



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas

Técnicas de inteligencia de datos para cuantificar la
rentabilidad de locales comerciales en la ciudad de Madrid

Trabajo Fin de Grado

Grado en Administración y Dirección de Empresas

AUTOR/A: Silvera Vies, Romina Andrea

Tutor/a: Debón Aucejo, Ana María

Cotutor/a: Ribal Sanchis, Francisco Javier

CURSO ACADÉMICO: 2021/2022

Índice general

Introducción	5
Motivación	5
Objetivos del TFG	6
Contexto	6
Definición de rentabilidad en inversiones inmobiliarias	7
Metodología	12
Obtención de la información	12
Lenguaje de programación R	12
Limpieza de bases de datos	13
Técnicas de estudio de la dependencia espacial	13
Índice global de Moran	18
Índice local de Moran	19
Descripción de los data sets	20
Análisis Exploratorio de Datos	24
Índices espaciales	48
Índice global de Moran	48
Índice local de Moran	53
Rentabilidades	58
Conclusiones	64

Anexo **66**

Referencias **68**

Índice de figuras

1	Ejemplo de la autocorrelación espacial y la importancia de la escala	16
2	Coordenadas locales comerciales en venta en la CCAA de Madrid	21
3	Coordenadas locales comerciales en alquiler en la CCAA de Madrid	22
4	Geometría de la ciudad de Madrid por distritos	22
5	Inmuebles comerciales en venta mapeados en la CCAA de Madrid	25
6	Inmuebles comerciales en alquiler mapeados en la CCAA de Madrid	26
7	Superficie de locales comerciales por mercado (en metros cuadrados)	29
8	Intensidad de oferta de venta por distritos	33
9	Intensidad de oferta de alquiler por distrito	34
10	Precio (en euros) por metro cuadrado. Base de venta	37
11	Precio (en euros) por metro cuadrado. Base de alquiler	38
12	Segmentación de locales comerciales por tipología	39
13	Cuota de locales pequeños en alquiler por distritos	40
14	Cuota de locales pequeños en venta por distritos	41
15	Porcentaje de locales medianos en venta por distritos	42
16	Porcentaje de locales medianos en alquiler por distritos	43
17	Porcentaje de locales grandes en venta por distritos	44
18	Porcentaje de locales grandes en alquiler por distritos	45
19	Porcentaje de locales industriales en venta por distritos	46
20	Porcentaje de locales industriales en alquiler por distritos	47
21	Órdenes para el calculo de los vecindarios	48
22	Vecindario utilizando el criterio de la reina	49

23	Simulación de Montecarlo para los datos de venta	52
24	Simulación de Montecarlo para los datos de alquiler	53
25	Mapa de significatividad (LISA), en relación con los precios de venta de inmuebles comerciales en la ciudad de Madrid	55
26	Mapa de significatividad (LISA), en relación con los precios de alquiler de inmuebles comerciales en la ciudad de Madrid	56
27	Mapa de clusters (LISA), en relación con los precios de venta de inmuebles comerciales en la ciudad de Madrid	57
28	Mapa de clusters (LISA), en relación con los precios de alquiler de inmuebles comerciales en la ciudad de Madrid	58

Índice de cuadros

1	Ejemplo flujos de caja de una inversión	8
2	Rentabilidad de la inversión inmobiliaria. 2T 2020 (Idealista)	10
3	Estadísticos resumen de variable precio (datos de venta)	27
4	Estadísticos resumen de variable precio (datos de alquiler)	27
5	Sumario de variable superficie en m^2 (datos de venta)	28
6	Sumario de variable superficie en m^2 (datos de alquiler)	28
7	Variable subtype	30
8	Valores faltantes por variable. Fichero de venta	30
9	Valores faltantes por variable. Fichero de alquiler	31
10	Oferta por mercado	31
11	Precios medios de venta por distritos (en euros)	35
12	Precios medios de alquiler por distritos (en euros)	36
13	Rentabilidad por metro cuadrado	59
14	Rentabilidad por metro cuadrado: Zona central	60
15	Rentabilidad por metro cuadrado: Zona media	61
16	Rentabilidad por metro cuadrado: Periferia este	62
17	Rentabilidad por metro cuadrado: Periferia oeste	63

Introducción

Anteriormente, la incursión en el mundo de programación de software recaía en un entorno inaccesible debido a la falta de experiencia y al ser ésta un área que, parecía alejada del terreno de ADE. A pesar de ello, el oportuno anexo de las asignaturas de Inteligencia de Negocios en mi último año como estudiante, no sólo aportaron bases fundamentales para el empleo de lenguajes de programación y manipulación de bases de datos, sino también la convicción de que, independientemente de las elecciones académicas, las cosas no se hacen más comprensibles a medida que las investigamos, sino más misteriosas, eso precisamente, es lo que promueve la realización de éste TFG, con el que se pretende compartir diversos resultados del análisis estadístico de la mano de uno de los principales lenguajes de programación en la actualidad.

Motivación

A diario nos exponemos a una creciente complejidad, producto de los veloces cambios en las estructuras y procesos, provocando que el aprendizaje rápido y continuo se convirtiera en un factor muy valioso dentro del ámbito laboral. La perspectiva analítica en el área de *business analytics*, a diario gana más crédito como bien estratégico en las empresas, responsable a su vez de responder al: ¿qué se está midiendo? y ¿se está midiendo lo que realmente importa?

La creación de valor puede tomar diversas vías: reducción de costes, revelación de oportunidades, mejor emplazamiento de recursos, independientemente de lo que se quiera lograr. Uno parte de grandes volúmenes de información, normalmente desorganizados y desestructurados (Laursen y Thorlund2010). En esta oportunidad, nos favorecemos de los lenguajes de programación y amplias posibilidades de diseño ofrecidas para un apto tratamiento de datos y posteriormente, hacer ofrecimiento de un informe que exprese los resultados obtenidos, de gran utilidad a la hora de la toma de decisiones económico-financieras, estrechamente ligadas a los indicadores clave de rendimiento (KPI).

Objetivos del TFG

- Proporcionar la descripción exploratoria del mercado inmobiliario de espacios comerciales en la ciudad de Madrid, tanto para propiedades en alquiler y venta.
- Aplicar diversas técnicas para la visualización, análisis e interpretación de la información.
- Determinar numéricamente la dependencia espacial a escala global y local mediante el índice de Moran.
- Calcular la rentabilidad bruta por metro cuadrado de los locales de cada uno de los 21 distritos que constituyen la ciudad de Madrid, así como para cada tipología en función de su extensión en metros cuadrados.

Contexto

Las técnicas que serán descritas en el apartado de metodología, se emplean sobre datos pertenecientes al 2020, periodo caracterizado por el cierre obligatorio del comercio no esencial y las limitaciones posteriores en términos de horarios y aforo.

De acuerdo con el [Instituto Nacional de Estadística](#), siendo responsable del 19,3% del PIB regional para 2020, el comercio de la Comunidad de Madrid mantuvo su representación como uno de los núcleos más productivos.

Ya sea en la capital o en los municipios, la oferta comercial es variable. Según el informe Comercio y Consumo (2022) disponible en la [web de la Comunidad de Madrid](#), los establecimientos no especializados resultan ser los más cuantiosos con un 34%, seguidos de los comercios de alimentos, bebidas y tabaco con el 20% y en tercer lugar se encuentran los comercios de artículos de uso doméstico ocupando un 18%.

El mismo informe indica que dentro del comercio especializado, rige el sector de la alimentación, destacando el papel de mercados y galerías comerciales de alimentación. También, el sector textil se distingue dentro de la estructura comercial por los ingresos generados en 2020, unos 1.967 millones de euros. Por otra parte, la Comunidad Autónoma de Madrid es líder en la acogida de franquicias (26%), de las cuales se reportó una facturación de 10.832,9 millones de euros.

En el mercado retail¹ en Madrid, el Barrio de Salamanca y la zona Centro son los ejes comerciales prime² de la ciudad, en cuanto al sector, la moda dominaba el mercado con el 37% de la cuota, seguida de la restauración con el 17,6%.

Definición de rentabilidad en inversiones inmobiliarias

Dentro de la Resolución del 1 de marzo del 2013, del [Instituto de Contabilidad y Auditoría de Cuentas](#), referente a las normas de registro y valoración del inmovilizado material, se provee la definición de las inversiones inmobiliarias, la cual se refleja a continuación:

«Son inversiones inmobiliarias las definidas en el Plan General de Contabilidad como activos no corrientes que sean inmuebles y que se posean para obtener rentas, plusvalías o ambas, en lugar de para: a) Su uso en la producción o suministro de bienes o servicios distintos del alquiler, o bien para fines administrativos; o b) Su venta en el curso ordinario de las operaciones de la empresa.»

Tradicionalmente, los ratios de rentabilidad se utilizan con la finalidad de analizar la capacidad de generación de resultado, dichas herramientas son las únicas que reducen a tamaño común la dimensión de entes, lo cual facilita la comparación entre estos.

Hay una amplia variedad de supuestos de valoración de inversiones, que a su vez pueden clasificarse en financieras y productivas (Brotos Martínez²⁰¹⁷), la compra de bonos o acciones son ejemplos de inversiones financieras, mientras que la compra de un local sirve de ejemplo de inversión productiva. Sea cual sea el tipo, una inversión supone la recuperación de un capital en un plazo fijo, incluso junto con una compensación (interés).

Toda inversión se asocia con un flujo de cobros recibidos y pagos efectuados, casi siempre, en tiempos diferentes. En definitiva, el inversor se decanta por el proyecto que presente cobros superiores a los pagos y, en caso necesario, recurre al descuento financiero de esos flujos de caja.

A continuación, definimos los elementos fundamentales de una inversión:

- Desembolso inicial D_0 : Es el paso inicial para adquirir activos fijos (maqui-

¹Relativo a la venta al por menor o comercio minorista.

²Predominante o principal.

naria, instalaciones, terrenos, etc.)

- Cobros C_i : Ingresos que supongan entrada de caja.
- Pagos P_i : Sólo gastos que supongan salida efectiva de caja, no incluye las amortizaciones de inmovilizado.
- Flujos netos de caja o "cash flow": Diferencia entre cobros y pagos.

Ejemplo: Se adquiere un local que supone una inversión inicial de 20.000 € y unos gastos de acondicionamiento de 1.000 €, 1.100 €, 1.200 € los primeros tres años y 1.400 € para los cinco años siguientes. Se espera percibir ingresos a partir del tercer año a razón de 1.000€ y que irán incrementando en un 10% hasta el octavo año.

La información anterior nos permite plasmar los flujos de caja en el Cuadro 1.

Año	Cobros	Pagos	Cash flow
0	0	-20.000	-20.000
1	0	-1.000	-1.000
2	0	-1.100	-1.100
3	1.000	-1.200	(1.000-1.200)= -200
4	1.100	-1.300	(1.100-1.300)= -200
5	1.210	-1.300	(1.210-1.300)= -90
6	1.331	-1.300	(1.331-1.300)= 31
7	1.464	-1.300	(1.464-1.300)= 164
8	1.610	-1.300	(1.610-1.300)= 310

Cuadro 1: Ejemplo flujos de caja de una inversión

La valoración de proyectos de inversión se pueden dividir en estáticos y dinámicos, los primeros no consideran el valor temporal del dinero, tomando así el supuesto de que el valor de 1 euro es el mismo hoy que en veinte años, análogamente, los dinámicos solucionan el problema del valor temporal del dinero mediante el descuento de los flujos de caja a una tasa adecuada (Brotons Martínez2017).

La Cámara de Comercio de Madrid (2021) contrasta los conceptos de rentabilidad financiera y **rentabilidad económica**, refiriéndose a la última como una relación entre los activos netos y el beneficio obtenido, sin haber deducido intereses e impuestos. Por lo tanto, determina la capacidad de generar beneficios del activo neto obviando la forma en la que se hayan financiado. Desde otra perspectiva, la rentabilidad financiera relaciona el beneficio neto y el patrimonio neto

o recursos propios, permitiendo así calcular la capacidad de los fondos propios para obtener beneficios.

Como se explicó antes, para calcular la **rentabilidad económica** es preciso conocer tanto el dato de los beneficios como el de la inversión realizada, que se corresponde con los activos netos. Así, planteamos el cálculo dividiendo la inversión realizada entre el beneficio obtenido como muestra la ecuación (1):

$$RE = \frac{\textit{Beneficio neto}}{\textit{Valor inicial inversión}} \cdot 100 \quad (1)$$

Un proyecto valorado en base a esta metodología, será aceptado cuando la tasa de rentabilidad sea superior a una tasa mínima fijada de antemano.

Aunque sean dos mercados diferentes el de venta y alquiler de inmuebles, estos están estrechamente relacionados y se utilizan en conjunto a la hora de cuantificar las rentabilidades medias de diversos productos en diferentes territorios. Con el supuesto de que un inversor adquiere un local, lo que supone un desembolso inicial D_0 , para luego poner dicho espacio en alquiler, lo cual conllevará una recepción de cobros periódicos C_i , podemos obtener la rentabilidad anual o, en nuestro caso, la rentabilidad por metro cuadrado.

Siendo uno de los objetivos del TFG, se obtendrán los índices anuales de rentabilidad económica por metro cuadrado de cada distrito y tipología de inmueble (pequeños, medianos...) a fin de compararlos. Para realizar lo propuesto se sigue la ecuación (1) aplicada a los datos disponibles:

- Numerador: Se toma el beneficio de la inversión, es decir, la renta anual a recibir en concepto de alquiler, para ello, multiplicamos el importe de alquiler mensual por 12, con el objeto de tratar con términos anuales y posteriormente dividimos la cantidad por la superficie en m^2 .
- Denominador: Precio de venta, que representa el coste total de inversión y también se divide por la extensión en metros cuadrados.

Tras obtener los numeradores y denominadores de cada observación, podremos agruparlos utilizando el paquete «dplyr» (Wickham, François, Henry, y Müller2022), en función del distrito o tipología de inmueble para llevar a cabo la relación.

Una metodología similar fue empleada en el Informe de Rentabilidad de Idealista (Redacción Idealista2021), realizado con los datos del segundo trimestre del

2020, los locales comerciales fueron de los productos estrella, ofreciendo un retorno de 9,4%, colocándose así por encima de otros inmuebles como garajes e incluso viviendas, no obstante, el producto líder del periodo fueron las oficinas con 11,4%.

Tal como muestra el Cuadro 2, que recopila los datos del informe, en casi todas las capitales se presenciaban los mayores porcentajes de retorno para los locales comerciales. En todo caso, las rentabilidades más bajas fueron del 7% en Teruel y Albacete..

Cuadro 2: Rentabilidad de la inversión inmobiliaria. 2T 2020 (Idealista)

Capital de provincia	Vivienda	Oficinas	Locales	Garajes
A Coruña	4.5	5.7	7.4	5.1
Albacete	5.7	-	7	4.2
Alicante	6.0	7.2	9.7	5.1
Almería	6.9	7.6	9	5.6
Ávila	6.1	-	11.1	10.3
Badajoz	6.1	7.3	7.9	5
Barcelona	4.4	6.3	8.5	6.3
Bilbao	5.0	6.4	9.6	4.3
Burgos	5.7	8.3	9.2	5.5
Cáceres	5.8	-	8.6	4.8
Cádiz	4.9	-	7.5	4.9
Castellón de la Plana	6.5	7.7	7.9	8
Ciudad Real	5.9	-	8.7	4.7
Córdoba	6.3	6.9	8.3	4.2
Cuenca	6.6	-	7.5	-
Donostia-San Sebastián	3.6	7.2	9.1	3.8
Girona	5.1	-	9	5.6
Granada	5.2	-	8.5	3.5
Guadalajara	6.5	-	8.8	5.4
Huelva	7.5	11.1	11.2	6.7
Huesca	6.1	-	9.6	-
Jaén	6.3	-	7.2	3.3
Las Palmas de Gran Canaria	5.9	6.2	8.5	5.3
León	5.9	7.4	8	5.5
Lleida	8.2	6.8	8.9	7.1
Logroño	5.5	7.5	8.1	4.6
Lugo	5.5	-	10.6	3.3
Madrid	4.6	6.8	7.8	4.8

Cuadro 2: Rentabilidad de la inversión inmobiliaria. 2T
2020 (Idealista) (continued)

Capital de provincia	Vivienda	Oficinas	Locales	Garajes
Málaga	5.3	6.8	8.9	4.9
Murcia	8.1	6.5	9.4	9.6
Ourense	4.9	6.8	8.3	3.1
Oviedo	6.0	6.9	10.8	5.6
Palencia	5.7	-	7.5	3.4
Palma de Mallorca	4.3	6.4	8.1	5.7
Pamplona/Iruña	4.9	7.2	9	6.5
Pontevedra	4.9	-	7.7	-
Salamanca	5.6	-	7.7	2.4
Santa Cruz de Tenerife	6.8	6.6	9.2	5.1
Santander	5.4	6.7	9.3	3.3
Segovia	6.4	-	8.2	3.8
Sevilla	5.9	14.1	9.7	6
Soria	6.3	-	-	-
Tarragona	5.9	-	9.9	4.9
Teruel	6.4	-	7	-
Toledo	6.4	10	8.4	7.4
València	5.8	6.1	8.3	4.7
Valladolid	5.7	8.2	8.7	5.5
Vitoria-Gasteiz	5.0	8.3	8.5	4
Zamora	5.7	-	7.8	4.8
Zaragoza	6.2	8.7	9.8	5.1

Metodología

El trabajo presentado tiene un carácter tanto cualitativo como cuantitativo. Por una parte, se ha inquirido en diversas fuentes referentes a la información espacial y los sistemas de información geográfica, así como en la estructuración de los índices que se emplean para determinar la correlación entre territorios.

La metodología cuantitativa se ha basado en la utilización del lenguaje de programación R, ciertas de sus librerías puntuales para el tratamiento de información espacial y diversas técnicas de visualización de datos.

Obtención de la información

El contenido de los espacios comerciales, tanto en venta como alquiler, fue obtenido en [Fotocasa](#), portal inmobiliario fundado en 1999 y de carácter líder en el ámbito nacional. La información fue descargada de la página web mediante el proceso denominado «*web scrapping*», perteneciente a la ciencia de datos, que permite la transformación y almacenamiento de datos para la realización de minería y otros diversos propósitos.

Cabe mencionar que la información data del año 2020.

Lenguaje de programación R

R (Team2022) es un lenguaje y entorno computacional de carácter estadístico y gráfico. Se define como un proyecto de GNU, sistema operativo de software libre. R pone a disposición del usuario un extensivo rango de técnicas estadísticas y gráficas, dotando de comodidad los procesos de manipulación, cálculo y exhibición de la información.

En el [sitio web oficial](#) de R, se listan las cualidades del lenguaje:

- Una instalación efectiva de manejo y almacenamiento de datos.
- Un conjunto de operadores para cálculos en arreglos, en particular matrices.
- Una colección grande, coherente e integrada de herramientas intermedias para el análisis de datos.
- Facilidades gráficas para el análisis y visualización de datos, ya sea en pantalla o en papel.

- Un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario y facilidades de entrada y salida.

Conjuntamente, RStudio (Team2021) es un entorno de desarrollo integrado (IDE) para R, disponible en ediciones abiertas o comerciales. Incluye una consola, editor de sintaxis para la ejecución directa de código, herramientas para visualización de gráficos, historial, depuración y la gestión del espacio del trabajo.

Limpieza de bases de datos

Una base de datos «limpia» no tiene un formato hermético, sino que depende de los objetivos planteados por el analista. Wickham (2014) ofrece una perspectiva útil sobre lo anterior, expresa que una base de datos «ordenada» es rectangular (o tabular), donde cada fila describe una unidad de análisis (observación) y cada columna aporta una medida (variable). Por ejemplo, en el fichero correspondiente a locales comerciales en alquiler, cada fila debería corresponder a una propiedad de esa naturaleza, mientras que cada columna aportaría información sobre la misma: superficie en metros cuadrados, precio en euros, número de aseos, ubicación, entre otros.

En este caso, ambos ficheros, tras ser descargados por medio del *web scraping*, deben ser rediseñados para su correcto estudio, lo cual implica un aspecto vital, el grado de reproducibilidad de la información, para evitar desacuerdos sobre los datos, otro analista debería ser capaz de empezar con la misma información «bruta» y tras ejecutar todos los pasos aplicados, poder tratar exactamente con la misma información. El acto de hacer que la investigación sea reproducible ha sido, en los últimos años, reconocido como una piedra angular del progreso científico (Buttrey y Whitaker2017).

Técnicas de estudio de la dependencia espacial

Como preámbulo a los indicadores en los que basaremos las conclusiones del presente trabajo, se indaga sobre la autocorrelación espacial, no solamente como concepto, también las perspectivas de notables geólogos dentro de la ciencia de información geográfica. Tras la asimilación de algunas ideas clave, se espera plasmar la repercusión que tuvieron las aplicaciones estadísticas en geografía a lo largo de 40 años de desarrollo y mejoras.

La estadística en los contextos geográficos

El geólogo Goodchild (2009) refleja vigorosamente en su artículo *What Problem? Spatial Autocorrelation and Geographic Information Science* el hecho que el conocimiento nomotético es más valorado por la ciencia que el conocimiento ideográfico, alegando que, un psicólogo que busca patrones y efectos en las operaciones del cerebro humano, hará contribuciones más estimadas cuando se demuestre que dichos efectos puedan aplicarse a toda la población humana. No obstante, sería imposible para un investigador demostrar generalidad con una confianza plena, eso conllevaría al estudio de todos los cerebros (incluyendo cerebros pasados y futuros), alternativamente, se apela a la utilización de muestras para generalizar los resultados a una población mayor.

El aparato de la estadística inferencial se desarrolló para dar base lógica y matemática a los procesos. Una muestra debe ser representativa de su población «padre», de lo contrario, la posibilidad de sesgo existe en los resultados, y para una muestra ser representativa es necesario que todos los miembros de la población «padre» tenga la misma probabilidad de ser elegidos. El paradigma presentado por Neyman (1934) argumenta que la presencia de autocorrelación espacial hace que los valores muestreados no sean independientes, haciendo que se cuente con menos información y que los grados de libertad sean muy elevados. Siguiendo la postura anterior, mucha literatura participa sobre la aplicación de métodos de inferencia donde el concepto de población no es articulado o cuando las muestras no son confeccionadas aleatoriamente e independientemente de una población.

Estos problemas son comunes en los denominados experimentos «naturales», análisis de datos que surgen en el mundo real sin ninguna selección o control experimental por parte del investigador, como la situación planteada por el artículo de Cliff, Ord, y Scott (1969) sobre el estudio de condados de EE. UU, siendo más de 3000, es imposible argumentar de manera convincente que dos condados se han dibujado independientemente de una población más grande, especialmente cuando los condados son adyacentes.

El ascenso de los sistemas de información geográfica (GIS) a principios de los 60's, trajo nuevas perspectivas en lo que respecta a la autocorrelación espacial. El ambiente digital y los sistemas de código trajeron consigo más nivel de enfoque y rigor. Durante la etapa inicial de los GIS, se utilizaba una representación uniforme que fragmentaba áreas geográficas en zonas irregulares de tipo aproximadamente homogéneo, el uso de la tierra sirvió de base para la fragmentación. Estos mapas, llamados *area-class maps* (Mark y Csillag 1989) consistían en mapeos desde la ubicación x hasta la clase $c(x)$, y se basan en la tendencia de las clases asignadas a puntos cercanos a estar positivamente correlacionados, permiti-

tiendo la formación de zonas con homogeneidad aproximada. La geo-estadística o la teoría de las variables regionalizadas (Goovaerts1997), aportó un marco formal para el tratamiento de estos datos, la autocorrelación ahora era vista como una función de distancia monótonamente decreciente, ya no se consideraba un problema sino una característica que serviría para la investigación de una gran gama de fenómenos espacialmente distribuidos.

La autocorrelación espacial

La autocorrelación espacial es un concepto estadístico que expresa, mediante un número, la relación entre valores de objetos, dependiendo de vínculos métricos o topológicos entre ellos. Pretende demostrar que las disposiciones de los datos no se deben al azar. Por otra parte, de existir dicha dependencia, posibilita la estimación de un punto desconocido o la obtención de un modelo estadístico para explicar una variable a partir de otras. En este último caso, un componente vinculado a la autocorrelación es agregado cuando no existe una variable que ya explicase dicho fenómeno. La existencia de este concepto, se reflejó en la primera ley de la geografía de Tobler (1970):

«Todo está relacionado a todo lo demás, pero las cosas próximas están más relacionadas que cosas distantes.»

Esta técnica se estudia a través del análisis espacial, el cual considera dos clases de información, por un lado, están los atributos de rasgos espaciales, pudiendo ser medidas (precipitaciones, temperatura, población) y variables cualitativas como: religión, tipo de suelo, etc. Por otro lado, la localización de cada atributo puede ser descrita, como su posición en un mapa, a través de sistemas de coordenadas o cualquier referencia geográfica. La autocorrelación, se considera una de las, relativamente, pequeñas técnicas que se ocupa tanto de atributos como de la ubicación (Goodchild1986).

El grado o intensidad de autocorrelación espacial en un patrón depende estrechamente de la escala (Goodchild1986). Consideremos un tablero de ajedrez, donde hay 64 celdas iguales, siendo 32 de color blanco y 32 de color negro, estas se ordenan en un patrón perfectamente homogéneo. La Figura 1 presenta 4 tableros, el primer tablero representa una independencia espacial total, mientras que el número 2 sirve de ejemplo de autocorrelación espacial perfecta. En lo que respecta a los demás tableros, se detecta una agrupación de atributos, no tan vigorosa como en el caso anterior, pero se puede suponer la existencia de cierta dependencia espacial según las celdas del tercer tablero. De los cuatro casos, se puede coincidir en que la disposición en el cuarto tablero parece ser la más aleatoria de todas, lo que daría cabida a una baja dependencia espacial.

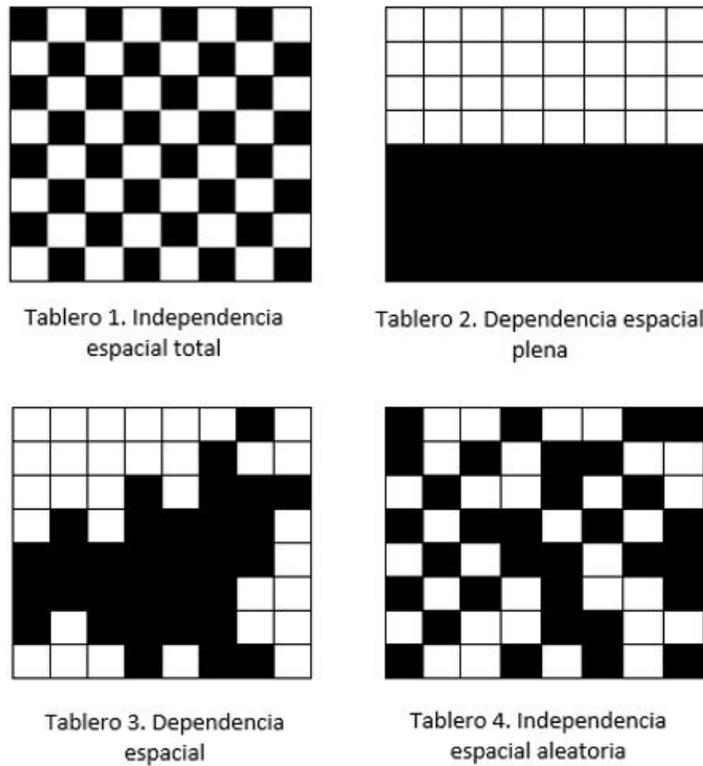


Figura 1: Ejemplo de la autocorrelación espacial y la importancia de la escala

Ahora, si subdividimos cada celda en cuatro o nueve partes iguales, cabe la posibilidad que los índices de dependencia varíen en términos numéricos, incluso de manera tan significativa hasta el punto de pasar inadvertido a simple vista. Aunque el patrón permanece constante, la escala ha sido modificada. Debemos ser conscientes que este tipo de medidas son específicas o concretas según la escala, es por ello que en este trabajo llevamos a cabo un análisis global y uno local para cada fichero de datos.

Souris (2019) expone en su obra, *Epidemiology and Geography*, los principales métodos para la conducción de análisis espaciales utilizados en el campo de epidemiología y la geografía médica. Los métodos directos, utilizados para analizar la distribución espacial de objetos o de sus valores directamente desde su ubicación, pueden listarse como:

- Análisis globales, que proveen información sobre la distribución espacial global.
- Análisis locales, que proveen información de variaciones locales de la distribución espacial.

- Detección de lugares en el espacio (puntos o superficies, como hot spots o clusters) que presentan un carácter específico respecto a la distribución espacial global.
- Estimación e interpolación, que permiten la caracterización de cualquier punto en el espacio respecto a la distribución espacial de valores conocidos de una función.

Siguiendo la naturaleza de otros tipos de análisis, el análisis espacial también aplica el principio del contraste de hipótesis. La cuestión a resolver se expresa mediante el enfrentamiento de dos hipótesis: una hipótesis nula (H_0) y una hipótesis alternativa (H_a). El principio se basa en la comparación de la situación que se observa en los datos bajo H_0 (suponiendo que H_0 es cierta). Se calcula un índice numérico que sintetiza el problema a resolver (test estadístico) que se utiliza para calcular el *p-valor*. Cuando los datos muestran evidencia clara en contra de los supuestos de la hipótesis nula, la magnitud del estadístico de prueba se vuelve demasiado grande o demasiado pequeña dependiendo de la hipótesis alternativa. Esto hace que el *p-valor* de la prueba se vuelva lo suficientemente pequeño como para rechazar la hipótesis nula.

La elección de la prueba de hipótesis a ejecutar, nos conducirá a un test estadístico u otro, pero en todo caso, la decisión sobre el rechazo o aceptación de la H_0 la dictamina el *p-valor*, cuyo propósito es cuantificar la significavidad estadística, es decir, la probabilidad de rechazar H_0 cuando es en realidad verdadera.

Como se expuso antes, el objetivo de los análisis globales es determinar características de la distribución espacial de eventos o atributos, para así profundizar en los procesos subyacentes que las provocan, las características referidas pueden ser: la extensión u orientación de una nube de puntos, la relación espacial entre objetos y sus valores (dependencia espacial) y los rasgos geométricos de las concentraciones de puntos.

Muchos índices de este tipo han sido desarrollados, la mayor parte de ellos derivan del índice de Mantel (1967) (ecuación (2)), que cuantifica las relaciones entre proximidad en el tiempo $|T_i - T_j|$, y proximidad en el espacio, medido por un número W_{ij} (basado en la distancia o proximidad), para todos los pares de puntos P_i, P_j . La significatividad estadística del índice se obtiene a través de una simulación de Monte Carlo.

$$I_{mantel} = \sum_{i,j,i>j}^n W_{i,j} * |T_i - T_j| \quad (2)$$

El aspecto que comparten la mayoría de índices es la utilización de un peso espacial, creado a partir de un parámetro métrico o topográfico, que una vez obtenido, expresa la dependencia espacial gracias a la ponderación de cada pareja de puntos.

Índice global de Moran

La herramienta I de Moran es una medida estadística desarrollada por Moran (1948), que analiza la autocorrelación espacial entre valores de vecinos más cercanos.

El índice se define como la media de los productos de los valores normalizados de los pares de puntos, ponderados por un «peso espacial» en función de la contigüidad o distancia entre ambos puntos del par:

$$I_{moran} = \frac{1}{S} \sum_{i,j,i>j} W_{i,j} \left(\frac{x_i - m}{\sigma} \right) * \left(\frac{x_j - m}{\sigma} \right),$$

donde m , es la media de x_i en D, σ es la desviación estándar de x_i en D, $W_{i,j}$ es el peso espacial del par de puntos (p_i, p_j) y S es la suma de los pesos espaciales.

En la definición inicial de Moran, el peso W_{ij} del par (p_i, p_j) es 1 si los dos objetos son adyacentes y 0 en caso contrario.

Proporciona resultados satisfactorios y tiende a elegirse sobre el estadístico de Geary, motivado a que el índice de Moran concuerda generalmente con las razones intuitivas. Sus valores oscilan entre -1 y +1, donde +1 significa autocorrelación perfecta. En definitiva, tenderá a ser > 0 cuando áreas cercanas tengan atributos familiares y < 0 en el caso contrario. En caso de ser igual a 0, se entiende como la independencia o ausencia absoluta de autocorrelación.

Sus resultados siempre son interpretados dentro del contexto de la hipótesis nula. Cuando el *p-valor* que devuelve esta herramienta es estadísticamente significativo, se puede rechazar la hipótesis nula, que establece que el atributo que se analiza está distribuido de forma aleatoria entre las entidades del área de estudio.

Índice local de Moran

Esta sección presenta los indicadores locales de asociación espacial (LISA), fruto de técnicas exploratorias desarrolladas a raíz de las capacidades de visualización de los sistemas de información geográfica. Ideales para averiguar la importancia de la variabilidad espacial de elementos locales, incluso si el fenómeno no presenta signos de autocorrelación global. Estos indicadores son muy prácticos a la hora de estudiar grandes dominios con marcadas diferencias regionales.

Los LISA miden la concentración de valores numéricos tomados de un conjunto de puntos (o calculados por agregación espacial). Siguen los mismos principios que los índices globales, pero el cálculo es efectuado para cada punto p_i usando los vecinos de ese punto, mientras que una simulación se hace para cada punto p_i , permutando los valores de los vecinos para aproximar la distribución bajo la hipótesis nula del índice y deducir el *p-valor*. Cuando el *p-valor* permite el rechazo de la hipótesis nula, existen varias posibilidades: p_i tiene un valor alto y está rodeado de valores altos («hot spot»), p_i tiene un valor bajo y los valores que lo rodean son bajos («cold spot») y el punto p_i está rodeado de valores opuestos así mismo («distinctive spot»).

El índice de Moran local (Anselin 1995) es una descomposición del índice de Moran global. Para cada punto p_i , se puede escribir lo siguiente:

$$I_{moran(p_i)} = \frac{x_i - \bar{X}}{S_i^2} \cdot \sum_{j=1, j \neq i}^n .W_{ij} \cdot (x_j - \bar{X}),$$

donde x_i es un atributo de i , \bar{X} es la media del atributo correspondiente y $w_{i,j}$ es el peso espacial de i y j .

Se utiliza para la caracterización de la autocorrelación local y nos permite diferenciar situaciones estadísticamente significativas por simulación, como en el caso general, aunque la simulación se refiere aquí sólo a los puntos implicados en el cálculo local.

Descripción de los data sets

Se cuenta con dos bases de datos:

- «a1» para el mercado de alquiler.
- «v1» para el mercado de venta.

Aunque ambas recolectan información sobre locales comerciales disponibles en la Comunidad de Madrid, nuestro análisis será sobre los inmuebles dentro de los límites de la ciudad de Madrid. La diferencia primordial de ambos ficheros es el tipo de oferta: venta o alquiler.

Una vez configuradas para que su visualización sea más sencilla, vemos que comparten las mismas 10 variables, listadas a continuación:

1. Id (id): Número identificador de cada registro en la base de datos.
2. Latitud (coordinates.latitude): Coordenada de latitud del inmueble.
3. Longitud (coordinates.longitude): Coordenada de longitud del inmueble.
4. Ubicación (location): Texto de la ubicación del local comercial.
5. Descripción (description): Texto descriptivo del local puesto en venta/alquiler.
6. Precio (rawPrice): Precio del local en cuestión, expresado en euros.
7. Subtypeld: Hasta el momento nos limitamos a decir que dicha variable es numérica, tomando valores de 0, 13 o 16. Por ahora, desconocemos el significado concreto de la misma.
8. Estado de conservación (conservationState): Variable de naturaleza numérica, tomando valores de 1 a 4. Presuponemos que, indica el grado de conservación del espacio ofertado.
9. Superficie (surface): Superficie del local en metros cuadrados.
10. N° de piso (floor): Número de piso/planta en la que se encuentra el local, en el caso de estar en un edificio.

El número de registros en cada fichero es distinto:

```
1 # dimension fichero de venta
2 dim(v1)
```

```
1 ## [1] 4087 10
```

```
1 # dimension fichero de alquiler
2 dim(a1)
```

```
1 ## [1] 2576 10
```

Se disponen de 4087 inmuebles puestos en venta y 2576 en alquiler en cada muestra, cifras que aportan un primer indicio sobre el mercado inmobiliario en la zona de estudio. Nótese que, por ahora, sólo contamos con los puntos de latitud y longitud. Con ésta información numérica podemos visualizar los puntos usando la librería «ggplot2» (Wickham2016) y como resultado obtenemos las Figuras 2 y 3, no obstante, no podemos determinar a qué distrito pertenecen a simple vista, tampoco saber cuántos locales hay por distrito, etc.

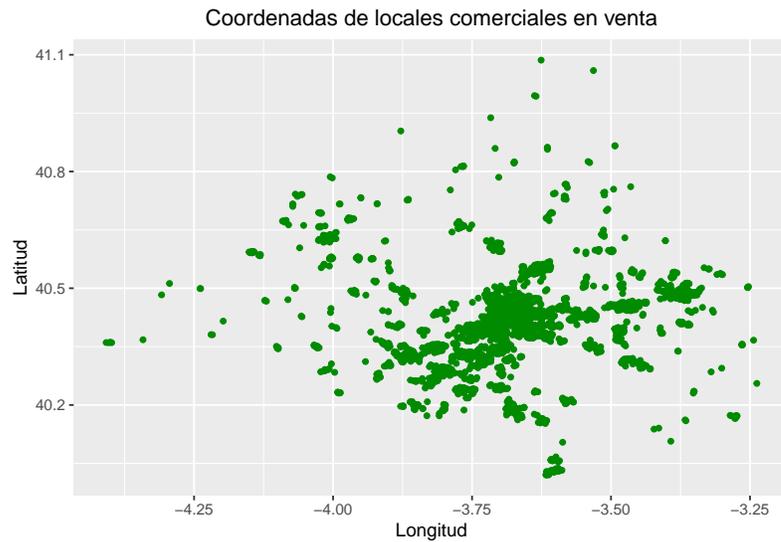


Figura 2: Coordenadas locales comerciales en venta en la CCAA de Madrid

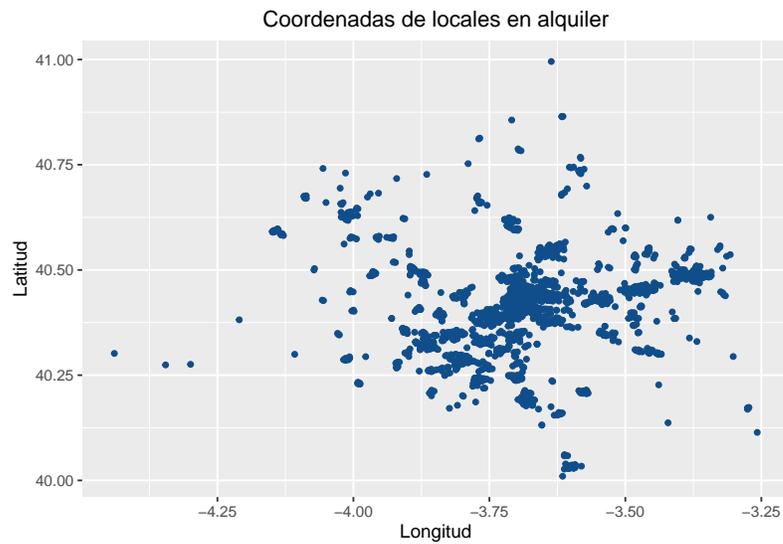


Figura 3: Coordenadas locales comerciales en alquiler en la CCAA de Madrid

Un paso importante para la ejecución del análisis exploratorio y posteriormente la determinación de dependencia geográfica, es la integración de un archivo tipo shapefile. Un shapefile (extensión *.shp) almacena atributos y características geográficas, consta de un número variable de información como pueden ser: zonales, lineales, puntuales, etc. Para la elaboración de este trabajo, hemos descargado un fichero shape desde el [Portal de datos abiertos de la Comunidad de Madrid](#), el mismo proporciona la división geográfica por distritos de la ciudad de Madrid. Por otra parte, una funcionalidad muy importante del archivo es brindarnos la geometría del área de estudio como muestra la Figura 4.



Figura 4: Geometría de la ciudad de Madrid por distritos

La unión de cada fichero de datos con el objeto shape, nos permitirá identificar fácilmente el distrito al que pertenece cada observación. El procedimiento anterior consta de transformar las bases de datos originales a bases de datos espaciales, para ello empleamos el la librería «sf» (Pebesma2018). Utilizando la función `st_as_sf`, en la cual se hace referencia a las columnas de longitud y latitud, con la especificación de un sistema de coordenadas que se aplicará a los elementos espaciales. Añadimos el sistema de referencia de coordenadas (crs) 4326 (WGS84- World Geodetic System 1984, usado en GPS).

Los crs proporcionan un método estandarizado para describir ubicaciones. Existen diferentes tipos y su utilización depende de la información disponible, su extensión geográfica, propósito, etc. La 4326 se usa cuando se disponen de puntos de longitud y latitud.

```
1 v1.sf<-st_as_sf(v1,coords=c("coordinates.longitude","coordinates.
  latitude"),crs=4326)
2 a1.sf<-st_as_sf(a1,coords=c("coordinates.longitude","coordinates.
  latitude"),crs=4326)
```

Tras la aplicación de dicha función, se comprueba que ambos objetos son de tipo espacial «sf» y data frame.

```
1 class(v1.sf)
1 ## [1] "sf" "data.frame"
1 class(a1.sf)
1 ## [1] "sf" "data.frame"
```

Luego, mediante `st_transform` se transfieren el resto de criterios geográficos, como el *datum* propio al objeto shape. Éste proporciona la información necesaria para anclar el resumen de coordenadas a la Tierra. El datum define un punto de origen de los ejes de coordenadas y define la dirección de los ejes. Piense en esto como la información base necesaria para «dibujar» coordenadas imaginarias en un globo terráqueo o mapa (Rummel y Teunissen1988). Esta tranferencia de información es necesaria a la hora de unir cada base con el objeto shape, ya que siendo diferentes las propiedades, los objetos no serán compatibles.

```
1 venta<-st_transform(v1.sf, crs = st_crs(shape))
2 alquiler<-st_transform(a1.sf, crs = st_crs(shape))
```

Ya podemos unir cada base de datos con el shape por medio de la función `st_join`.Dicha unión da lugar a dos objetos «sf» y data frame a los que llamaremos «venta» y «alquiler»

```
1 venta<-st_join(venta,shape)
2 alquiler<-st_join(alquiler,shape)
```

Análisis Exploratorio de Datos

Antes de tomar decisiones basándose en la información disponible, uno debe poseer un grado de entendimiento de la misma. El análisis exploratorio y descriptivo (EDA), hace posible una mejor comprensión de los datos. Éste paso es el más relevante, en el que especialmente las organizaciones, cuyas operaciones están basadas en la información, deberían enfocar su energía y recursos.

Existen más de 10.000 paquetes disponibles en R para todo tipo de tareas y dominios. Algunos están vinculados con el EDA y otros lo están indirectamente. Se permite la división del modelado de datos en 3 fases ordinales (Datar y Garg2019):

- Pre-modelado: Se compone de procesos como la visualización y transformación de datos, imputación de valores faltantes y detección de outliers.
- Modelado: Este paso intermedio se basa en modelos de regresión, clasificación, clustering y series temporales.
- Post-modelado: Realiza la interpretación del output.

A modo de demostración, las Figuras 5 y 6 comprimen todas las observaciones (4087 del fichero de venta y 2576 del fichero de alquiler) en la geometría de la Comunidad Autónoma de Madrid, con la finalidad de comprobar que hay áreas con muy escasa oferta (Sierra Norte, Cuenca Alta de Manzanares, Sierra Oeste y Comarca de Las Vegas), concentrándose en la Área Metropolitana del territorio. Recordemos que se han preservado los inmuebles ubicados dentro de los límites de la ciudad aportados por el archivo shape, ya que el estudio tiene como objetivo cuantificar la dependencia espacial de la ciudad como conjunto así como la de los 21 distritos comprendidos.

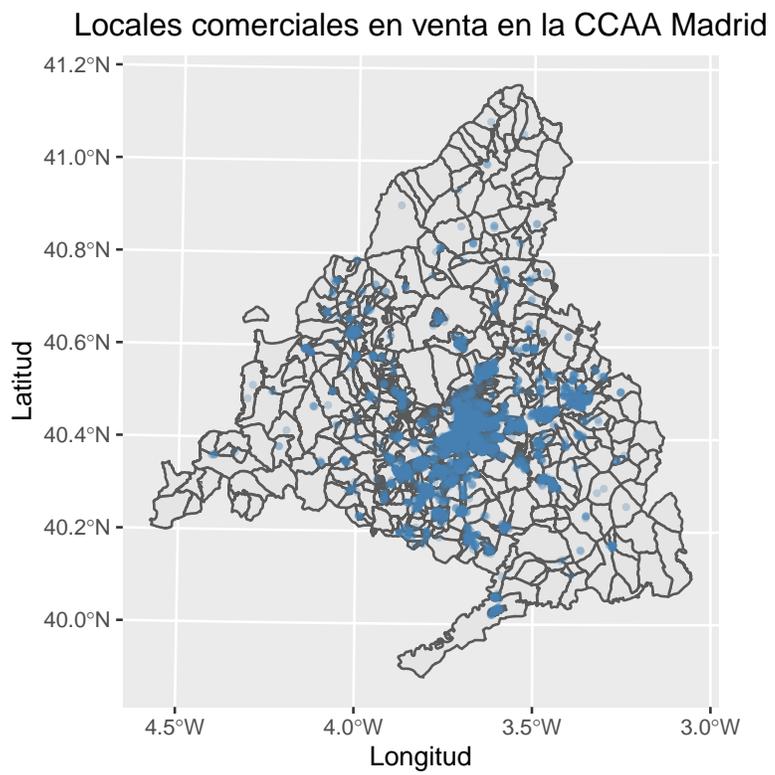


Figura 5: Inmuebles comerciales en venta mapeados en la CCAA de Madrid

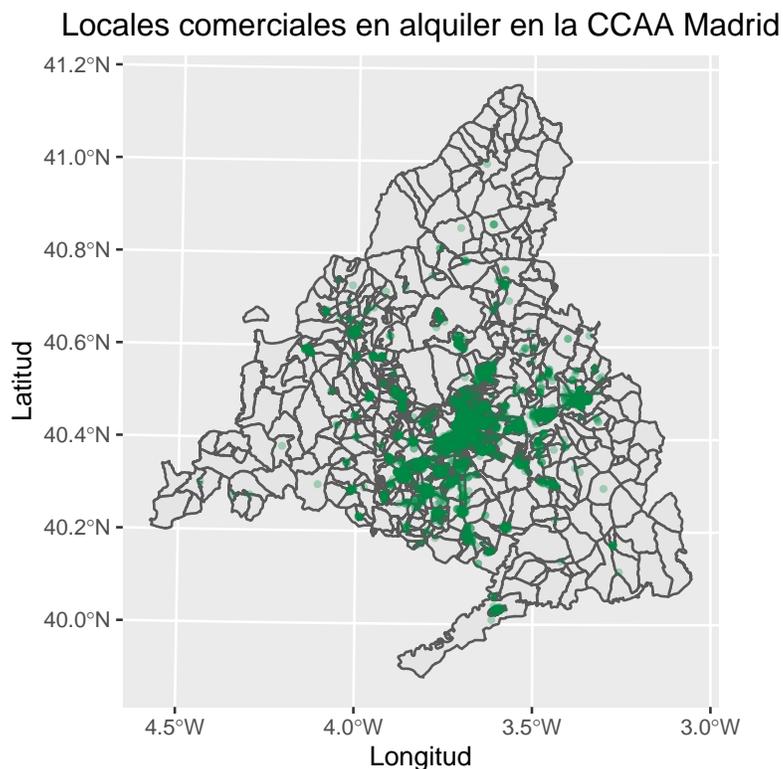


Figura 6: Inmuebles comerciales en alquiler mapeados en la CCAA de Madrid

Tras la eliminación de inmuebles fuera de la ciudad de Madrid, la base de venta cuenta con 1547 observaciones, mientras que la de alquiler cuenta con 955, manteniéndose la idea que la oferta de venta es superior a la de alquiler.

Antes de profundizar en relaciones bivariantes, procedemos a hacer una descripción de los objeto «venta» y «alquiler».

En primer lugar, la variable *id* presenta valores distintos para cada registro, su decodificación en ambos ficheros podrían deberse a una secuencia lógica o podría estar motivada a otros aspectos (ubicación, superficie, precio...), en todo caso, sirve de identificativo para el correcto manejo de quienes pueden gestionar periódicamente los registros.

```
1 length(unique(venta$id))
```

```
1 ## [1] 1547
```

```
1 length(unique(alquiler$id))
```

```
1 ## [1] 955
```

Respecto a los precios, tratándose de una variable económica, podría esperarse una distribución asimétrica y efectivamente hay un rango extremo de precios en ambos ficheros. Puntualmente hablando, vemos en el Cuadro 3 que uno puede adquirir una propiedad comercial por un mínimo de 12.500 euros (basándonos en la muestra), mientras que la propiedad más costosa roza casi los 9 millones de euros. En términos de alquiler, el Cuadro 4 refleja que el alquiler mensual más barato es de 190 € mientras que la cifra más onerosa sobrepasa los 188 mil euros mensuales al tratarse de una propiedad de proporciones colosales.

Summary	Precio
Min	12.500
1er cuartil	130.000
Mediana	230.000
Media	434.360
3er cuartil	457.000
Max	8.990.000

Cuadro 3: Estadísticos resumen de variable precio (datos de venta)

Summary	Precio
Min	190
1er cuartil	950
Mediana	1.600
Media	3.731
3er cuartil	3.000
Max	188.623

Cuadro 4: Estadísticos resumen de variable precio (datos de alquiler)

Por otro lado, hacemos mención al intervalo donde se concentran el 50% de valores para cada mercado, en el caso de venta, los precios se encuentran entre 130.000 € y 457.000 €, para el alquiler, el 50% de las mensualidades están entre los 950 € y 3.000 € euros, esto sin hacer distinción en cuanto al distrito, superficie u otras cualidades que posea el espacio ofertado.

De una manera más formal, la dispersión de un conjunto de datos respecto a su media puede ser calculada con el **coeficiente de variación (CV)**, el cual se obtiene dividiendo la desviación estándar muestral (s) entre la media de los datos (\bar{x}). El CV es un estadístico que podemos representar como porcentaje o sin unidades, para así realizar comparaciones cuando las escalas de valores

no son compatibles. En las siguientes líneas de código vemos que el CV de los precios de alquiler es muy superior a los de venta.

```
1 cv_venta<-round(sd(venta$rawPrice)/mean(venta$rawPrice)*100,2)
2 cv_venta

1 ## [1] 150.3

1 cv_alquiler<-round(sd(alquiler$rawPrice)/mean(alquiler$rawPrice)*100,2)
2 cv_alquiler

1 ## [1] 263.08
```

Haciendo un breve repaso de la variable superficie (*surface*) que, igual a la anterior variable, se caracteriza por una marcada asimetría. A partir del *summary()* de la variable para cada mercado, se confeccionan los Cuadros 5 y 6, permitiéndonos ver que el local en venta más pequeño tiene $5 m^2$ al tratarse de un quiosco de prensa, estando en el extremo opuesto, un inmueble de $12.500 m^2$. El rango de superficies para los espacios en alquiler parte de los $10 m^2$ hasta los $34.295 m^2$.

Medida	m^2
Min	5
1er cuartil	74
Mediana	135
Media	257
3er cuartil	255
Max	12.500

Cuadro 5: Sumario de variable superficie en m^2 (datos de venta)

Medida	m^2
Min	10
1er cuartil	74
Mediana	121
Media	326
3er cuartil	220
Max	34.295

Cuadro 6: Sumario de variable superficie en m^2 (datos de alquiler)

La dispersión estadística de las superficies dada por los coeficientes de variación, nuevamente muestra que hay más dispersión en los datos de alquiler en

relación a la media. Otro aspecto a mencionar es que las extensiones máximas de los inmuebles no pasan desapercibidas, numéricamente, la propiedad en alquiler con mayor cantidad de metros cuadrados es 230 veces mayor al rango intercuartílico. En la modalidad de venta, el valor máximo sólo es 69 veces mayor al rango intercuartílico.

```
1 cv_venta_surface<-round(sd(venta$surface)/mean(venta$surface)*100,2)
2 cv_venta_surface
```

```
1 ## [1] 215.09
```

```
1 cv_alq_surface<-round(sd(alquiler$surface)/mean(alquiler$surface)*
  100,2)
2 cv_alq_surface
```

```
1 ## [1] 484.62
```

Aunque no sean excesivas las propiedades con superficies tan extremas, en definitiva, distorsionan la distribución haciendo el sesgo a la derecha muy superior, complicándose la total apreciación de la densidad de los datos, tal como se observa en la curva de densidad en la Figura 7.

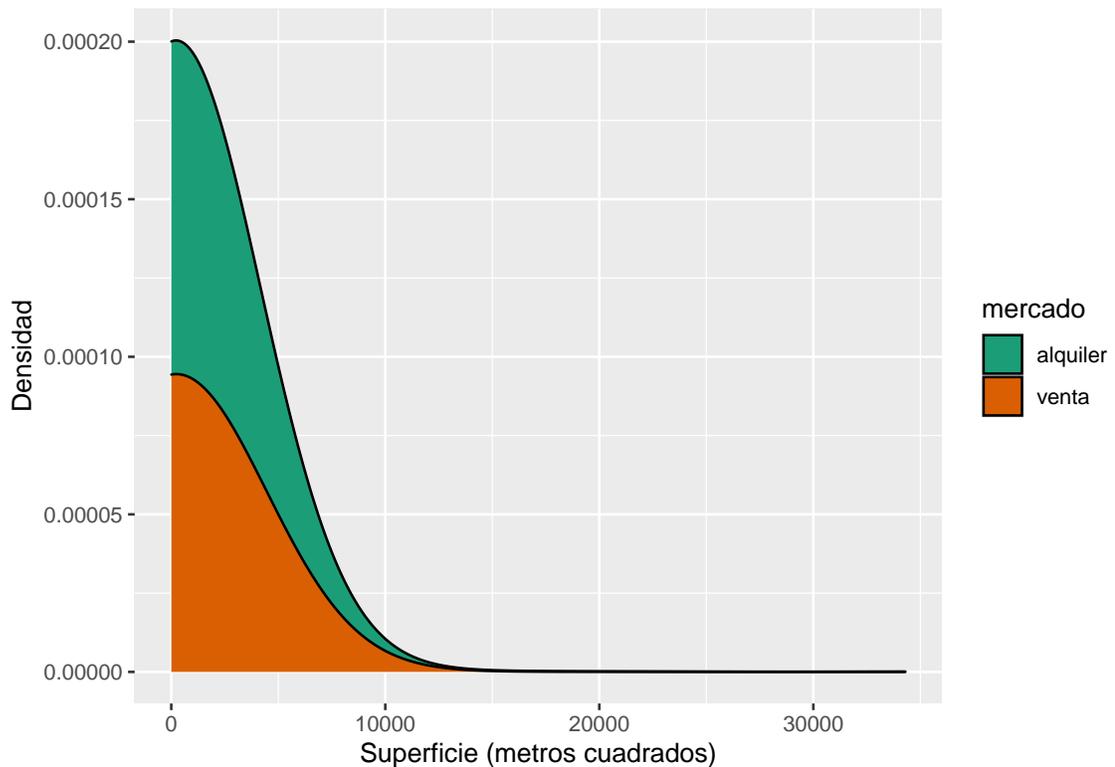


Figura 7: Superficie de locales comerciales por mercado (en metros cuadrados)

Si siguiendo con variables que podrían clasificarse como categóricas, vemos las estructuras de *subtypeld* y *conservationState*:

La variable *subtypeld*, resumida en el Cuadro 7, recoge valores numéricos y adjudica a los registros tres posibles valores [0,13,16], el valor tomado con una frecuencia superior al 90% tanto en los datos de alquiler como de venta, es el 0, seguido del 16 y por último, el 13 es el valor más escaso entre las observaciones. Carecemos del significado exacto de la variable.

Valor	Freq Venta	Freq Alquiler
0	93.54	93.61
13	0.02	0.11
16	6.44	6.28

Cuadro 7: Variable subtype

ConservationState al igual que *floor*, relucen por la gran tasa de valores no disponibles, con la función *freq.na()* de la librería «questionR» (Barnier, Briatte, y Larmarange2020) se consulta el porcentaje de valores faltantes por variable, con los cuales confeccionamos los Cuadros 8 y 9. Posteriormente, se comprueba en el fichero de venta, la no disponibilidad de más de la mitad de información para las variables mencionadas anteriormente, mientras que en el fichero de alquiler, vemos un caso similar.

Cuadro 8: Valores faltantes por variable. Fichero de venta

	Absoluto	Relativo (%)
conservationState	1016	66
floor	899	58
description	31	2
id	0	0
rawPrice	0	0
surface	0	0
subtypeld	0	0
shape_area	0	0
shape_len	0	0
distrito	0	0
location	0	0
geometry	0	0

Cuadro 9: Valores faltantes por variable. Fichero de alquiler

	Absoluto	Relativo (%)
conservationState	520	54
floor	466	49
description	16	2
id	0	0
rawPrice	0	0
surface	0	0
subtypeld	0	0
shape_area	0	0
shape_len	0	0
distrito	0	0
location	0	0
geometry	0	0

Tras comentar de la distribución de precios y otras variables, conduciremos el análisis incorporando la segmentación por distritos, que ha sido posible gracias a las coordenadas de latitud y longitud para llevar a cabo la transformación espacial.

El Cuadro 10 agrupa tanto el número de locales por distrito como el porcentaje de la oferta, además, dado que los distritos están ordenados por orden alfabético, se ha aplicado un formato que configura el color de las celdas de modo que, distritos con ofertas más elevadas serán más oscuras, lo opuesto para distritos con bajas concentraciones de inmuebles.

Cuadro 10: Oferta por mercado

Distrito	Venta		Alquiler	
	Locales	Porcentaje	Locales	Porcentaje
Arganzuela	56	3.62%	50	5.24%
Barajas	14	0.90%	9	0.94%
Carabanchel	140	9.05%	52	5.45%
Centro	111	7.18%	126	13.19%
Chamartin	70	4.52%	70	7.33%
Chamberi	92	5.95%	87	9.11%
Ciudad Lineal	104	6.72%	51	5.34%
Fuencarral	55	3.56%	47	4.92%
Hortaleza	40	2.59%	34	3.56%
Latina	83	5.37%	45	4.71%

Cuadro 10: Oferta por mercado (*continued*)

Distrito	Locales	Porcentaje	Locales	Porcentaje
Moncloa	46	2.97%	32	3.35%
Moratalaz	10	0.65%	10	1.05%
Puente de Vallecas	102	6.59%	23	2.41%
Retiro	44	2.84%	37	3.87%
Salamanca	98	6.33%	93	9.74%
San Blas	77	4.98%	40	4.19%
Tetuan	130	8.40%	49	5.13%
Usera	54	3.49%	14	1.47%
Vicalvaro	55	3.56%	4	0.42%
Villa de Vallecas	86	5.56%	52	5.45%
Villaverde	80	5.17%	30	3.14%

Teniendo conocimiento de los porcentajes de oferta por distrito, otra herramienta que nos ayudaría a extraer información útil serían las Figuras 8 y 9, las cuales representan la geometría del área de estudio y las distintas intensidades de la oferta, obtenidas tras dividir el total de inmuebles de cada distrito entre la cantidad total de inmuebles. Ambas figuras realizadas con la librería «tmaps» (Tennekes2021) exponen claramente si áreas con alta o baja concentración de locales se encuentran rodeadas de áreas similares o no.

Oferta de locales en venta

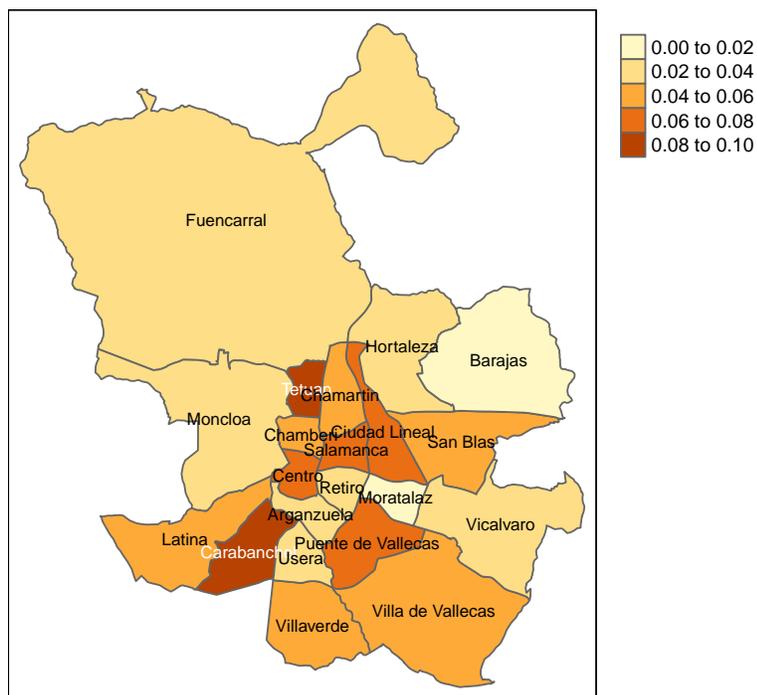


Figura 8: Intensidad de oferta de venta por distritos

Dentro de los límites de la ciudad de Madrid, la concentración de propiedades ocurre principalmente en el interior para ambos mercados, aunque con más incidencia en distritos de la periferia sur en el caso de los locales en venta. En términos generales, la Figura 8, nos permite ver que la oferta es diluida en el norte, se concentra en el centro y valores medio altos se hacen presente en el sur. Carabanchel es el distrito periférico con cuota más significativa. Los distritos cercanos a éste: Latina y Villaverde poseen valores medio altos, no obstante, Usera y Arganzuela presentan porcentajes inferiores.

En el norte, se manifiestan volúmenes bajos en Moncloa, Fuencarral y Hortaleza pero no tanto como en Barajas.

En el centro de la ciudad, destaca Tetuán, existiendo polaridad entre sus distritos próximos, algunos cuentan con pesos inferiores, como Fuencarral (3.56%) y Moncloa (2.97%), otros cuentan con porcentajes más elevados como Chamartín con 4.52% y Chamberí que roza el 6%.

Claramente, Centro, Salamanca y Ciudad Lineal forman un espacio de alta concentración de oferta de inmuebles comerciales en venta, próximo a este foco,

San Blas-Canillejas adopta un valor aceptable (casi 5%) a diferencia de Arganzuela (3.62%), Retiro (2.84%) y Moratalaz (0.65%).

Oferta de locales en alquiler

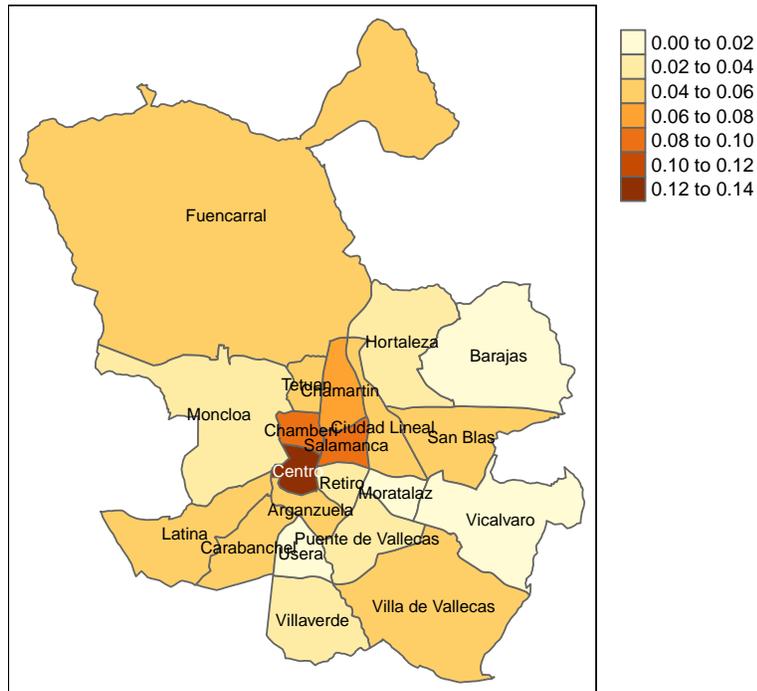


Figura 9: Intensidad de oferta de alquiler por distrito

En referencia a las propiedades en alquiler, en la Figura 9, Centro refleja una vasta superioridad respecto al resto de distritos, aunque áreas cercanas como Chamberí y Salamanca cuentan con valores medio altos.

Volvemos a remarcar la presencia de *outliers* en los precios, pero ahora podemos saber con certeza dónde se localizan estos valores tan elevados. Como paso ordinario, aportamos los Cuadros 11 y 12, que reflejan información sobre los precios de cada mercado, tanto los precios medios, los mínimos y máximos de cada distrito junto con la columna «Diferencia», que como su nombre indica, es la diferencia entre el precio mínimo y máximo de ese distrito.

Cuadro 11: Precios medios de venta por distritos (en euros)

Distrito	Precio medio	MIN	MAX	Diferencia
Salamanca	1013921	49000	7000000	6951000
Centro	603101	35000	3350000	3315000
Fuencarral	591792	18000	3900000	3882000
Retiro	591136	33000	2213000	2180000
Moncloa	552531	59500	4000000	3940500
Chamartin	534747	55300	2900000	2844700
Chamberi	500201	35000	2800000	2765000
Arganzuela	481069	70000	1990000	1920000
Tetuan	464634	49700	2300000	2250300
Villaverde	439634	30000	8990000	8960000
Hortaleza	410668	16500	2999000	2982500
San Blas	379089	20900	3400000	3379100
Villa de Vallecas	345884	13000	3700000	3687000
Ciudad Lineal	342213	20900	5950000	5929100
Carabanchel	276894	16500	3125000	3108500
Latina	263526	12500	2500000	2487500
Usera	253898	40000	2250000	2210000
Moratalaz	229388	85000	474875	389875
Barajas	169080	29260	600000	570740
Vicalvaro	166533	40000	678000	638000
Puente de Vallecas	162605	37000	1000000	963000

Dentro del mercado de venta, Salamanca encabeza la lista (Cuadro 11), con un precio medio de 1.013.921 €, y con una dispersión que limita con los 7.000.000 €, seguida de Centro, Fuencarral y Retiro con precios medios aproximadamente 40% inferiores que al de Salamanca. El formato de la columna «Diferencia» nos permite ver que Villaverde tiene el rango de precios más prolongado que se acerca a los 9.000.000 €.

Si volvemos a Carabanchel y Tetuán, quienes acaparaban gran parte del porcentaje de la oferta de venta, el primero podría clasificarse como uno de los distritos más baratos al posicionarse entre los últimos puestos del listado, gracias a un precio medio de 276.894 €, por otro lado, Tetuán se encuentra en el top diez de distritos con precios medios más altos, con 464.634 €.

Barajas y Moratalaz ocupaban los últimos lugares en cuanto a cuota de oferta

de locales en venta. También estos ofrecen algunos de los precios más accesibles de la ciudad, 169,080 € y 229.387 €. Hay distritos que presentan cifras menores a las ya mencionadas, como Vicálvaro (166.533 €) y Puente de Vallecas (162.605 €).

Cuadro 12: Precios medios de alquiler por distritos (en euros)

Distrito	Precio medio	MIN	MAX	Diferencia
Vicalvaro	17188	300	66551	66251
Villaverde	11776	350	188623	188273
Moratalaz	5988	350	50000	49650
Villa de Vallecas	5239	250	65642	65392
Centro	3584	400	25000	24600
Salamanca	3318	380	22000	21620
Tetuan	3235	450	13000	12550
San Blas	3201	300	15000	14700
Moncloa	3094	800	14000	13200
Carabanchel	2865	275	30000	29725
Chamartin	2783	600	18000	17400
Barajas	2738	700	11000	10300
Hortaleza	2313	250	10000	9750
Retiro	2271	400	8300	7900
Usera	2191	450	12000	11550
Chamberi	2012	480	6500	6020
Arganzuela	1835	600	6000	5400
Ciudad Lineal	1754	250	17000	16750
Fuencarral	1702	350	9000	8650
Latina	1146	190	4380	4190
Puente de Vallecas	982	300	3300	3000

Pasando al mercado de alquiler, pasamos a ver el Cuadro 12, que recopila el importe mensual de alquiler de cada distrito. Vicálvaro, distrito con menos locales ofertados, presenta el alquiler medio más caro de 17.188 €/mes, Villaverde y Moratalaz suceden al anterior con cifras de 11.776 €/mes y 5.988 €/mes. Cabe mencionar que Villaverde tiene una mayor dispersión en sus precios de alquiler, al ser su diferencia más elevada.

Centro, que contenía la mayor cuota de oferta de alquiler, se encuentra dentro del top cinco de precios medios de alquiler más caros con 3,584 €/mes, aunque

al comparar dicha cantidad con las anteriores a esta, nos percatamos de una clara inferioridad. Entre las últimas posiciones, encontramos a Fuencarral (1.702 €/mes), Latina (1.146 €) y Puente de Vallecas con 987 €/mes de media.

Adicionalmente, ofrecemos las Figuras 10 y 11, correspondientes a las distribuciones del precio por metro cuadrado para cada mercado. Observando la longitud de las cajas de los *boxplots* podemos detectar donde se encuentra mayor dispersión de precios por metro cuadrado.

En el mercado de venta, Salamanca y Centro presentan los rangos más amplios de la ciudad. También, vemos distribuciones que podríamos clasificar de simétricas, al estar la mediana prácticamente en el centro de las cajas, como en el caso de Villa de Vallecas, Usera, Tetuán, San Blas-Canillejas, Puente de Vallecas y Arganzuela. No obstante, vemos que en todos los distritos hay colosales valores extremos, excepto en Moratalaz. Por último, hacemos mención a Vicálvaro, cuyos valores de precio/metro cuadrado presentan una gran asimetría negativa.

Por parte del mercado de alquiler, la Figura 11 indica que los distritos centrales ofrecen una mayor dispersión de precios por metro cuadrado, donde destacamos a Salamanca y a Centro. Por otra parte, los distritos que parecen tener precios por metro más accesibles son los periféricos como Carabanchel, Vicálvaro y Villa-verde.

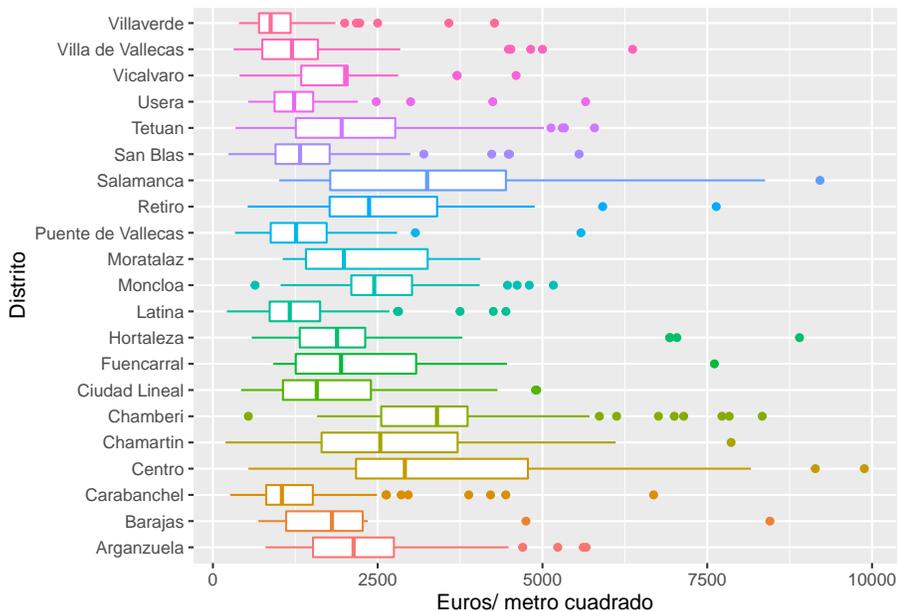


Figura 10: Precio (en euros) por metro cuadrado. Base de venta

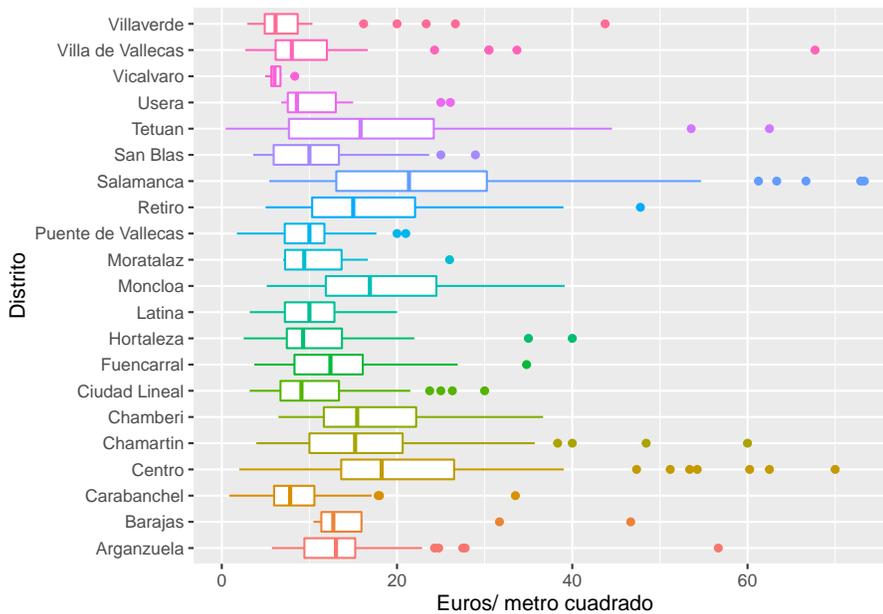


Figura 11: Precio (en euros) por metro cuadrado. Base de alquiler

En situaciones puntuales, es una buena estrategia ignorar los valores extremos, ya sea eliminándolos cuando son pocos o imputándoles valores dictados por el rango intercuartil. También, valores tan extremos pueden ser producto de un error de registro. Trantándose de inmuebles comerciales, su precio está sujeto a sus características particulares, superficie, infraestructura, seguros, equipación, entre otros, por tanto no podemos esperar una disposición perfecta de valores.

Por ejemplo, el espacio más costoso en Moncloa (4 millones de euros), ubicado en Paseo Pintor Rosales, de $760 m^2$, puede atribuir su precio, no solo a su gran extensión sino también a peculiaridades como la posesión de licencia para cafetería y espectáculo, sistemas de insonorización, de protección contra incendios, instalaciones adaptadas a minusválidos e instalaciones eléctricas según la normativa.

Se analizó la variable cuantitativa *surface*, y considerando la notable disparidad en su distribución, decidimos crear una variable categórica que asigne cada observación a un grupo en relación con su extensión en metros cuadrados, la cual llamaremos *tipo*:

Tipos de locales:

- Pequeños $\leq 150 m^2$
- Medianos 151 a $350 m^2$

- Grandes 351 a 550 m^2
- Industriales <550 m^2

Una vez segmentados los locales, analizamos el volumen por tipologías. La Figura 12 muestra que las propiedades de hasta 150 m^2 son el grupo predominante en ambos mercados, con más del 50% de la cuota en ambos casos. En segundo lugar, siguen los locales medianos con un 30,2% (venta) y 26,5% (alquiler). El orden que esperaríamos es que los locales industriales fuesen la tipología más inusual, sin embargo, estos espaciosos inmuebles superan a los grandes, aunque por una ventaja reducida.

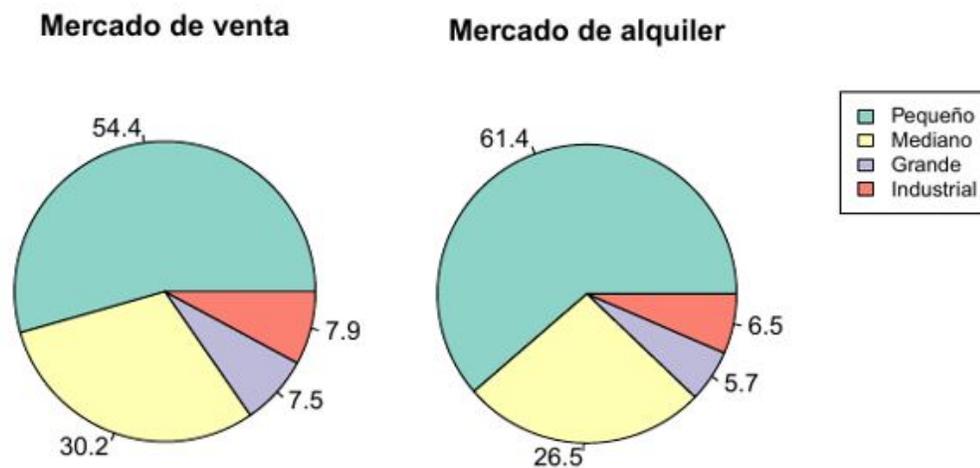


Figura 12: Segmentación de locales comerciales por tipología

Ubicación de los locales pequeños

De la adjudicación territorial de los inmuebles pequeños podemos decir que, en cuanto al alquiler, las circunscripciones líderes son metropolitanas con cuotas superiores a 10% (Figura 13), entre las que se destaca Centro con 14.33%, seguida por Chamberí (10.92%) y Salamanca (10.07%). Se destaca la presencia de zonas no metropolitanas ni periféricas en el top 10, en el caso de Ciudad Lineal y Arganzuela con la misma cuota de 5.6%, también se aprecia a Fuencarral con 5.8%.

En términos de venta (Figura 14), Puente de Vallecas cuenta con la cuota más sustancial de locales (8.91%). Dentro de los 10 distritos que más acogen dichos espacios, hay un reparto balanceado entre zonas centrales y periféricas, también hacemos mención a Ciudad Lineal posicionándose dentro de las cinco zonas con más locales de la mencionada tipología con un 7.13%.

Entre las circunscripciones con menos oferta de locales, ya sea del mercado de alquiler o venta, encontramos distritos medios, como Moratalaz y Usera, como periféricos: Villaverde, Moncloa, Barajas, Vicálvaro. Excepcionalmente, encontramos al distrito céntrico Retiro dentro de las zonas con menos oferta en términos de venta (1.9%).

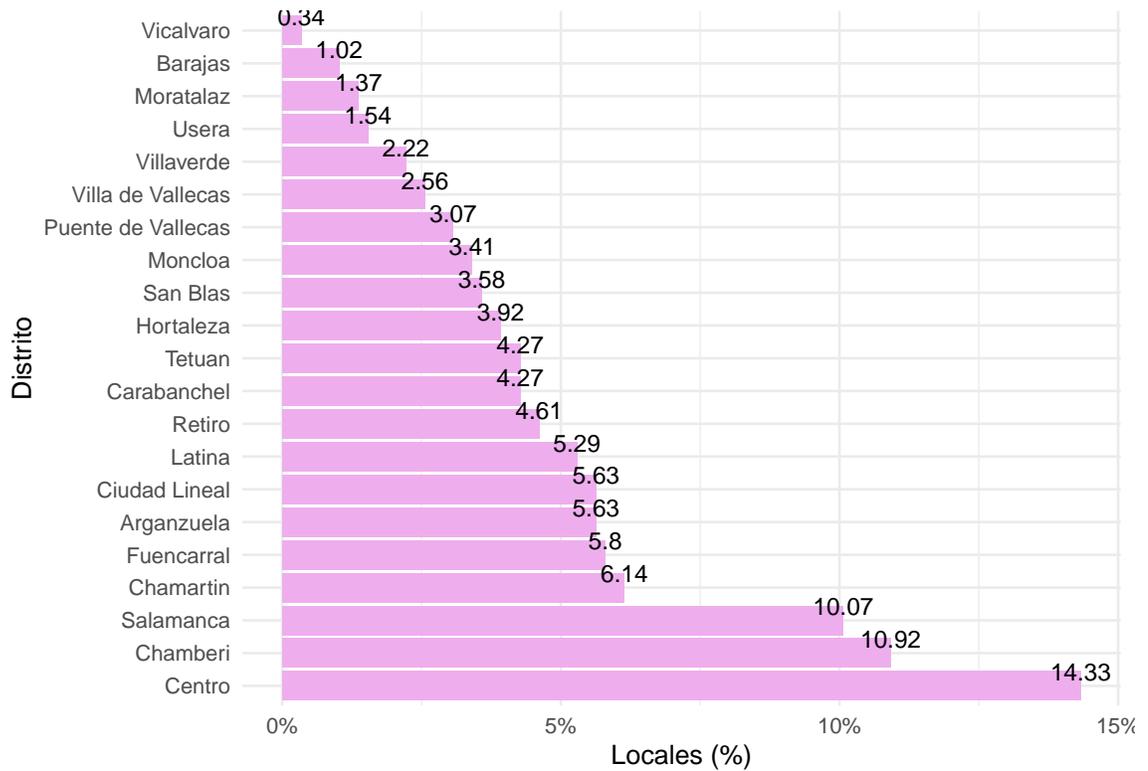


Figura 13: Cuota de locales pequeños en alquiler por distritos

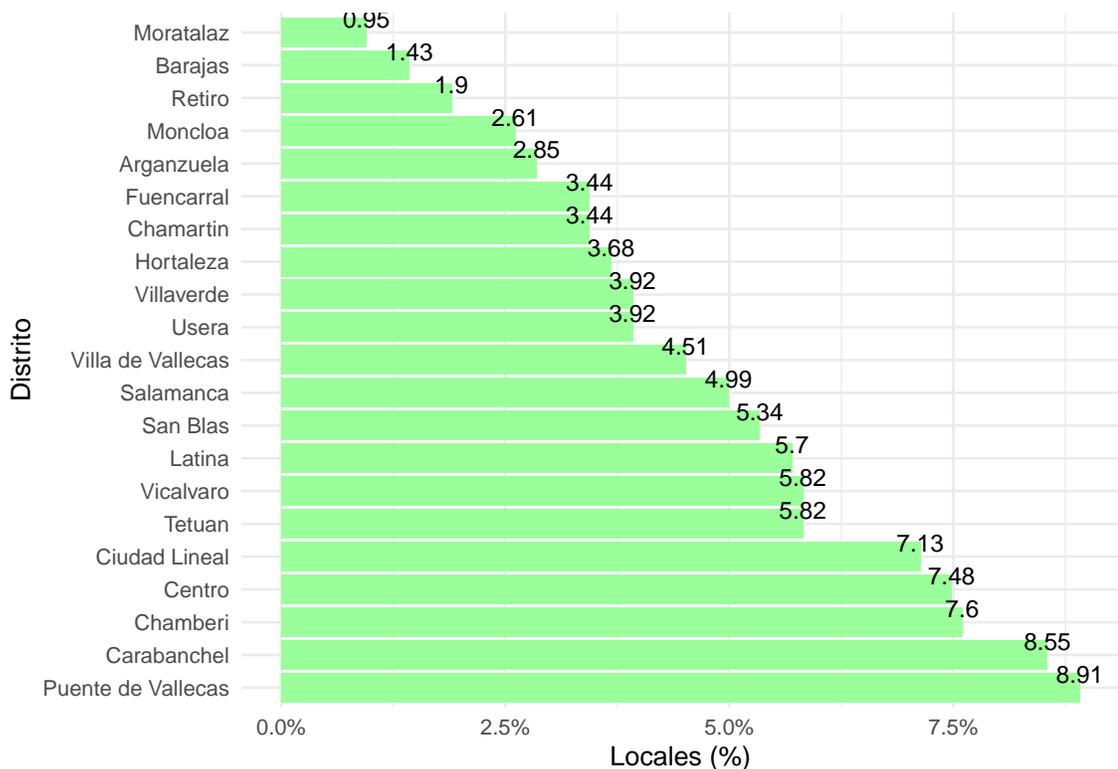


Figura 14: Cuota de locales pequeños en venta por distritos

Ubicación de los locales medianos

El reparto de locales en venta, según la Figura 15, es liderado por Tetuán (13.49%), con una moderada ventaja respecto a Carabanchel (9.42%), que ocupa el segundo lugar, y consecuentemente el resto de demarcaciones. Además de Tetuán, las céntricas Salamanca y Centro se encuentran en el top cinco, mientras que los distritos siguientes son en su mayoría periféricos o limitantes con la metrópolis.

En cuanto al alquiler, los tres primeros distritos son céntricos y concentran un poco más del 35% de locales (Figura 16), por otro lado, Villa de Vallecas posee un 9.49% de los locales.

Asimismo, encontramos coincidencias para ambos mercados respecto a los distritos con menos cuota de locales medianos, siendo el caso de Usera, Moratalaz, Barajas y Vicalvaro.

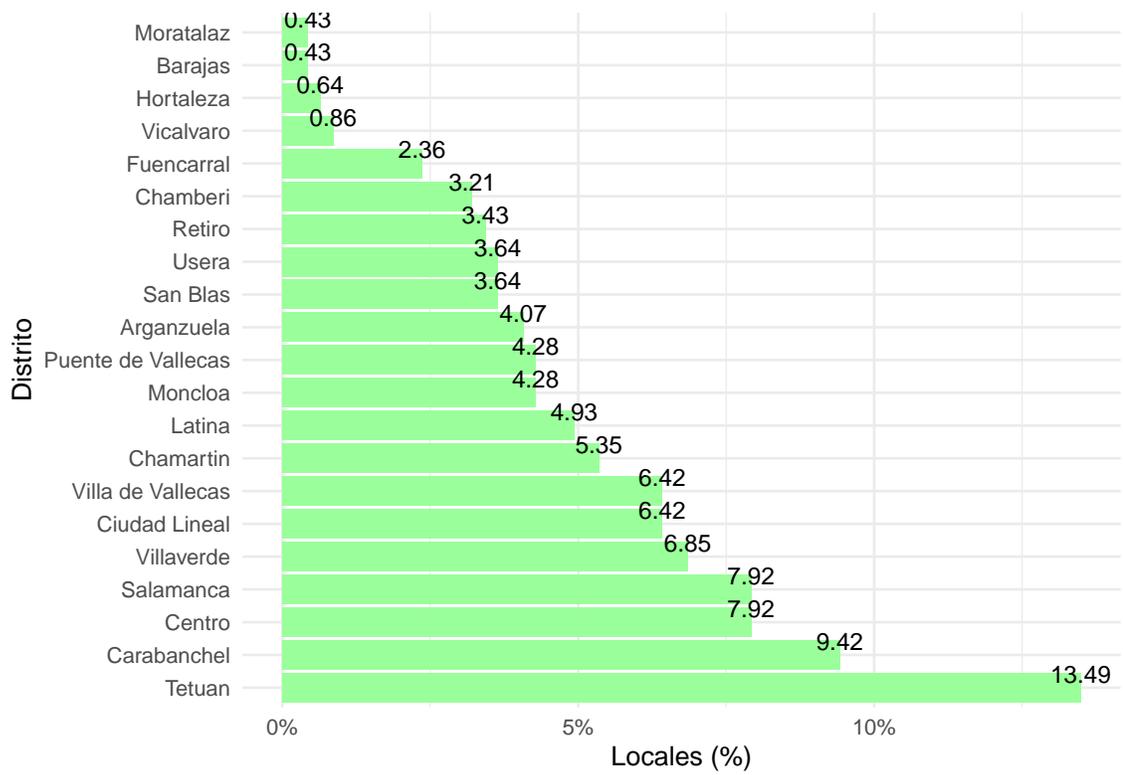


Figura 15: Porcentaje de locales medianos en venta por distritos

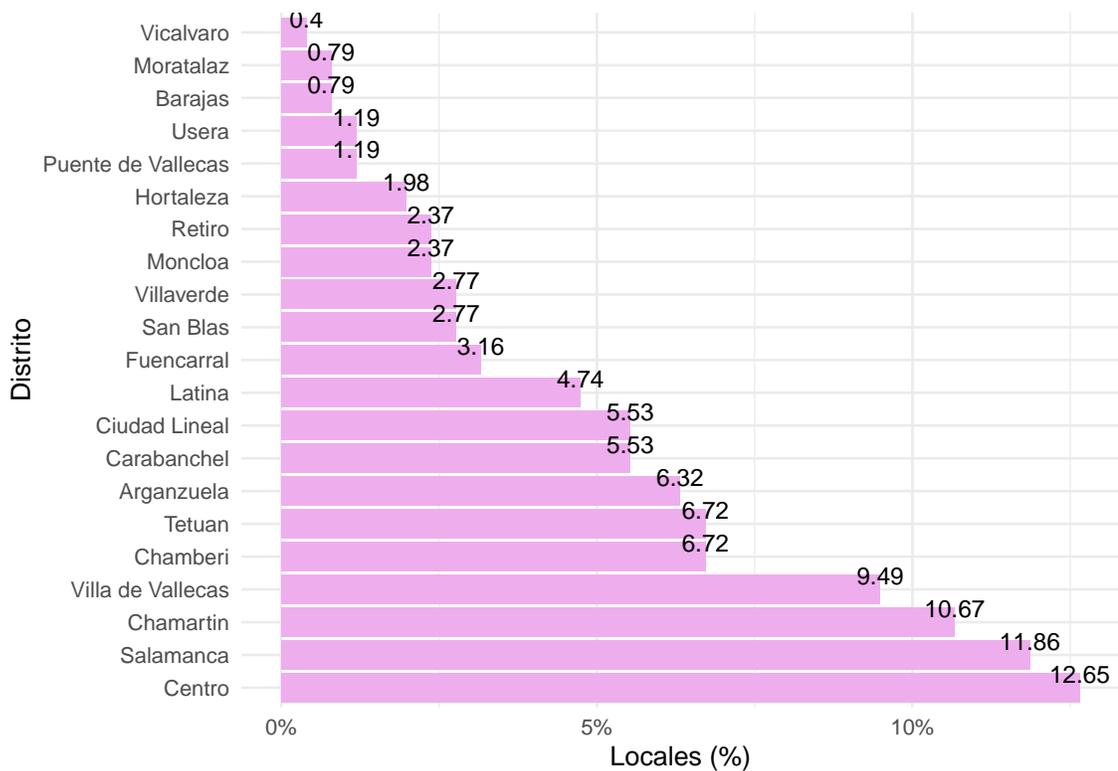


Figura 16: Porcentaje de locales medianos en alquiler por distritos

Ubicación de los locales grandes

Comparando las Figuras 17 y 18, vemos una mezcla de zonas centrales y periféricas donde se concentran la mayor parte de este tipo de inmuebles, como Tetúan, con más opciones de venta, mientras que Centro y Chamberí ofrecen más oportunidades de alquiler.

Hay rasgos comunes en ambos mercados, como la nula oferta en Barajas y Moratalaz. Hay un mínimo porcentaje de locales en distritos periféricos como en Usera, San Blas-Canillejas, Hortaleza, Arganzuela, Puente de Vallecas y Vicálvaro.

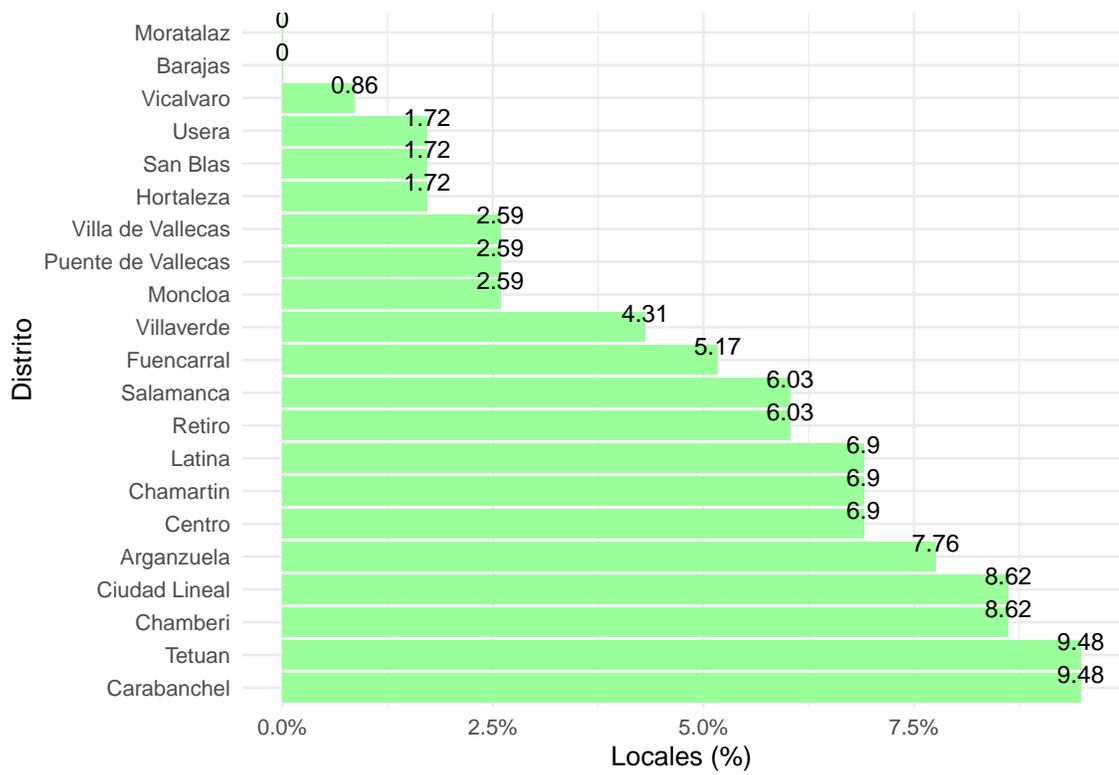


Figura 17: Porcentaje de locales grandes en venta por distritos

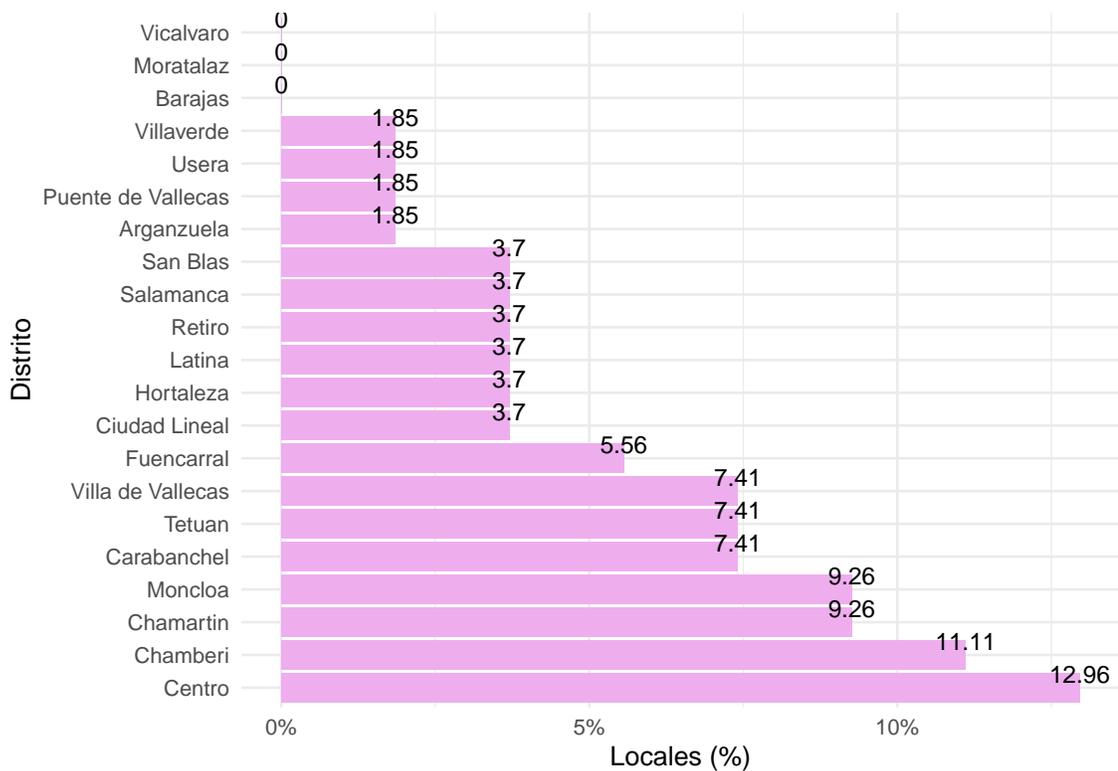


Figura 18: Porcentaje de locales grandes en alquiler por distritos

Ubicación de los locales industriales

Bien sea en términos de compra o alquiler (Figuras 19 y 20), en la periferia es donde se encuentran la mayoría de espacios de más de $550 m^2$. Villa de Vallecas, San Blas, Carabanchel y Villaverde destacan por concentrar la mayoría de estos inmuebles. Destacamos a la central Salamanca por ser el cuarto distrito con más espacios industriales disponibles en régimen de venta (9.84%).

Hay nula oferta de este tipo de locales en Moratalaz (venta y alquiler), en Latina, Chamberí y Arganzuela hay reducidas opciones de venta y nulas de alquiler. Barajas no ofrece la venta de estos locales y una única opción de alquiler.

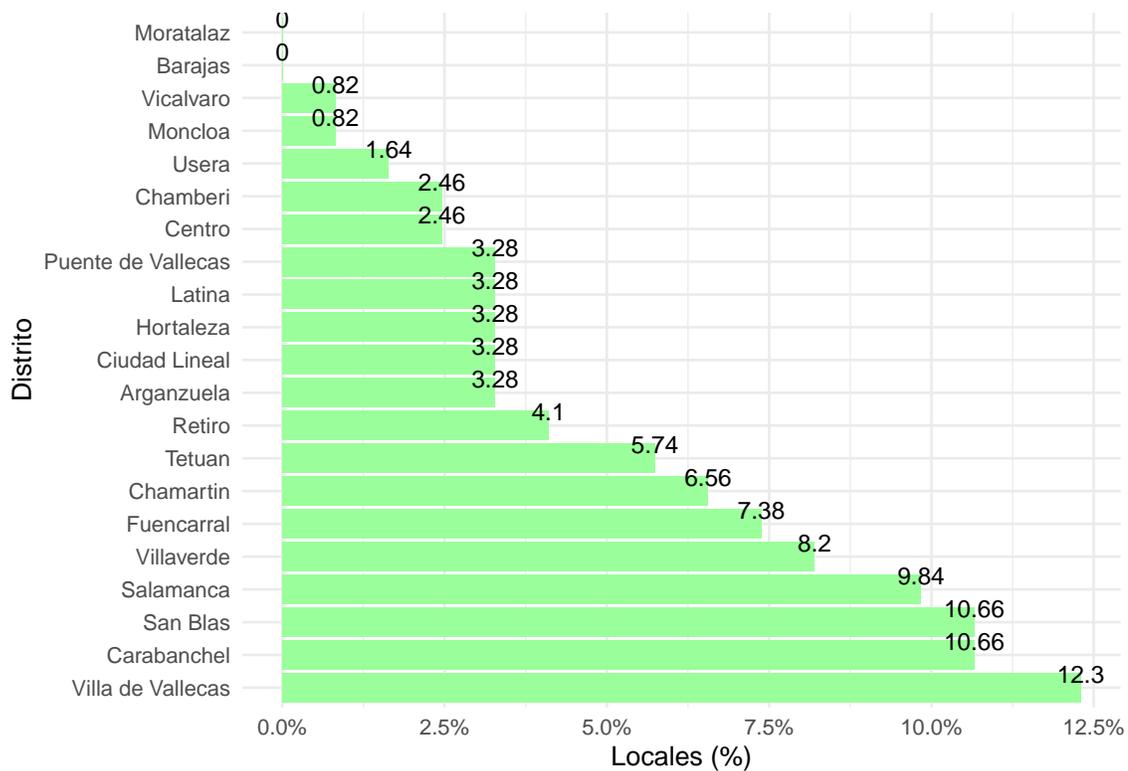


Figura 19: Porcentaje de locales industriales en venta por distritos

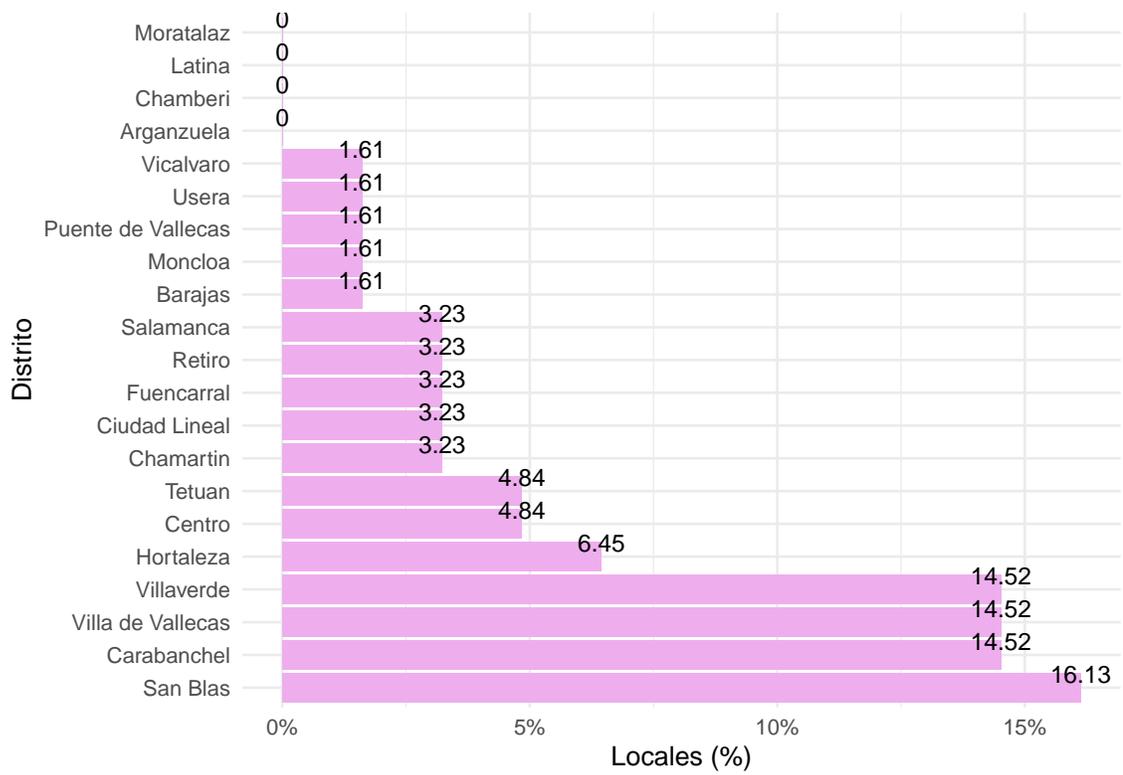


Figura 20: Porcentaje de locales industriales en alquiler por distritos

Índices espaciales

Índice global de Moran

Hay diversas maneras de definir regiones vecinas: una es la cantidad de bordes/fronteras comunes o bien la distancia entre un punto de referencia con cada región. La estructura de los vecinos puede ser almacenada en una matriz binaria W , donde cada entrada w_{ij} es 1 o 0; 1 si el par de regiones son vecinas, y 0 si el par de regiones no son vecinas.

De la matriz se excluyen los *self neighbors* (vecino de uno mismo), la diagonal principal es 0. La matriz es de dimensión $n \times n$, donde n es el número de regiones.

Los vecinos pueden ser definidos por medio de órdenes que referencian los movimientos de ciertas piezas de ajedrez. La Figura 21 hace referencia a dichos órdenes:

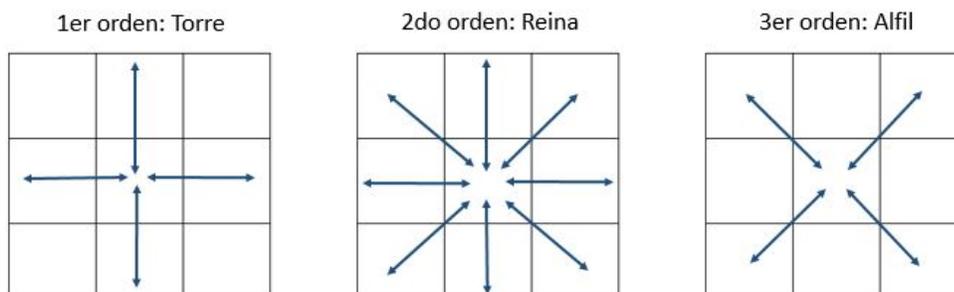


Figura 21: Órdenes para el cálculo de los vecindarios

Con ayuda de la librería «spdep» (Bivand, Pebesma, y Gomez-Rubio2013), creamos una matriz espacial tipo nb , para ambos sets. Utilizamos el orden de la reina, por lo que fijamos el parámetro queen (reina) como «TRUE».

```
1 ## Definir poligonos (vecinos)
2 nb_venta <- poly2nb(venta)
3 nb_alquiler <- poly2nb(alquiler)
```

Al representar gráficamente la existencia de relaciones entre vecindarios con una línea, obtenemos otro diagrama. Nodos que representan regiones y las líneas son vínculos conectando los nodos al ser las regiones vecinas como muestra la Figura 22.

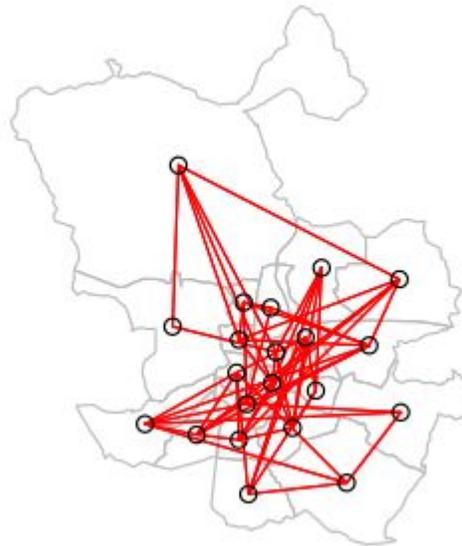


Figura 22: Vecindario utilizando el criterio de la reina

El mismo resultado es almacenado en una lista debido a que la función para calcular el estadístico de Moran requiere un objeto tipo lista (lw).

```
1 lw_venta <- nb2listw(nb_venta, zero.policy = T)
2 lw_alquiler <- nb2listw(nb_alquiler, zero.policy = T)
```

El contraste de hipótesis a realizar sobre ambas bases de datos es planteado como:

H_0 : *Los precios de los locales son independientes espacialmente*

H_a : *Hay interdependencia espacial en los precios de los locales comerciales*

Aplicaremos dos métodos al conducir el test, uno analítico y una simulación Monte Carlo.

Método analítico

Se requiere la función `moran.test()` para ambos sets de datos. Una vez ejecutada la función, consultamos las salidas de ambos objetos «globalventa» y «globalalquiler».

```

1 globalventa<-moran.test(venta$rawPrice, lw_venta)
2 globalalquiler<-moran.test(alquiler$rawPrice, lw_alquiler)

1 globalventa

1 ##
2 ## Moran I test under randomisation
3 ##
4 ## data: venta$rawPrice
5 ## weights: lw_venta
6 ##
7 ## Moran I statistic standard deviate = 26.197, p-value < 2.2e-16
8 ## alternative hypothesis: greater
9 ## sample estimates:
10 ## Moran I statistic      Expectation      Variance
11 ##      3.692153e-02      -6.468305e-04      2.056528e-06

1 globalalquiler

1 ##
2 ## Moran I test under randomisation
3 ##
4 ## data: alquiler$rawPrice
5 ## weights: lw_alquiler
6 ##
7 ## Moran I statistic standard deviate = 5.785, p-value = 3.626e-09
8 ## alternative hypothesis: greater
9 ## sample estimates:
10 ## Moran I statistic      Expectation      Variance
11 ##      9.883449e-03      -1.048218e-03      3.570819e-06

```

Para las propiedades en alquiler de la I de Moran es 0,00988. Para los datos de locales en venta, se obtiene una I de 0,0363.

Dichos coeficientes, dan la impresión de una inminente independencia espacial al ser tan próximos a 0. No obstante, los índices anteriores no son suficientes para descartar la H_a , debemos observar los *p-valor* obtenidos para ambos test, estos al ser tan inferiores indican que la probabilidad de que los conjuntos de datos estudiados se deban al azar es prácticamente 0. Si bien recordamos, al inicio de la sección de autocorrelación espacial, mencionamos la importancia de la escala durante la identificación de patrones, por ello procedemos a realizar simulaciones de Montecarlo, cuyo objetivo es corroborar la significatividad estadística del contraste de hipótesis y como último paso, se calcula el índice de Moran local para cada distrito.

Simulación Monte Carlo

Aunque el método anterior es rápido, puede ser sensible a irregularidades distribuidas en los polígonos. Para llevar a cabo esta alternativa, empleamos la función `moran.mc()`, cuya sintaxis contiene el parámetro n , el número de simulaciones que deseamos se lleve a cabo en la simulación.

El enfoque de éste método es asignar valores aleatorios del atributo seleccionado (precio) a los polígonos en cada permutación, lo que conlleva a un índice de Moran por cada simulación. Como resultado se tiene una distribución de coeficientes de Moran bajo la hipótesis nula que nos permite contrastar más fácilmente la autocorrelación de los datos.

Una vez más, se realizan dos simulaciones para ambos ficheros de datos, cuyos resultados se almacenan en dos objetos llamados «MC» y «MC2» para los inmuebles en venta y alquiler respectivamente.

```
1 MC<-moran.mc(venta$rawPrice, lw_venta, nsim = 999, alternative = "  
  greater")  
2 MC
```

```
1 ##  
2 ## Monte-Carlo simulation of Moran I  
3 ##  
4 ## data: venta$rawPrice  
5 ## weights: lw_venta  
6 ## number of simulations + 1: 1000  
7 ##  
8 ## statistic = 0.036922, observed rank = 1000, p-value = 0.001  
9 ## alternative hypothesis: greater
```

```
1 MC2<-moran.mc(alquiler$rawPrice, lw_alquiler, nsim = 999, alternative = "  
  greater")  
2 MC2
```

```
1 ##  
2 ## Monte-Carlo simulation of Moran I  
3 ##  
4 ## data: alquiler$rawPrice  
5 ## weights: lw_alquiler  
6 ## number of simulations + 1: 1000  
7 ##  
8 ## statistic = 0.0098834, observed rank = 1000, p-value = 0.001  
9 ## alternative hypothesis: greater
```

Tras mil simulaciones para cada caso, los *p-valor* generados son mínimos.

Pudiendo los procesos anteriores ser graficados, las curvas representan los valores de las *I*'s de Moran que podríamos esperar si los precios siguieran un

patrón aleatorio en los distritos. Nótese que ambos *Moran I Statistic* (Figuras 23 y 24), se posicionan a la derecha de la distribución aleatoria, pero la distancia en los datos de venta es mucho mayor a la de los datos de alquiler, indicando una mayor agrupación (clústers) en los precios de venta.

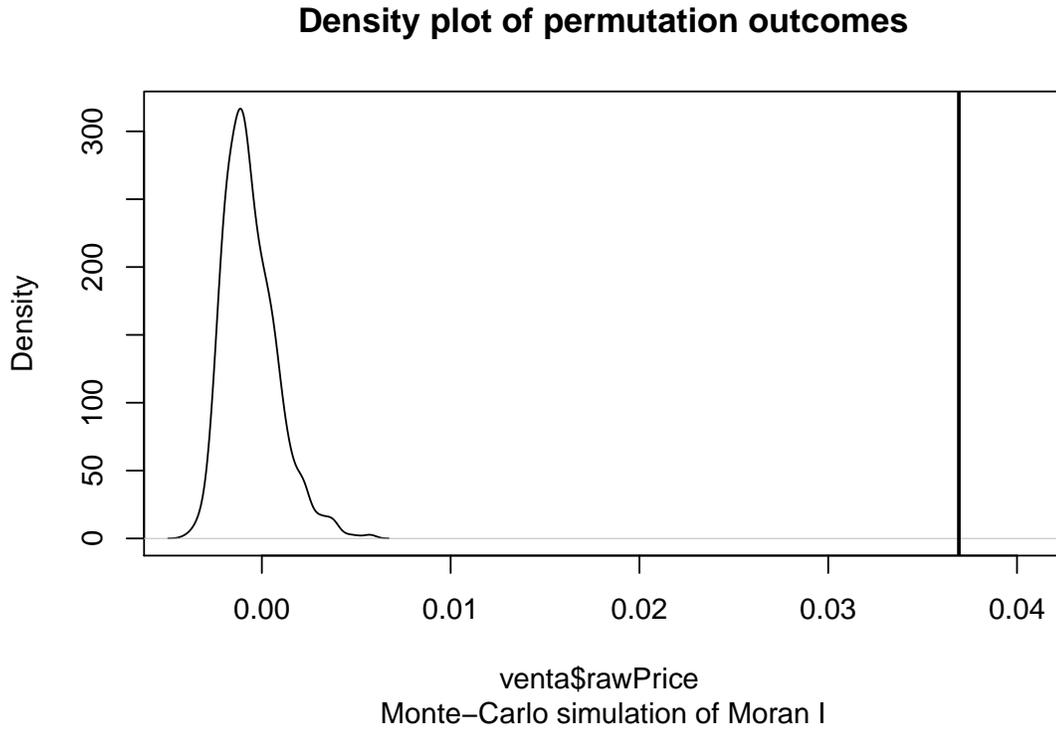


Figura 23: Simulación de Montecarlo para los datos de venta

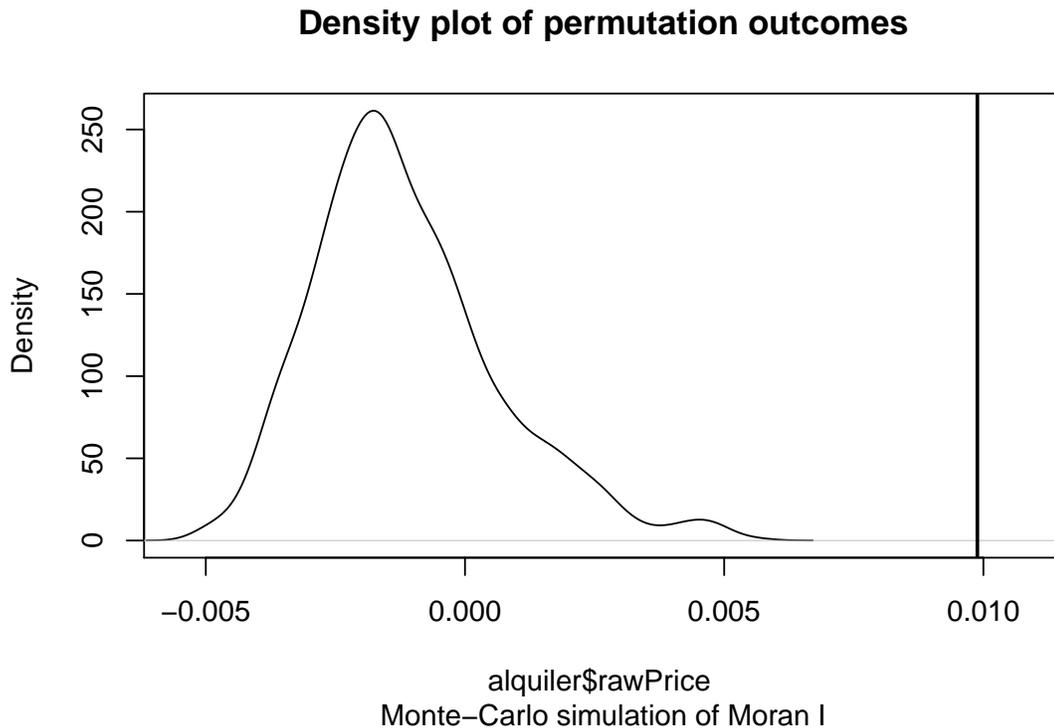


Figura 24: Simulación de Montecarlo para los datos de alquiler

Índice local de Moran

El I de Moran local identifica agrupaciones espaciales o valores atípicos para comprender la contribución de cada unidad espacial al estadístico. Fue desarrollado por Anselin (1995) y se considera un tipo de Indicadores Locales de Asociación Espacial (LISA). Los LISA tienen dos propósitos en el análisis de datos espaciales exploratorios (ESDA): indican grupos espaciales locales y realizan análisis de sensibilidad (identifican valores atípicos).

Anselin define a los LISA como:

- El LISA para cada observación da una indicación de la extensión del agrupamiento espacial significativo de valores similares alrededor de cada observación; y
- La suma de LISA para todas las observaciones es proporcional a un indicador global de asociación espacial.

Al igual que con las medidas globales, los LISA prueban si el patrón espacial observado de una variable de interés entre áreas es extremo, probable o esperado, dada una distribución geográfica aleatoria de la variable.

Se calcula con la siguiente fórmula:

$$I_i = Z_i \sum_j W_{ij} * Z_j$$

Donde Z_i y Z_j son observaciones en desviaciones de la media, W_{ij} es la matriz de peso espacial.

La interpretación es similar a la aplicada para el análisis global. Valores positivos de I_i sugieren la existencia de un cluster espacial, donde valores similares se agrupan (sean altos o bajos), y los valores negativos de I_i representan un cluster con valores análogos.

Las Figuras 25 y 26 muestran el nivel de significatividad estadística, pudiendo considerarse como la contribución que cada región aporta al resultado de la autocorrelación espacial global, siendo las zonas de color verde oscuro las más significativas.

Esto se determina utilizando un procedimiento de aleatorización de Monte Carlo.

Viendo la Figura 25, observamos consistencia en las unidades espaciales, es decir, si todo el mapa tomara un color amarillo pálido excepto Arganzuela y Retiro, significaría que la autocorrelación global se debe únicamente a la covarianza de estos dos distritos. En este caso, vemos distritos altamente significativos, especialmente ubicados al sur: Carabanchel, Villa de Vallecas y Vicálvaro junto con Moratalaz y Salamanca más en el centro de la ciudad.

Al norte de la ciudad, se diluye el aporte de los distritos al estadístico global pero entre los menos significativos se listan: Moncloa, Tetuán, Retiro, Arganzuela, Ciudad Lineal y San Blas-Canillejas.

Avanzando a los resultados de precios de alquiler, la Figura 26 presenta un panorama muy distinto al anterior, en primer lugar, la mayor parte de unidades espaciales con contribución más significativa al índice global se encuentran en los extremos de la ciudad, como Fuencarral al norte y Latina al oeste, por otro lado, Vicálvaro y Villa de Vallecas al sureste, vuelven a darse como distritos con alta significatividad. En cuanto la zona central, destacamos a Centro y con menos intensidad a Ciudad Lineal.

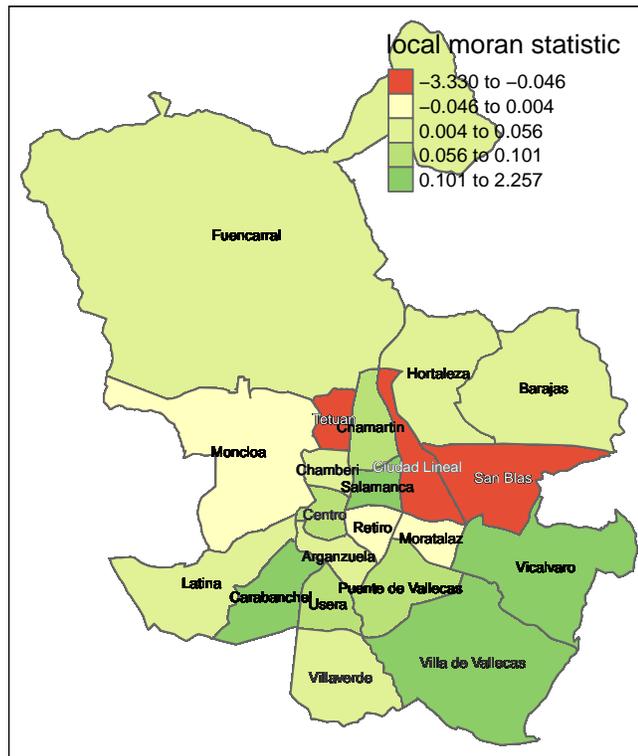


Figura 25: Mapa de significatividad (LISA), en relación con los precios de venta de inmuebles comerciales en la ciudad de Madrid

Valores medios se reparten en la parte media de la ciudad (Moncloa, Chambería, Chamartín, Retiro, Arganzuela, San Blas y Barajas). Vemos un conglomerado de distritos muy poco significativos al sur, conformado por: Carabanchel, Usera, Villaverde, Puente de Vallecas y Moratalaz. Finalmente vemos las centrales Tetuán y Salamanca son de las que menos contribuyen al estadístico global junto con Hortaleza al sur.

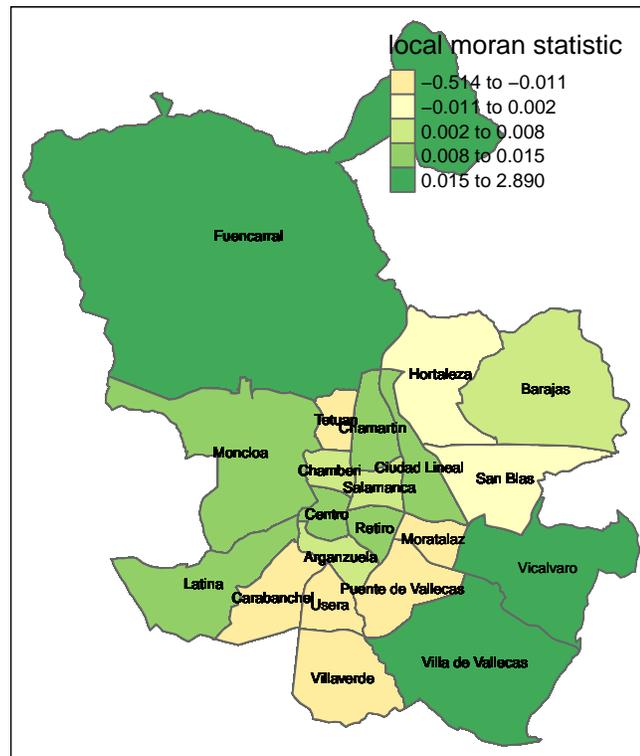


Figura 26: Mapa de significatividad (LISA), en relación con los precios de alquiler de inmuebles comerciales en la ciudad de Madrid

Para confirmar la presencia de asociación a nivel local, se muestra en las Figuras 27 y 28 la formación de clústers tanto intra distritales así como unidades espaciales que presentan patrones atípicos respecto a sus vecinos.

En el mercado de venta (Figura 27) se observa la existencia de diversos clústers espaciales:

- Precios superiores a la media global rodeados de precios que también siguen dicho comportamiento (hot spots), dispuestos en la zona central de la ciudad donde encontramos a Chamberí, Centro, Chamartín y Salamanca.

- Precios inferiores a la media global rodeados de valores similares (cold spots) al sur de la ciudad. Dicho clúster comprime a: Carabanchel, Usera, Arganzuela, Villaverde, Villa de Vallecas, Puente de Vallecas y Vicálvaro.
- San Blas-Canillejas sigue un patrón atípico en cuanto a los precios de venta de locales comerciales. Presenta un comportamiento Low-High, lo cual traduce a que sus valores son inferiores a la media global y se encuentran rodeados por valores opuestos (que superan a la media).

En cuanto a los demás distritos, que presentan un color blanco, podemos decir que sus precios no siguen ningún patrón uniforme, o al menos, no tan significativo como los anteriormente nombrados.

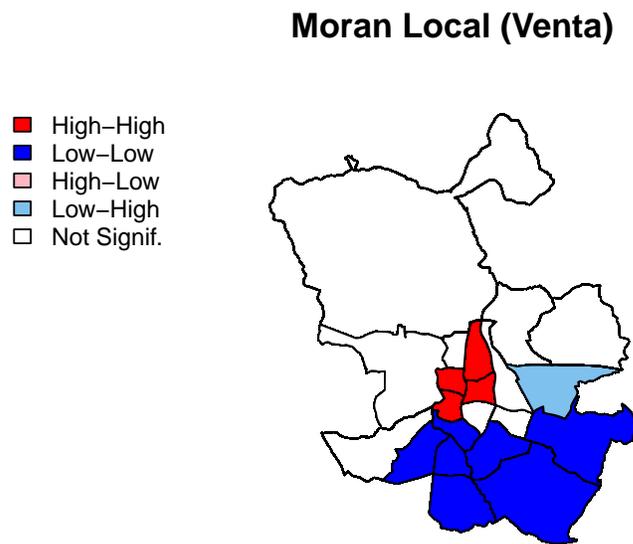


Figura 27: Mapa de clusters (LISA), en relación con los precios de venta de inmuebles comerciales en la ciudad de Madrid

Para el mercado del alquiler, no vemos patrones uniformes que aglomeren a más de un distrito, sino más bien comportamientos atípicos de carácter individual como el «hot spot» en Villa de Vallecas, queriendo decir que los precios dentro de dicha unidad espacial son principalmente altos.

Lo opuesto ocurre en Chamartín, existiendo homogeneidad significativa en los precios, que son bajos.

Por último, Carabanchel representa un punto Low-High, queriendo decir que sus precios son inferiores a la media global y presentan disparidad con los distritos vecinos

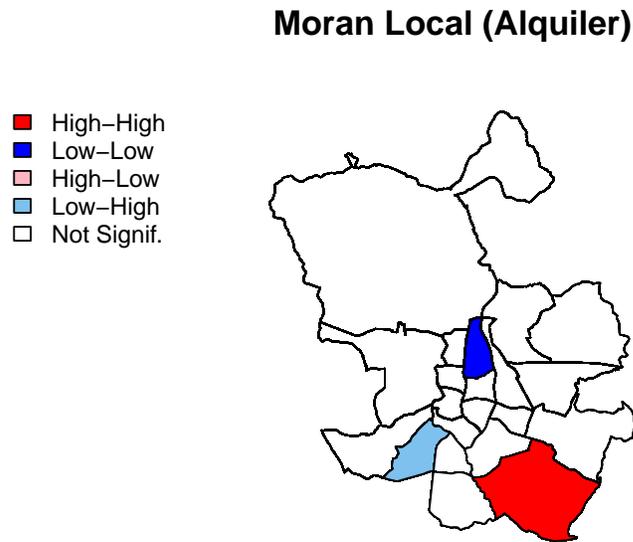


Figura 28: Mapa de clusters (LISA), en relación con los precios de alquiler de inmuebles comerciales en la ciudad de Madrid

Rentabilidades

Analizaremos la rentabilidad anual (en euros) por metro cuadrado, con el objetivo de averiguar dónde se encuentran los mayores retornos por metro cuadrado en la ciudad de Madrid.

Teniendo presente la posición geográfica, veremos los indicadores de rentabilidad de cada distrito, posteriormente, se desglosan los índices en función de los

Cuadro 13: Rentabilidad por metro cuadrado

Distrito	Rentabilidad %
Moratalaz	18.1
Villa de Vallecas	11.0
Barajas	9.9
Usera	9.9
Tetuan	9.7
Villaverde	9.2
Latina	9.0
Puente de Vallecas	9.0
Carabanchel	8.6
San Blas	8.6
Retiro	8.0
Arganzuela	7.2
Centro	7.2
Chamartin	7.1
Fuencarral	6.9
Moncloa	6.8
Ciudad Lineal	6.6
Salamanca	6.1
Hortaleza	6.0
Chamberi	5.6
Vicalvaro	4.2

tipos de locales en cada uno de estos (pequeños, medianos, grandes e industriales). Recordemos que los espacios pequeños son los más numerosos e incluso existen distritos sin locales grandes y/o industriales o con un número insuficiente de propiedades que impide el cómputo de de la rentabilidad.

Tras emplear la fórmula reflejada en el apartado de metodología, disponemos del Cuadro 13 que indica, de manera descendente, los retornos (en euros) por metro cuadrado de cada distrito.

Moratalaz ocupa el primer lugar de la lista con 18,1%, dando lugar a una brecha de más de 10 puntos porcentuales respecto a Villa de Vallecas y las sucesivas áreas. Representando a la zona central, Tetuán ocupa el quinto lugar, con un 9,7%, seguida de Villaverde, perteneciente a la periferia oeste con 9,2%.

Respecto al los retornos más bajos, Vicálvaro aparenta ser el distrito menos rentable con 4,2%, anteriormente a éste, encontramos a distritos centrales como Chamberí (5,6%) y Salamanca (6,1%), así como de la periferia este en el caso de

Hortaleza (6,0%).

A continuación, desglosamos dichos indicadores en función de las dimensiones de los locales, dictados mediante la extensión en metros cuadrados, con el objetivo de proporcionar cifras más acertadas. Adicionalmente, se ha fragmentado el análisis por la disposición geográfica de los distritos:

- Zona central (Cuadro 14)
- Zona media (distritos no centrales ni periféricos) (Cuadro 15)
- Periferia este (Cuadro 16)
- Periferia oeste (Cuadro 17)

Zona central: Tetuán, Chamberí, Centro, Arganzuela, Retiro, Salamanca y Chamartín

Cuadro 14: Rentabilidad por metro cuadrado: Zona central

Distrito	Tamaño			
	Pequeño	Mediano	Grande	Industrial
Arganzuela	0.066	0.058	0.081	0.000
Centro	0.070	0.074	0.053	0.069
Chamartin	0.063	0.072	0.050	0.091
Chamberi	0.056	0.048	0.054	0.000
Retiro	0.074	0.055	0.061	0.055
Salamanca	0.067	0.040	0.075	0.110
Tetuan	0.087	0.118	0.036	0.111

En lo que respecta a los espacios pequeños y medianos, los de mayor rentabilidad se sitúa en Tetuán, con índices de 8,7% y 11,8% respectivamente. El resto de rentabilidades de los locales pequeños se muestran más armónicas, comprendiendo valores entre 5,6% (Chamberí) hasta 7,4% (Retiro). Las superficies medianas reportan datos más dispares, habiendo una gran brecha entre el porcentaje máximo y el mínimo, éste último correspondiendo a Salamanca (4,0%) y muy próximo a ese valor encontramos a Chamberí (4,8%).

En cuanto a los locales grandes, los de Arganzuela aparecen como los más rentables (8,1%), en Tetuán se encuentran los espacios grandes con menos retorno (3,6%). Por último, Tetuán acoge los locales industriales más rentables (11,1%) seguido del distrito de Salamanca con 11,0%.

Sobre los locales industriales, para Arganzuela y Chamberí, no se pueden obtener los índices puesto que dichos distritos no acogen locales de esta tipología.

Zona media: Ciudad Lineal, Moratalaz, Puente de Vallecas y Usera

Cuadro 15: Rentabilidad por metro cuadrado: Zona media

Distrito	Tamaño			
	Pequeño	Mediano	Grande	Industrial
Ciudad Lineal	0.068	0.051	0.051	0.165
Moratalaz	0.060	0.084	0.000	0.000
Puente de Vallecas	0.091	0.078	0.106	0.041
Usera	0.085	0.172	0.060	0.097

Los espacios pequeños ubicados en Puente de Vallecas y Usera reportan los mayores porcentajes, 9,1% y 8,5% respectivamente, mientras que los Moratalaz y Ciudad Lineal mantienen sus índices entre 6,0% y 6,8%.

Sobre los locales medianos, Moratalaz había reportado un retorno exorbitado de 88,9%, a causa de un local cuyo importe mensual en concepto de alquiler era igual a 50,000.00 €, la cifra era tan extrema que distorsionaba la magnitud de la ratio. Es por ello que, se decidió obtener un índice más ajustado a través del reemplazo del outlier con el valor del rango intercuartílico, como resultado se obtuvo un porcentaje de 8,4%. Pese a la modificación, el distrito líder con locales medianos más rentables es Usera con un 17,2%. Dentro de los menos rentables están Puente de Vallecas (7,8%) y Ciudad Lineal (5,1%).

Puente de Vallecas muestra mejor rentabilidad en términos de locales grandes al obtener un 10,6% y con gran ventaja respecto a Usera (6,0%) y Ciudad Lineal (5,1%). La situación opuesta a lo anterior se evidencia a la hora de comparar la productividad económica de los locales industriales, Puente de Vallecas es el menos rentable con un 4,1% siendo rebasado por Usera (9,7%) y Ciudad Lineal (16,5%).

Moratalaz carece de espacios grandes e industriales, consecuentemente, no se puede calcular el ratio de rentabilidad.

Periferia Este: Hortaleza, Barajas, San Blas- Canillejas, Vicálvaro y Villa de Vallecas

Cuadro 16: Rentabilidad por metro cuadrado: Periferia este

Distrito	Tamaño			
	Pequeño	Mediano	Grande	Industrial
Barajas	0.106	0.116	0.000	0.000
Hortaleza	0.070	0.024	0.083	0.041
San Blas	0.099	0.056	0.205	0.085
Vicalvaro	0.046	0.061	0.000	0.095
Villa de Vallecas	0.100	0.102	0.198	0.082

Hay altas tasas de rentabilidad entre locales pequeños, comprendiendo entre 7,0% y 10,6%, siendo Barajas el poseedor del mayor índice, análogamente, aunque Vicálvaro se aleja de ésta tendencia y figura como el distrito con menor retorno por metro cuadrado (4,6%). En el grupo de locales medianos, Barajas presenta más rentabilidad con 11,6% aunque muy cercano se encuentra Villa de Vallecas con 10,2%, Hortaleza es el distrito donde los locales mediados ofrecen el menor retorno.

Respecto a grandes superficies, el número de propiedades es insuficiente para el cómputo del índice en Barajas y Vicálvaro, no obstante, se dispone de los índices de Hortaleza (8,3%) y San Blas-San Blas-Canillejas (2,5%). Para obtener el índice en Villa de Vallecas se ha sustituido un valor extremo, perteneciente al mercado de alquiler, por el valor del rango intercuartílico resultando en un retorno del 19,8%, de lo contrario se habría obtenido un índice desmesurado.

Vicálvaro aparenta tener los espacios industriales más rentables de la zona (9,5%), seguido de San Blas-Canillejas y Villa de Vallecas, ambos con ratios superiores a 8,0%, quedando Hortaleza en la última posición (4,1%), de nuevo, Barajas no cuenta con espacios industriales en venta, por lo que no puede computarse el índice.

Periferia Oeste: Carabanchel, Fuencarral, Latina, Moncloa y Villaverde

Cuadro 17: Rentabilidad por metro cuadrado: Periferia oeste

Distrito	Tamaño			
	Pequeño	Mediano	Grande	Industrial
Carabanchel	0.091	0.080	0.102	0.093
Fuencarral	0.069	0.052	0.055	0.063
Latina	0.088	0.082	0.059	0.000
Moncloa	0.061	0.093	0.038	0.048
Villaverde	0.104	0.079	0.055	0.067

Los índices de los espacios pequeños oscilan entre el 6,1% y 10,4%, siendo Villaverde líder del grupo mientras Moncloa se posiciona de última. En la tipología de superficies medianas, Moncloa presenta la tasa más elevada (9,3%) y Fuencarral la más baja (5,25%), mientras que el resto de la zona presenta índices altos entre 7,9% y 8,2%.

Hay un descenso en los porcentaje de locales grandes, la mayoría oscila entre 3,8% y 5,9%, aunque Carabanchel destaca con un 10,2%, al igual que con su alto índice para locales industriales de 9,3%. Los índices de locales industriales del resto de distritos parten desde 4,8% en caso de Moncloa, seguido de Fuencarral (6,3%) y Villaverde (6,7%). Para el distrito Latina, no se disponen de locales en alquiler, por lo que no se puede llevar a cabo el cálculo de la rentabilidad.

Conclusiones

Este trabajo se ha llevado a cabo con datos reales de locales en alquiler y venta de la ciudad de Madrid y tras el empleo de herramientas disponibles del entorno de programación R, ha sido posible conocer la estructura por tipologías, cuotas de ofertas por distrito, distribución de precios en el territorio y demás características esenciales para poder brindar un apto análisis de mercado que se alinea con los objetivos planteados. Algunos aspectos que destacan son:

- Más del 50% de locales que conforman ambas muestras tienen una extensión de hasta $150 m^2$.
- Centro es el distrito con más cantidad de inmuebles comerciales ofertados, con una cuota conjunta de 20.37%, siendo mayor la cuota de alquiler con 13.19%.
- La oferta de locales en alquiler se concentra fuertemente en el área central de la ciudad (32.04% del total), específicamente, en Centro, Chamberí y Salamanca, siendo estos tres distritos los mayores poseedores de locales pequeños, medianos y grandes.
- Carabanchel es el distrito con más locales en venta (9.05%) seguido de Tetuán (8.05%)
- En términos de rentabilidad por metro cuadrado en función de las zonas:
 - Zona central: La tipología con mayor retorno por metro cuadrado es la «pequeña» con 6.90% y el distrito más rentable es Tetuán con un retorno medio por metro cuadrado de 8.8% (calculado a partir de todos sus índices de rentabilidad por metro cuadrado).
 - Zona media: La tipología con mayor retorno por metro cuadrado es la «mediana» con 9.63% y el distrito más rentable es Usera con un retorno medio por metro cuadrado de 10.35% (calculado a partir de todos sus índices de rentabilidad por metro cuadrado).
 - Periferia este: La tipología con mayor retorno por metro cuadrado es la «grande» con 9.60% y el distrito más rentable es Villa de Vallecas con un retorno medio por metro cuadrado de 11.9% (calculado a partir de todos sus índices de rentabilidad por metro cuadrado y sin tomar en cuenta los valores extremos).
 - Periferia oeste: La tipología con mayor retorno por metro cuadrado es la «pequeña» con 8.26% y el distrito más rentable es Carabanchel con un retorno medio por metro cuadrado de 9.15% (calculado a partir de todos sus índices de rentabilidad por metro cuadrado).

Los pasos tomados para la obtención de gráficas, cuadros y estadísticos son reproducibles y válidos para datos de diferentes periodos.

Además, las herramientas de autocorrelación, que no sólo agrupan observaciones similares sino que comunican el grado de correlación, sirvieron de mucha ayuda a la hora de hallar patrones que no se aprecian a simple vista, lo cual explica su notoria aplicación en campos como la ecología y demografía.

Tras el cálculo del índice global de Moran, obtuvimos un estadístico muy próximo a 0, pero con un *p-valor* muy significativo, lo cual nos condujo a realizar un análisis a escala más reducida con el objetivo de descubrir patrones subyacentes que no son tan fáciles de detectar a simple vista con el análisis global.

Con el análisis local de los precios de venta, identificamos patrones significativos en los precios del año 2020. Los precios más elevados se distribuyeron uniformemente en el centro de la ciudad, más concretamente en Chamberí, Centro, Chamartín y Salamanca. Por otro lado, los precios más bajos se aglomeraron al sur de la ciudad.

Siguiendo con el mercado de venta, fue posible la detección de un outlier espacial: San Blas-Canillejas, el cuál contiene precios heterogéneos. El resto de distritos carece de signos de autocorrelación con áreas vecinas.

Pasando al mercado de alquiler de locales comerciales, identificamos comportamiento individuales significativos, como el «hot spot» en Villa de Vallecas, donde los alquileres seguían una tendencia al alza, un «cold spot» en Chamartín y determinamos que Carabanchel es un outlier espacial con disparidad significativa en sus precios respecto a los distritos próximos.

Cabe señalar que, tras contrastar los índices locales, el mercado de alquiler es el que más presentó aleatoriedad en sus precios en comparación al mercado de venta.

Recordemos que el período de tiempo en que recaen los datos estudiados fue sumamente excepcional, con marcadas repercusiones no sólo en la oferta comercial de Madrid, también a escala nacional. Los descensos de índices de ventas, sobre todo en el comercio minorista y de la ocupación evolucionaron de forma distinta en los distritos observados, eso explica las diferencias entre los análisis locales de ambos mercados.

A pesar de la utilidad indudable de nuestros resultados, este estudio tiene limitaciones que tienen que ver con el periodo utilizado, al ser un estudio cross-section, cabría esperar un análisis continuado de datos que nos permitiera seguir validando nuestros resultados.

Anexo

OBJETIVOS DE DESARROLLO SOSTENIBLE



Reflexión sobre la relación del TFG con los ODS en general y con el/los ODS más relacionados.

Tras la realización del Trabajo de Fin de Grado coincido en que la Inteligencia de Negocios, en inglés Business Intelligence (BI), sirve como puente entre gran variedad de sistemas permitiendo un enfoque basado en los datos para la toma de decisiones.

Por Inteligencia de Negocios, nos referimos al conjunto de programas, herramientas y procesos utilizados para analizar información con el fin de suministrar aportes que puedan ayudar a una organización a tomar mejores decisiones. La metodología BI posee una estructura óptima para una visión conjunta de las gestiones y actividades que tradicionalmente eran controladas de forma individual, mediante los ERP (Enterprise Resource Planning) y los CRM (Customer Relationship Management).

Considerando la creciente necesidad de simplificar y unificar el acceso a la información, una de las ventajas de esta rama es que la información esté organizada, dotando de agilidad y objetividad la toma de decisiones. Como contrapartida, el hecho que los datos estén centralizados permite una optimización de recursos. Otra garantía de BI es la presentación de información de forma atractiva para que se pueda entender con facilidad. En función de los criterios de cada compañía y/o sector, la información también puede personalizarse. Como consecuencia de todo lo anterior, la comunicación entre los distintos departamentos de la organización mejora notablemente y aumenta su productividad.

Igualmente, el BI es de gran ayuda para otras actividades como la eficiencia del stock, la fidelización de clientes, la atracción de nuevos usuarios o la posibilidad de evitar errores.

Esencialmente, los procesos llevados a cabo en este TFG equivalen a las primeras etapas del BI, la recolección de la información, la cual se espera que sea relevante y correcta. Para empezar, los datos en bruto no pueden utilizarse directamente para la toma de decisiones. Como tal, la inteligencia de datos permite un adecuado procesamiento con herramientas de extracción y métodos analíticos

apropiados para transformar los datos en información y conocimiento listos para que actúen los tomadores de decisiones.

Los datos pueden ingerirse y entregarse a diversas fuentes y destinos. El tipo de datos que impulsan las plataformas de BI puede variar desde informes en papel hasta conversaciones telefónicas, chats por correo electrónico, archivos almacenados en una computadora o archivos de audio/video. Siguiendo una canalización de datos, las empresas realizan operaciones de extracción, transformación y carga para procesar datos estructurados, no estructurados y semiestructurados y luego transferirlos a almacenes de datos en la nube.

Agradezco enormemente la dedicación y esfuerzos de todos los docentes que lograron materializar su apuesta por la Inteligencia de Negocios en la facultad, para así dotar a jóvenes de las aptitudes que se demandan fuertemente en el mercado laboral actual en una extensa gama de sectores, igualmente proporcionándoles una base sobre la que pueden continuar su aprendizaje por sus propios medios a través de material y recursos electrónicos. Lo anterior ha sido una acción fundamental para el fomento del **Objetivo de Desarrollo Sostenible N° 8 de las Naciones Unidas** en relación con el **trabajo decente y el desarrollo económico**.

Las competencias derivadas de los mecanismos BI ofrecen una gran oportunidad de transición a un empleo ya que son más las empresas que se apresuran a convertirse en data-driven. La necesidad de trabajadores capaces de recolectar, analizar y visualizar grandes cantidades de información continúa creciendo exponencialmente.

Este crecimiento será impulsado por factores como aumento en la infraestructura y los servicios de informática en la nube; el aumento de las inversiones en actividades de investigación y desarrollo relacionadas con la inteligencia artificial y el aprendizaje automático; y la creciente demanda de científicos de datos en varios sectores verticales de la industria, como servicios financieros, venta minorista y bienes de consumo, fabricación y logística.

Referencias

L. Anselin (1995) Local Indicators of Spatial Association—LISA *Geographical analysis*, 27(2)93-115. doi:[10.1111/j.1538-4632.1995.tb00338.x](https://doi.org/10.1111/j.1538-4632.1995.tb00338.x)

J. Barnier, F. Briatte, y J. Larmarange (2020) *questionr: Functions to Make Surveys Processing Easier*. Obtenido de <https://CRAN.R-project.org/package=questionr>

R. S. Bivand, E. Pebesma, y V. Gomez-Rubio (2013) *Applied spatial data analysis with R, Second edition*. Springer, NY. Obtenido de <https://asdar-book.org/>

J. M. Brotons Martínez (2017) *Supuestos de Valoración de Inversiones*. Universidad Miguel Hernández.

S. Buttrey y L. R. Whitaker (2017) *A data scientist's guide to acquiring, cleaning and managing data in R* (1st edition.). Wiley.

A. D. Cliff, J. K. Ord, y A. J. Scott (1969) *London Papers in Regional Science* (pp. 25-55). Pion.

C. de Comercio de Madrid (2021) La Rentabilidad Económica. ¿Qué es y para qué sirve? www.mba-madrid.com. Obtenido de <https://www.mba-madrid.com/economia/rentabilidad-economica/>

D. G. de Comercio y Consumo (2022) El comercio de Madrid en cifras. Available at <https://www.comunidad.madrid/inversion/comercio/actividad-comercial-madrid> (2022/05/21).

R. Datar y H. Garg (2019) *Hands-on exploratory data analysis with R: become an expert in exploratory data analysis using R packages* (1st edition.). Packt Publishing.

M. F. Goodchild (1986) *Spatial Autocorrelation* (1st edition.). Geo Books, Norwich.

M. F. Goodchild (2009) What Problem? Spatial Autocorrelation and Geographic Information Science *Geographical analysis*, 41(4)411-417. doi:[10.1111/j.1538-4632.2009.00769.x](https://doi.org/10.1111/j.1538-4632.2009.00769.x)

P. Goovaerts (1997) *Geostatistics for natural resources evaluation /*. Oxford University Press,

G. H. Laursen y J. Thorlund (2010) *Business Analytics at the Analytical Le-*

vel *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting*, 93-135.

N. Mantel (1967) The Detection of Disease Clustering and a Generalized Regression Approach *Cancer research (Chicago, Ill.)*, 27(2)209-220.

D. M. Mark y F. Csillag (1989) The Nature of Boundaries on «Area-Class» Maps *Cartographica.*, 2665-77.

P. A. Moran (1948) The interpretation of statistical maps *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2)243-251.

J. Neyman (1934) On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection *J. R. Stat. Soc.*, 97558-625.

E. Pebesma (2018) Simple Features for R: Standardized Support for Spatial Vector Data *The R Journal*, 10(1)439-446. doi:[10.32614/RJ-2018-009](https://doi.org/10.32614/RJ-2018-009)

Redacción Idealista (2021, julio) La rentabilidad de la inversión en vivienda sufre una significativa caída tras la pandemia. Idealista. Obtenido de <https://www.idealista.com/news/inmobiliario/vivienda/2021/07/05/791270-la-rentabilidad-de-la-inversion-en-vivienda-sufre-una-significativa-caida-tras-la>

R. Rummel y P. Teunissen (1988) Height datum definition, height datum connection and the role of the geodetic boundary value problem *Bulletin géodésique*, 62(4)477-498.

M. Souris (2019) *Epidemiology and geography : principles, methods and tools of spatial analysis*. ISTE.

R. Team (2021) *RStudio: Integrated Development Environment for R*. Boston, MA RStudio, Inc. Obtenido de <http://www.rstudio.com/>

R. Team (2022) *R: What is R?* Vienna, Austria R Foundation for Statistical Computing. Obtenido de <https://www.r-project.org/about.html/>

M. Tennekes (2021) *tmertools: Thematic Map Tools*. Obtenido de <https://CRAN.R-project.org/package=tmertools>

W. R. Tobler (1970) A Computer Movie Simulating Urban Growth in the Detroit Region *Economic geography.*, 46234-40.

H. Wickham (2014) Tidy Data, 591-23. doi:[10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10)

H. Wickham (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Obtenido de <https://ggplot2.tidyverse.org>

H. Wickham, R. François, L. Henry, y K. Müller (2022) *dplyr: A Grammar of Data Manipulation*. Obtenido de <https://CRAN.R-project.org/package=dplyr>