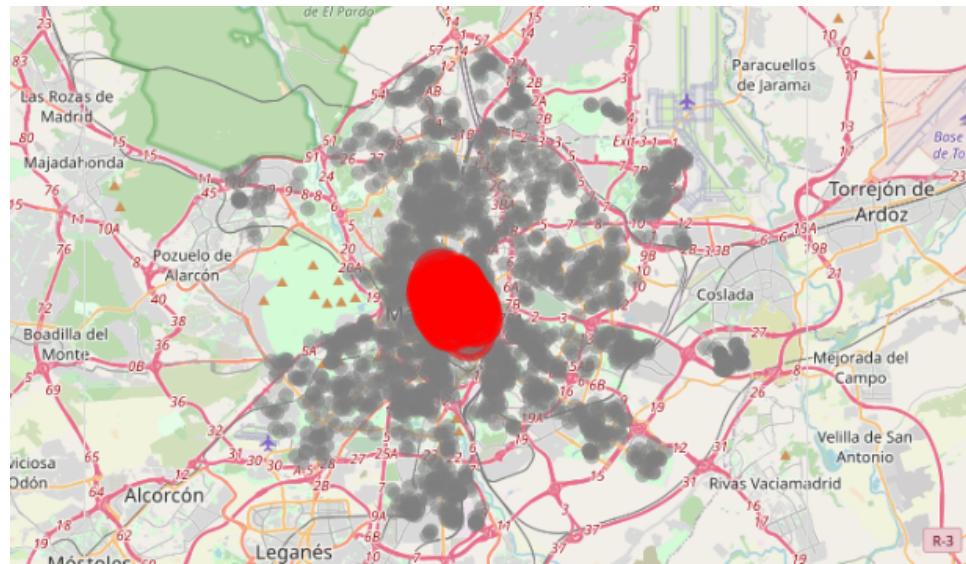


FIGURA E.5: Mapa SatScan para el modelo de error espacial.
Clusters con $p\text{-value} < 0.1$.

Apéndice F

GWR

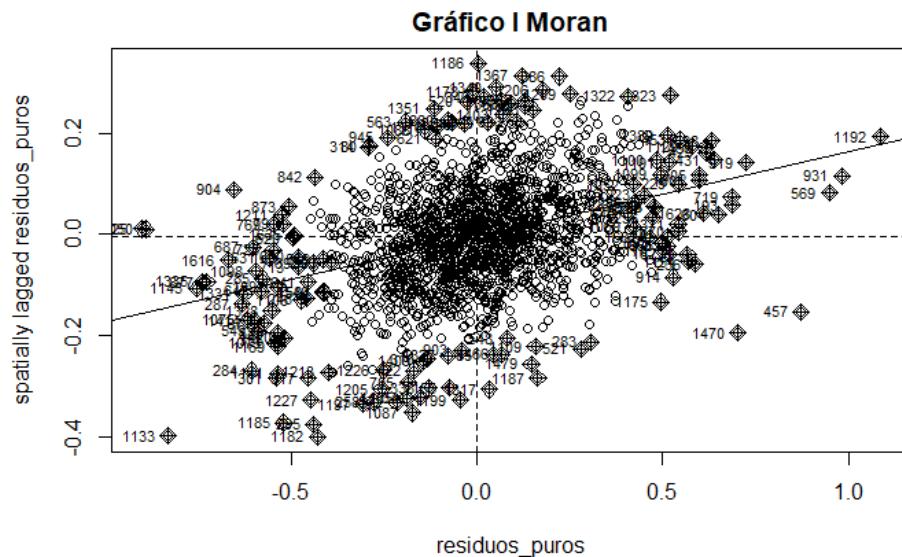


FIGURA F.1: Gráfico *I de Moran* para el modelo geográficamente ponderado.

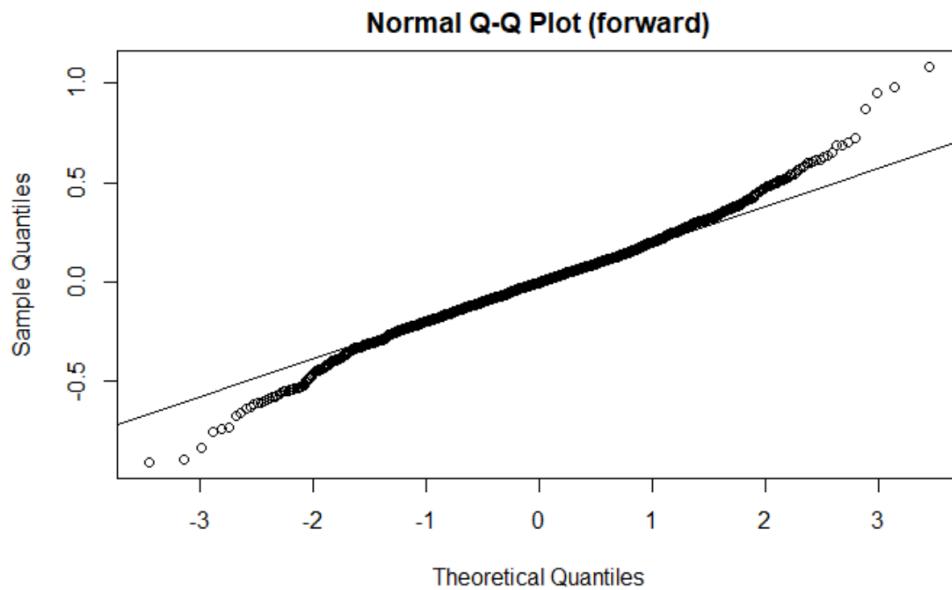


FIGURA F.2: Gráfico *QQ plot* para el modelo geográficamente ponderado.

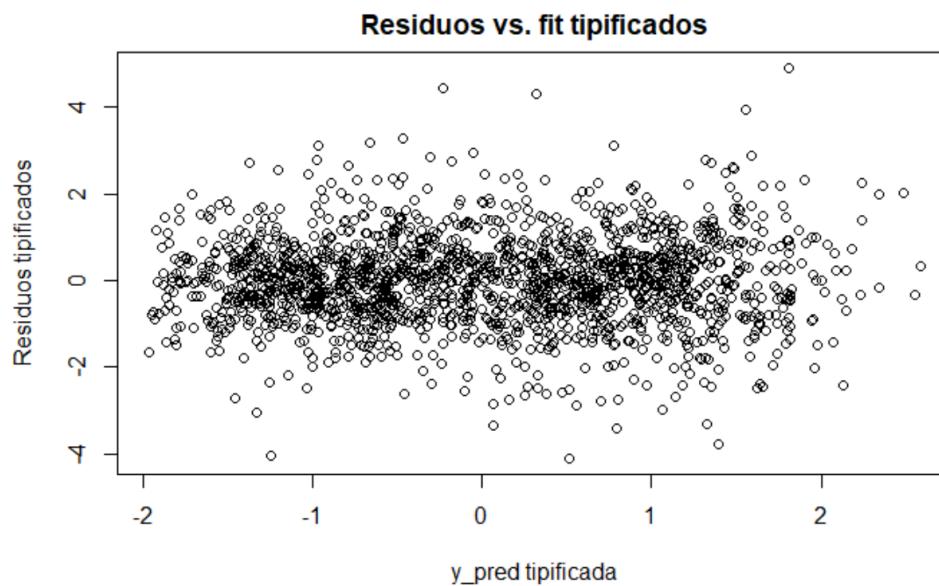


FIGURA F.3: Gráfico de residuos para el modelo geográficamente ponderado.

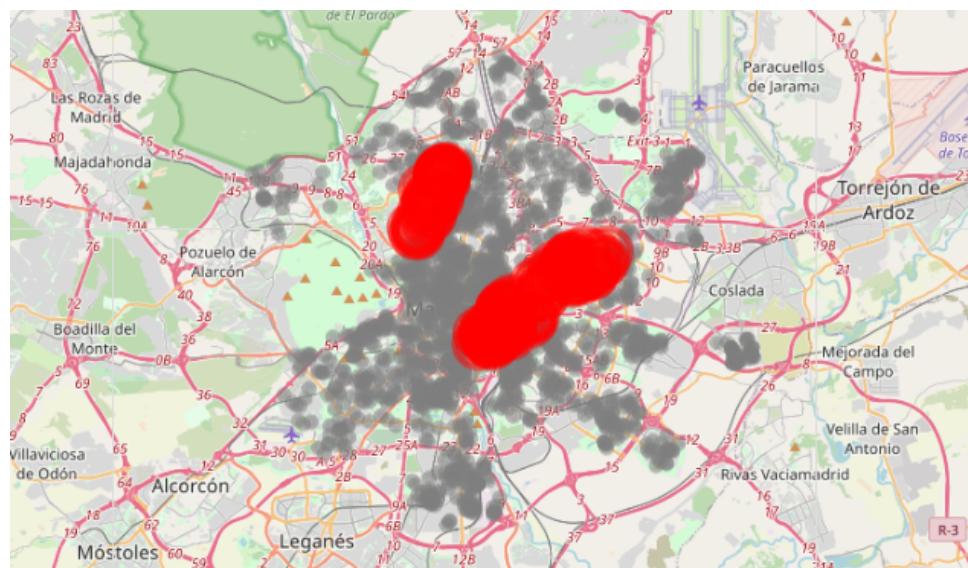


FIGURA F.4: Mapa *SatScan* para el modelo geográficamente ponderado. *Clusters* con $p\text{-value} < 0.1$

Apéndice G

GB

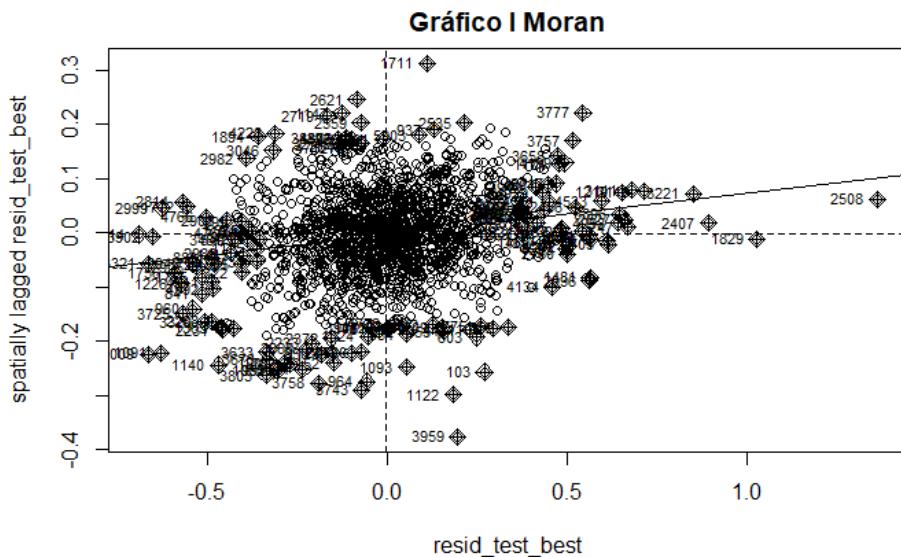


FIGURA G.1: Gráfico *I de Moran* para el modelo *Gradient Boost*.

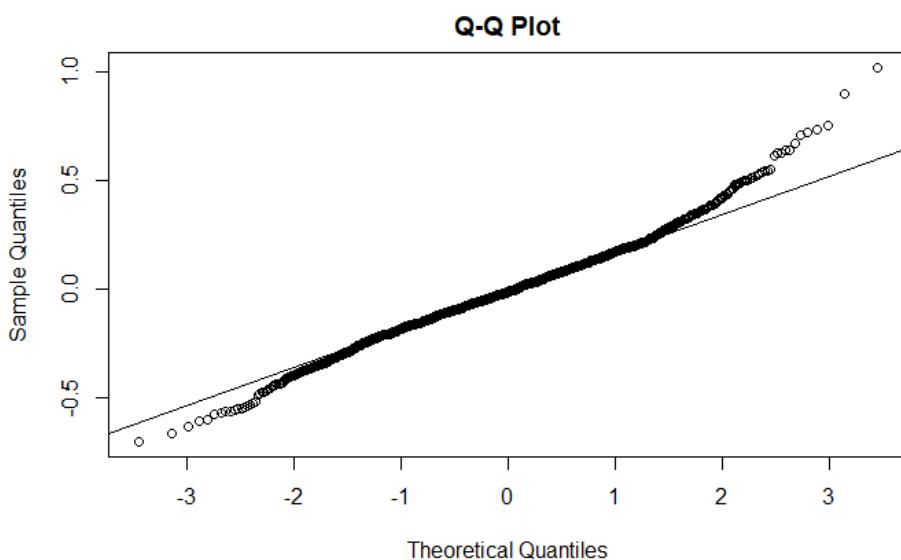


FIGURA G.2: Gráfico *QQ plot* para el modelo *Gradient Boost*.

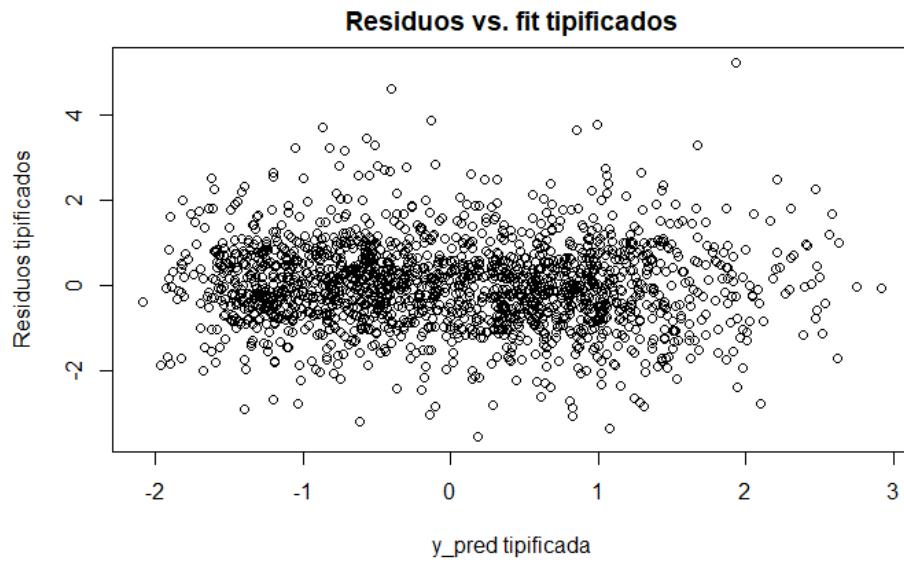


FIGURA G.3: Gráfico de residuos para el modelo *Gradient Boost*.

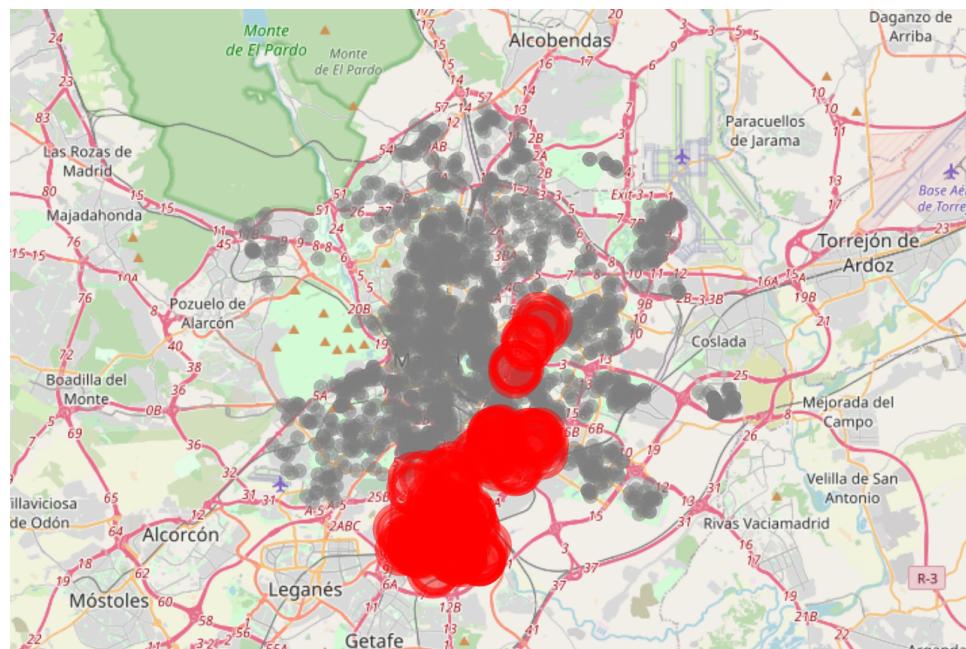


FIGURA G.4: Mapa *SatScan* para el modelo *Gradient Boost*.
Cluster con p-value <0.1.

Bibliografía

- [1] Jean HP Paelinck y col. *Spatial econometrics*. Vol. 1. Saxon House, 1979.
- [2] L. Anselin. *Spatial Econometrics: Methods and Models*. 1988. DOI: <http://doi.org/10.1007/978-94-015-7799-1>.
- [3] Luc Anselin. "Thirty years of spatial econometrics". En: *Papers in Regional Science* 89.1 (2010), págs. 3-25. DOI: <https://doi.org/10.1111/j.1435-5957.2010.00279.x>. eprint: [https://rsaiconnect.onlinelibrary.wiley.com/doi/abs/10.1111/j.1435-5957.2010.00279.x](https://rsaiconnect.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1435-5957.2010.00279.x). URL: <https://rsaiconnect.onlinelibrary.wiley.com/doi/abs/10.1111/j.1435-5957.2010.00279.x>.
- [4] A.D. Cliff y J.K. Ord. *Spatial Processes: Models & Applications*. Pion, 1981. ISBN: 9780850860818. URL: <https://books.google.es/books?id=Mi0OAAAAQAAJ>.
- [5] L. Anselin. "Model Validation in Spatial Econometrics: A Review and Evaluation of Alternative Approaches". En: *International Regional Science Review* 11.3 (1988), págs. 279-316. DOI: <10.1177/016001768801100307>. eprint: <https://doi.org/10.1177/016001768801100307>. URL: <https://doi.org/10.1177/016001768801100307>.
- [6] Paul Krugman. "Increasing Returns and Economic Geography". En: *Journal of Political Economy* 99.3 (1991), págs. 483-499. DOI: <10.1086/261763>. eprint: <https://doi.org/10.1086/261763>. URL: <https://doi.org/10.1086/261763>.
- [7] Robin A. Dubin. "Spatial Autocorrelation: A Primer". En: *Journal of Housing Economics* 7.4 (1998), págs. 304-327. ISSN: 1051-1377. DOI: <https://doi.org/10.1006/jhec.1998.0236>. URL: <https://www.sciencedirect.com/science/article/pii/S105137798902364>.
- [8] Kelley Pace, Ronald Barry y C F Sirmans. "Spatial Statistics and Real Estate". En: *The Journal of Real Estate Finance and Economics* 17.1 (1998), págs. 5-13. URL: <https://EconPapers.repec.org/RePEc:kap:jrefec:v:17:y:1998:i:1:p:5-13>.
- [9] Sabyasachi Basu y Thomas G. Thibodeau. "Analysis of Spatial Autocorrelation in House Prices". En: *The Journal of Real Estate Finance and Economics* 17.1 (1998), págs. 61-85. DOI: <https://doi.org/10.1023/A:1007703229507>.
- [10] José Luis Campos Echeverría. *La burbuja inmobiliaria española*. Abr. de 2008.
- [11] Gonzalo Bernardos Domínguez. "Creación y destrucción de la burbuja inmobiliaria en España". En: *Información Comercial Española, ICE* 850 (2009), págs. 23-40.
- [12] José Montero. "El precio medio del metro cuadrado de la vivienda libre: Una aproximación metodológica desde la perspectiva de la

- Geoestadística." En: *Estudios de Economía Aplicada* 22 (ene. de 2004), págs. 1-18.
- [13] José Montero y Gema Fernández-Avilés. "La importancia de los efectos espaciales en la predicción del precio de la vivienda. Una aplicación geoestadística en España". En: *Papeles de Economía Española ISSN 0210-9107 Papeles de Economía Española* (ene. de 2017), págs. 104-124.
- [14] María del Carmen Morillo Balsara, Francisco García Cepeda y Sandra Martínez Cuevas. "The application of spatial analysis to cadastral zoning of urban areas: an example in the city of Madrid." En: *Survey Review* 49.353 (2017), págs. 83-92. URL: <http://oa.upm.es/40201/>.
- [15] R. Freeman y Dl Reed. "Stockholders and Stakeholders: A New Perspective on Corporate Governance". En: *California Management Review* 25 (abr. de 1983). DOI: [10.2307/41165018](https://doi.org/10.2307/41165018).
- [16] A. Humphrey. "SWOT Analysis for Management Consulting". En: *SRI Alumni Newsletter* (2005).
- [17] Rezaeian M. "Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary". En: *Journal of epidemiology and community health* 61 (2007), págs. 98-10. DOI: [10.1136/jech.2005.043117](https://doi.org/10.1136/jech.2005.043117).
- [18] CHRISTIAN RODRIGUEZ FUENTES y col. "La proximidad geográfica en el contagio del fracaso empresarial en la pyme: Una aplicación empírica con el modelo probit espacial". Español. En: *Estudios de Economía Aplicada* (2016). ISSN: 1133-3197. URL: <https://www.redalyc.org/articulo.oa?id=30147485007>.
- [19] J. Nellis y J. Longbottom. "An Empirical Analysis of the Determination of House Prices in the United Kingdom". En: *Urban Studies* 18 (1981), págs. 21 -9.
- [20] Denise DiPasquale y William C. Wheaton. "Housing Market Dynamics and the Future of Housing Prices". En: *Journal of Urban Economics* 35.1 (1994), págs. 1-27. ISSN: 0094-1190. DOI: <https://doi.org/10.1006/juec.1994.1001>. URL: <https://www.sciencedirect.com/science/article/pii/S0094119084710011>.
- [21] Jesse M. Abraham y Patric H. Hendershott. "Bubbles in Metropolitan Housing Markets". En: *Journal of Housing Research* 7.2 (1996), págs. 191-207. ISSN: 10527001. URL: [http://www.jstor.org/stable/24832859](https://www.jstor.org/stable/24832859).
- [22] Stephen Malpezzi. "A Simple Error Correction Model of House Prices". En: *Journal of Housing Economics* 8.1 (1999), págs. 27-62. ISSN: 1051-1377. DOI: <https://doi.org/10.1006/jhec.1999.0240>. URL: <https://www.sciencedirect.com/science/article/pii/S1051137799902401>.
- [23] G. Jud y Dan Winkler. "The Dynamics of Metropolitan Housing Prices". En: *Journal of Real Estate Research* 23 (feb. de 2002), págs. 29-46.
- [24] Paloma Taltavull de La Paz. "Determinants of housing prices in Spanish cities". En: *Journal of Property Investment & Finance* 21 (abr. de 2003), págs. 109-135. DOI: [10.1108/14635780310469102](https://doi.org/10.1108/14635780310469102).
- [25] Leslie Rosenthal. "House prices and local taxes in the UK". En: *Fiscal Studies* 20.1 (mar. de 1999), págs. 61-76.
- [26] Coro Chasco. "Geografía y precio de la vivienda en los municipios urbanos de España". En: *CLM Economía-Revista Económica de Castilla-La Mancha* 12 (mayo de 2008), págs. 243-272.

- [27] Paul E. Bidanset y John R. Lombard. "Evaluating Spatial Model Accuracy in Mass Real Estate Appraisal: A Comparison of Geographically Weighted Regression and the Spatial Lag Model". En: *Cityscape* 16.3 (2014), págs. 169-182. ISSN: 1936007X. URL: <http://www.jstor.org/stable/26326913>.
- [28] Mateusz Tomal. "Modelling Housing Rents Using Spatial Autoregressive Geographically Weighted Regression: A Case Study in Cracow, Poland". En: *ISPRS International Journal of Geo-Information* 9.6 (2020). ISSN: 2220-9964. URL: <https://www.mdpi.com/2220-9964/9/6/346>.
- [29] Marco Helbich y col. "Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria". En: *Urban Studies* 51.2 (2014), págs. 390 -411. DOI: [10.1177/0042098013492234](https://doi.org/10.1177/0042098013492234). eprint: <https://doi.org/10.1177/0042098013492234>. URL: <https://doi.org/10.1177/0042098013492234>.
- [30] P. Whittle. "On Stationary Processes in the Plane". En: *Biometrika* 41.3 - 4 (1954), págs. 434-449. ISSN: 00063444. URL: <http://www.jstor.org/stable/2332724>.
- [31] Chris Brunsdon, A. Stewart Fotheringham y Martin E. Charlton. "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity". En: *Geographical Analysis* 28.4 (1996), págs. 281-298. DOI: <https://doi.org/10.1111/j.1538-4632.1996.tb00936.x>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.1996.tb00936.x>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1996.tb00936.x>.
- [32] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine." En: *The Annals of Statistics* 29.5 (2001), págs. 1189 -1232. DOI: [10.1214/aos/1013203451](https://doi.org/10.1214/aos/1013203451). URL: <https://doi.org/10.1214/aos/1013203451>.
- [33] Jerome Friedman. "Stochastic Gradient Boosting". En: *Computational Statistics & Data Analysis* 38 (feb. de 2002), págs. 367-378. DOI: [10.1016/S0167-9473\(01\)00065-2](https://doi.org/10.1016/S0167-9473(01)00065-2).
- [34] S.A. Glantz y B.K. Slinker. *Primer of Applied Regression and Analysis of Variance*. McGraw-Hill, Health Professions Division, 1990. ISBN: 9780070234079. URL: <https://books.google.es/books?id=SHH3GwAACAAJ>.
- [35] P. A. P. Moran. "Notes on Continuous Stochastic Phenomena". En: *Biometrika* 37.1/2 (1950), págs. 17-23. ISSN: 00063444. URL: <http://www.jstor.org/stable/2332142>.
- [36] Carlos M. Jarque y Anil K. Bera. "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". En: *Economics Letters* 6.3 (1980), págs. 255-259. ISSN: 0165-1765. DOI: [https://doi.org/10.1016/0165-1765\(80\)90024-5](https://doi.org/10.1016/0165-1765(80)90024-5). URL: <https://www.sciencedirect.com/science/article/pii/0165176580900245>.
- [37] Joseph I. Naus. "Approximations for Distributions of Scan Statistics". En: *Journal of the American Statistical Association* 77.377 (1982), págs. 177 -183. DOI: [10.1080/01621459.1982.10477783](https://doi.org/10.1080/01621459.1982.10477783). eprint: <https://www.tandfonline.com/doi/pdf/10.1080/01621459>.

1982.10477783. URL: <https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477783>.