Trabajo Fin de Máster

Análisis de datos agregados por área Contaminación atmosférica y desigualdad socioeconómica

Diana Cristina Castro Armenta

Máster en Técnicas Estadísticas

Curso 2025-2026

Propuesta de Trabajo Fin de Máster

Título en galego: Análise de datos agregados por área. Contaminación atmosférica e desigualdade socioeconómica

Título en español: Análisis de datos agregados por área. Contaminación atmosférica y desigualdad socioeconómica

English title: Areal data analysis. Air pollution and socioeconomic inequality

Modalidad: Modalidad A

Autora: Diana Cristina Castro Armenta, Universidad de Santiago de Compostela

Director/a: María Xosé Rodríguez Álvarez, Universidad de Vigo ; Tomás Raimundo Cotos Yáñez, Universidad de Vigo

Breve resumen del trabajo:

La contaminación del aire es uno de los mayores riesgos ambientales para la salud humana a escala global. Según análisis de la Agencia Ambiental Europea (EEA), en 2022 la exposición a NO_2 y $PM_{2,5}$ fue responsable de un elevado número de muertes prematuras en Europa, asociadas principalmente a enfermedades cardiovasculares y respiratorias atribuibles a estos contaminantes. Desde la perspectiva de la justicia ambiental, resulta relevante estudiar si la exposición a la contaminación atmosférica se distribuye de forma desigual entre la población, especialmente en relación con el nivel socioeconómico de las áreas de residencia. Este trabajo busca aplicar herramientas de análisis de datos agregados por área para estudiar esta cuestión en el contexto de la comunidad autónoma de Galicia. En particular, se analizará la posible asociación entre los niveles de NO_2 y $PM_{2,5}$, y la renta per cápita media municipal, con el objetivo de valorar si existen patrones territoriales compatibles con una distribución desigual de la exposición a estos contaminantes según el nivel de renta. Para ello, se emplearán técnicas descriptivas, cartográficas y estadísticas adecuadas para datos agregados por área, teniendo en cuenta la estructura espacial de la información disponible. Los resultados podrán contribuir a una mejor caracterización de posibles desigualdades socioambientales en Galicia.

Recomendaciones:

Se recomienda haber cursado la materia Estadística Espacial y tener buen conocimiento del lenguaje de programación R.

Otras observaciones:

Doña María Xosé Rodríguez Álvarez, Ramón y Cajal Fellow de la Universidad de Vigo y don Tomás Raimundo Cotos Yáñez, Profesor Contratado Doctor de la Universidad de Vigo, informan que el Trabajo Fin de Máster titulado

Análisis de datos agregados por área: contaminación atmosférica y desigualdad socioeconómica

fue realizado bajo su dirección por doña Diana Cristina Castro Armenta para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal. Además, doña María Xosé Rodríguez Álvarez, don Tomás Raimundo Cotos Yáñez y doña Diana Cristina Castro Armenta

√ sí □ no

autorizan a la publicación de la memoria en el repositorio de acceso público asociado al Máster en Técnicas Estadísticas.

En Santiago de Compostela, a 17 de junio de 2026.

La directora:
Doña María Xosé Rodríguez Álvarez

RODRIGUEZ ALVAREZ MARIA JOSE - 35567986G
Digitally signed by :RODRIGUEZ ALVAREZ MARIA JOSE - 35567986G
Reason:
Date: 2026/07/17 13:35:59 +02'00'

El director:
Don Tomás Raimundo Cotos Yáñez

Firmado por COTOS YAÑEZ, TOMAS RAIMUNDO (FIRMA) el día 17/07/2026 con un

La autora:
Doña Diana Cristina Castro Armenta

Firmado por CASTRO ARMENTA DIANA CRISTINA - ****0169* el día 17/07/2026 con un certificado emitido por AC

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **la autora declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,... sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

En primer lugar, deseo expresar mi gratitud a mi familia y, en especial, a las mujeres que, a lo largo de varias generaciones, me han enseñado a superar cada obstáculo, incluso cuando se pensaba haber llegado al límite. Abuelas, madre y pseudomadres, hermanas, siempre les estaré agradecida.

También quisiera expresar mi gratitud a mis profesores de la Universidad de Santiago de Compostela, que a lo largo de este máster me han transmitido tantos conocimientos. Puede que no siempre reciban el reconocimiento que merecen, pero su labor docente nunca es en vano. Y extiendo un agradecimiento muy especial a la Universidad de Santiago de Compostela y la dirección del máster por darme la oportunidad de realizar este máster y este trabajo.

Por último, deseo dar un agradecimiento muy afectuoso a mi pareja, por el apoyo que me ofreces cada día. Sin ti, no habría llegado hasta aquí.

Índice general

Resumen

Resumen en español

Este trabajo aplicó estadística espacial a datos agregados por área con el fin de caracterizar la distribución de NO_2 y $PM_{2,5}$ en los 313 municipios de Galicia y evaluar su relación con la renta bruta per cápita, analizando posibles patrones de desigualdad en la carga ambiental.

El análisis exploratorio (ESDA) mostró autocorrelación espacial positiva en las tres variables con clústeres de alta contaminación en áreas metropolitanas y baja contaminación en el zonas rurales. El análisis bivariado reveló una asociación positiva entre los contaminantes y los retardos espaciales de la renta. Sin embargo, el análisis local identificó un grupo de nueve municipios con contaminación elevada y un entorno socioeconómicamente desfavorecido.

Los modelos espaciales Besag-York-Mollié (BYM) confirmaron la asociación positiva entre renta y contaminación, con coeficientes de renta positivos y significativos para ambos contaminantes. El diagnóstico de la componente no estructurada evidenció ausencia de autocorrelación espacial, y la comparación mediante DIC y WAIC respaldó la especificación espacial, validando el modelo BYM.

En conjunto, los niveles de NO_2 y $PM_{2,5}$ se mantienen por debajo de los límites recomendados por la Organización Mundial de la Salud, por lo que no se derivan implicaciones inmediatas para políticas correctivas ligadas a la Agenda 2030.

English abstract

This thesis applied spatial statistics to areal data to characterize the distribution of NO_2 and $PM_{2,5}$ across Galicia's 313 municipalities and to assess their relationship with gross income per capita, examining possible patterns of socioeconomic inequality with a high environmental burden.

Exploratory spatial data analysis (ESDA) revealed positive spatial autocorrelation in all three variables, with high-pollution clusters in metropolitan areas and low-pollution clusters in rural areas. Bivariate analysis uncovered a positive association between pollutants and the spatial lags of income. However, local-scale analysis identified a group of nine municipalities with high pollution and a socioeconomically vulnerable setting.

Besag-York-Mollié (BYM) spatial models confirmed the positive association between income and pollution, with positive and significant income coefficients for both pollutants. Diagnostics on the unstructured component showed an absence of spatial autocorrelation, and comparison via DIC and WAIC supported the spatial specification, validating the BYM model.

Overall, NO_2 and $PM_{2,5}$ levels remain below the limits recommended by the World Health Organization, so no immediate implications for corrective policy linked to the 2030 Agenda follow from this work.

Prefacio

A través de los años, con diferentes propuestas e inumerables acciones, la humanidad ha buscado mejorar las condiciones ambientales e intentar llegar a una convivencia sostenible con el ambiente, entendiéndose por desarrollo sostenible al equilibrio entre el desarrollo económico y social que respeta los ecosistemas naturales, como se menciona en el reporte de Brundtland de la Organización de las Naciones Unidas (ONU):

> *...satisfacer las necesidades de las generaciones presentes sin comprometer la capacidad de las generaciones futuras para satisfacer sus propias necesidades (ONU, 1987).*

Desde la Cumbre de la Tierra en 1992, los países miembros de las Naciones Unidas han buscado crear acuerdos para alcanzar un desarrollo sostenible a través de diferentes mecanismos como convenciones, agendas, asambleas, entre otros. En septiembre del 2015 se estableció la Agenda 2030 para el Desarrollo Sostenible proporcionando un plan global a seguir por todos sus miembros y hacer frente a los retos globales, no solo en cuestión de ambiente y cuidado del planeta, sino para las personas y su prosperidad. Esta agenda consiste en 17 Objetivos de Desarrollo Sostenible (ODS o SDG por sus siglas en inglés), objetivos que buscan de manera general terminar con la pobreza, reducir la desigualdad e impulsar el crecimiento económico al mismo tiempo que se preservan recursos naturales y se mitiga el cambio climático (Desarrollo-Sostenible, 2023).

Este trabajo nació de una inquietud doble: por un lado, la voluntad de aplicar los métodos estadísticos aprendidos durante el Máster en Técnicas Estadísticas a un problema con impacto social directo aplicando los ODS; por otro, la constatación de que la desigualdad ambiental rara vez se analiza con rigor cuantitativo fuera de las grandes metrópolis.

La marcada dualidad de Galicia entre áreas metropolitanas y extensas zonas rurales fue lo que hizo atractivo trabajar con esta región. Junto con la disponibilidad de datos municipales de calidad y la heterogeneidad de los 313 municipios, la región me ofreció las condiciones idóneas para explorar técnicas de estadística espacial de datos agregados en un contexto real y retador, que enriquecieron considerablemente mi aprendizaje.

Este trabajo se estructura en dos partes. La primera presenta la introducción y objetivos en el capítulo primero, y después, en los capítulos dos y tres se desarrollan los fundamentos teóricos y metodológicos del análisis de datos agregados por área, desde la exploración espacial descriptiva hasta las técnicas de modelado. La segunda parte aplica esos métodos a los datos de contaminación atmosférica y renta municipal de Galicia, con el objetivo de valorar la existencia de patrones compatibles con una distribución desigual.

Capítulo 1

Introducción

La calidad del aire es uno de los principales determinantes ambientales de la salud humana. Según estimaciones de la Agencia Europea de Medio Ambiente, en el 2022, la exposición prolongada a dióxido de nitrógeno (NO_2) y a partículas finas ($PM_{2,5}$) fue responsable de aproximadamente 66000 y 269000 muertes prematuras, respectivamente (Soares et al., 2024). A pesar de décadas de regulación ambiental, la distribución de estos contaminantes en el territorio Europeo no es homogénea: su concentración depende de factores geográficos, industriales y de movilidad que varían notablemente entre regiones, y cuya caracterización espacial resulta imprescindible para el diseño de políticas públicas eficaces.

Dentro de este contexto, la estadística espacial para datos agregados por área ofrece herramientas que permiten identificar clústeres de exposición elevada. Mediante el modelado de la dependencia espacial entre unidades geográficas, como municipios o secciones censales, es posible identificar estos clústeres y cuantificar la asociación entre la concentración del contaminante y variables socioeconómicas. Este enfoque constituye la base metodológica para este trabajo.

1.1. Motivación

La Agenda 2030 para el Desarrollo Sostenible, adoptada por las Naciones Unidas en 2015, establece un marco global de 17 Objetivos de Desarrollo Sostenible (ODS). Dentro de estos 17 objetivos, se incorporan objetivos que buscan erradicar la pobreza, reducir la desigualdad y preservar el medio ambiente (Desarrollo-Sostenible, 2023). Más allá de los objetivos ambientales, la Agenda 2030 incorpora una dimensión de equidad que resulta central para este trabajo. El ODS 10 busca reducir las desigualdades dentro de los países y entre ellos, mientras que el ODS 11 promueve ciudades y comunidades inclusivas, seguras y resilientes. Ambos objetivos convergen en el concepto de *justicia ambiental*: la idea de que los beneficios de un entorno saludable y los riesgos derivados de la contaminación deben distribuirse de forma equitativa en la sociedad, sin que los grupos más vulnerables soporten una carga desproporcionada (Walker, 2012). Esta problemática lleva a preguntas como: ¿hay comunidades más vulnerables que otras a riesgos de incendios, inundaciones o impactos del cambio climático? ¿Pueden todas las personas tener acceso a áreas verdes o solo algunas? ¿Está involucrada la población en decisiones ambientales o toman las decisiones solo unas pocas personas en el poder? Este tipo de cuestiones forma parte del concepto complejo que es la justicia ambiental, que por su importancia moderna tiene un alcance extenso en diferentes industrias y contextos. Sin embargo, para el propósito de este trabajo, solo es necesario comprender la faceta de distribución equitativa de la carga ambiental en la sociedad.

1.2. Justificación

En la actualidad, la evidencia disponible sugiere que la distribución equitativa de la carga ambiental está lejos de ser la norma. Estudios realizados en grandes metrópolis españolas, como Madrid (Hewitt

et al., 2024) y Barcelona (Moreno-Jimenez et al., 2016), han documentado asociaciones espaciales entre la exposición a contaminantes atmosféricos y el nivel socioeconómico de los barrios, señalando que las poblaciones más vulnerables tienden a residir en las zonas de mayor contaminación. Por otro lado, este tipo de análisis son escasos para regiones de menor tamaño o de carácter mixto (urbano-rural), donde los patrones de exposición y capacidades institucionales difieren de las grandes ciudades. El territorio de Galicia, comunidad autónoma del noroeste de España, combina distintos tipos de áreas con extensas zonas rurales y metrópolis costeras constituyendo un caso de estudio de especial interés. Los 313 municipios de la región presentan una gran heterogeneidad demográfica y socioeconómica que ofrecen una gran diversidad. Esta diversidad representa un escenario idóneo para explorar si la distribución espacial de los contaminantes está asociada a la desigualdad socioeconómica municipal, y para evaluar en qué medida las herramientas de análisis de datos de área pueden aportar evidencia útil para la toma de decisiones en materia de política ambiental. Esto considerando, que la relación esperada de sectores socioeconómicamente vulnerables con una carga ambiental alta, puede no cumplirse en el contexto mixto (rural-urbano) que presenta Galicia.

1.3. Tipos de datos espaciales

Para llevar a cabo un análisis de datos espaciales, es importante entender qué son los datos espaciales, qué representan y qué tipos de datos existen. El análisis estadístico de fenómenos georreferenciados parte de un marco común: un proceso espacial representado como una colección de variables aleatorias:

$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^2\}, \tag{1.1}$$

donde D es la región de estudio y $Z(\mathbf{s})$ la variable aleatoria en la localización $\mathbf{s}$. Las características de la región de estudio D y la forma en que las observaciones se asocian a localizaciones determinan el tipo de datos espaciales con los que se trabaja. La clasificación clásica distingue tres tipos (Cressie, 1993; Moraga, 2023):

Datos geoestadísticos. Este tipo de datos se presenta cuando el dominio D es continuo y fijo: la variable podría observarse en cualquier punto del espacio, pero solo se dispone de mediciones en un conjunto finito de localizaciones. El objetivo habitual al trabajar con datos geoestadísticos, es predecir el valor del proceso en localizaciones no muestreadas mediante técnicas de interpolación como el *kriging*. Ejemplos típicos de estos datos son la temperatura superficial del suelo o la precipitación acumulada en una cuenca.

Datos puntuales. Al hablar de datos puntuales, lo que se modela no es el valor de una variable, sino la propia localización aleatoria de los eventos dentro de un dominio fijo. El interés recae en determinar si los eventos se agrupan en clústeres, se dispersan en un patrón regular o siguen un patrón aleatorio. Estos datos hacen posible analizar si la variable presenta algún patrón espacial. Ejemplos de datos puntuales son la distribución de incendios forestales o la localización de casos de una enfermedad infecciosa.

Datos agregados por área. En este tipo de datos, la región de estudio se divide en un número finito de áreas o polígonos $\{A_1, \ldots, A_n\}$, usualmente unidades administrativas, y la variable de interés se observa como un único valor agregado por polígono. Este tipo de datos puede utilizarse para representar la cantidad de individuos con una enfermedad (recuentos) agrupados por municipios o regiones censales, por ejemplo. La forma más habitual de representar los datos de área es a través de mapas coropléticos, coloreando cada polígono de acuerdo con los valores de la variable. Es el tipo más frecuente en epidemiología y constituye la base metodológica de este trabajo. Su tratamiento en profundidad se desarrolla en el Capítulo 2.

1.4. Objetivos

La distribución equitativa de la carga ambiental puede abordarse desde un análisis espacial con datos agregados por área y una metodología que permita transformar los compromisos de los ODS en políticas efectivas. Es por ello que el **objetivo general** de este trabajo es: aplicar técnicas de estadística espacial para datos agregados por área con el fin de caracterizar la distribución espacial de la contaminación atmosférica en Galicia y evaluar su relación con la desigualdad socioeconómica municipal. De forma específica, se persiguen los siguientes objetivos:

1. Presentar los fundamentos teóricos del análisis de datos agregados por área, desde la definición de vecindario y la matriz de pesos espaciales hasta los modelos bayesianos de suavizado areal.
2. Realizar un análisis exploratorio de datos espaciales (ESDA) de las concentraciones de NO_2 y $PM_{2,5}$, y de la renta per cápita en los 313 municipios gallegos para datos del 2023, identificando clústeres y patrones de autocorrelación espacial.
3. Ajustar modelos espaciales areales y compararlos con modelos de referencia no espaciales para seleccionar la especificación más adecuada.
4. Contrastar la hipótesis de justicia ambiental evaluando la asociación espacial entre la exposición a contaminantes y la renta bruta per cápita municipal utilizando mapas BiLISA.
5. Proporcionar una base metodológica y empírica que contribuya al seguimiento de los ODS en el contexto gallego.

1.5. Estructura del trabajo

El presente trabajo se organiza en dos partes. La **Parte I** desarrolla los fundamentos metodológicos necesarios para el análisis. El Capítulo 2 introduce los conceptos teóricos del análisis de datos agregados por área: la definición formal del proceso areal, los problemas metodológicos inherentes a la agregación, la construcción del vecindario y la matriz de pesos espaciales, los métodos del análisis exploratorio y los modelos espaciales bayesianos.

Por otro lado, la **Parte II** aplica esos métodos al caso de estudio. El Capítulo 3 describe el preprocesado de datos, la construcción del vecindario para los 313 municipios gallegos. El Capítulo 4 el análisis exploratorio completo, incluyendo los mapas coropléticos, la autocorrelación espacial global y local mediante el índice I de Moran, el análisis bivariado mediante BiLISA y presenta los modelos de referencia no espaciales y los modelos bayesianos con componente espacial BYM. Finalmente, el Capítulo 5 recoge las conclusiones del trabajo, sus limitaciones y las posibles líneas de investigación futuras.

Parte I

Metodología

Capítulo 2

Cuestiones y fundamentos teóricos

Como se introdujo en la Sección 1.3, los datos de área surgen cuando la región de estudio se divide en un número finito de polígonos y la variable de interés se agrega a nivel de cada unidad. Este capítulo desarrolla en profundidad los conceptos y herramientas específicos de este tipo de datos. Para ello, en este capítulo se consideran como referencia general Cressie (1993) y Moraga (2023).

El capítulo se estructura en las siguientes secciones: la introducción de los conceptos teóricos del análisis de datos agregados por área; la definición formal del proceso areal; los desafíos metodológicos inherentes a dicha agregación; la elaboración del vecindario y la matriz de pesos espaciales; el análisis exploratorio y los modelos bayesianos espaciales.

2.1. Datos agregados por área

Los datos agregados por área, areales, *lattice*, o reticulares surgen cuando la región de estudio D se divide en un número finito de subregiones o polígonos $\{A_1, \ldots, A_n\}$ no solapados, y la variable de interés se observa como un único valor agregado por polígono o unidad. Así, el **proceso areal** se puede representar como:

$$Z(A_1) \ldots, Z(A_n) \tag{2.1}$$

donde el área i-ésima se denota por A_i y $Z(A_i)$ es la variable aleatoria que puede representar un conteo, una proporción, una media, una tasa u otras medidas asociadas a dicha área. En los datos areales, el índice i es discreto y solo toma n valores posibles, uno por unidad o polígono. De manera equivalente, en (2.1) se puede escribir como $Z_i \equiv Z(A_i)$, resumiendo el proceso areal como el vector $\mathbf{Z} = (Z_1, \ldots, Z_n)^\top$. Por otro lado, el valor de la observación en el área A_i se denota como $z(A_i)$ y el conjunto de datos observados como:

$$z(A_1) \ldots, z(A_n) \tag{2.2}$$

En ambas notaciones A_i sirve como identificador del polígono o área i. Frecuentemente, se toma como identificador al centroide del área o capital. Al igual que el proceso estocástico, las observaciones de datos agregados pueden representarse equivalentemente como $z_i \equiv z(A_i)$ y resumirse en el vector $\mathbf{z} = (z_1, \ldots, z_n)^\top$.

De esta manera, la variación espacial de la variable de interés se estudia a través de las relaciones entre áreas, introduciendo a la noción de **vecindad** como el elemento central del análisis de datos areales. Los límites de las áreas suelen ser impuestos por divisiones administrativas como municipios, provincias, secciones censales, entre otros; aunque también, pueden ser rejillas regulares o unidades definidas por el investigador. Este tipo de datos es especialmente frecuente en epidemiología y ciencias ambientales, donde la información se agrega por razones de disponibilidad. Ejemplos representativos incluyen la tasa de desempleo por provincia o la incidencia de una enfermedad por municipio.

2.1.1. Problemas metodológicos inherentes a la agregación de datos

Al trabajar con datos agregados se introduce una serie de problemas que el analista debe identificar antes de abordar cualquier método. A continuación, se describen esas consideraciones, así como, recomendaciones y métodos para prevenirlas o mitigarlas.

Problema de la unidad areal modificable

El *modifiable area unit problem* (MAUP) surge cuando los resultados del análisis pueden variar sustancialmente al cambiar la unidad espacial de análisis, ya sea por cambio de escala de agregación (**efecto escala**) o por redefinición de los límites de los polígonos (**efecto zonificación**). Por ejemplo, el patrón de concentración de contaminantes observado a nivel de provincia puede diferir del observado a nivel municipal (escala) o el precio de vivienda puede variar debido a la delimitación del barrio o zona censal (zonificación). Por este motivo, las conclusiones obtenidas con una determinada unidad espacial deben interpretarse en ese marco, evitando hacer interpretaciones muy específicas debido al MAUP.

Problemas por diferencias de exposición

Esta consideración surge de estudios llevados a cabo a nivel poblacional para investigar la relación de factores de exposición y resultados de salud. Denominados **estudios ecológicos**, se enfocan en analizar datos agregados para poblaciones o grupos, resultando en asociaciones que pueden ser ciertas al nivel grupal o de la población, pero no ser verdaderas para los individuos del grupo. Esta situación da lugar a un sesgo que se considera un caso especial del MAUP ya que es similar al efecto de agregación. Para mitigarlo existen metodologías de desagregación como el modelo de fusión Bayesiano de Moraga et al. (2017).

Efecto de áreas pequeñas

En unidades con población de referencia reducida, un único evento puede producir tasas extremas que distorsionan el patrón espacial observado. Este problema, denominado *small area problem* (SAP), se mitiga habitualmente mediante técnicas de suavizado, ya sea estabilizando tasas considerando unidades vecinas o ponderando con la tasa global en función del tamaño poblacional.

2.1.2. Dependencia espacial

La **dependencia espacial** es la tendencia de los valores de una variable a estar correlacionados con los de las áreas próximas. Esta propiedad está recogida en la **Primera Ley de Geografía** de Tobler (1970):

> *Everything is related to everything else, but near things are more related than distant things. [Todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las lejanas.]*

En datos agregados por área, la dependencia espacial se manifiesta cuando los valores observados en unidades vecinas tienden a parecerse más entre sí que los de unidades lejanas, al compartir condiciones socioeconómicas, climáticas o ambientales. Esta estructura no cumple con el supuesto de independencia de las observaciones que exigen los modelos de regresión ordinarios. Cuando existe dependencia espacial, los residuos de un modelo no espacial presentan autocorrelación, lo que sesga la estimación de los coeficientes e invalida los intervalos de confianza y las inferencias derivadas. Para identificar si existe o no dependencia espacial, se puede calcular la autocorrelación espacial con el índice de Moran I.

Esta es la razón por la que el análisis de datos de área requiere métodos específicos que incorporen explícitamente la estructura de vecindad. Las siguientes secciones presentan esos métodos; como paso previo, las siguientes secciones describen las herramientas que permiten formalizar la proximidad entre áreas: el vecindario y la matriz de adyacencia.

2.2. Primeros pasos para el análisis de datos

Como en cualquier análisis, designar el alcance y los objetivos que se desean lograr es esencial para poder llevar a cabo un análisis de datos. Específicamente para el análisis espacial de datos agregados por área, es útil plantear preguntas como:

- ¿Se desea simplemente describir mediante mapas los datos y observar si hay patrones? Por ejemplo, interesa observar clústers o saber la distribución de una especie en una región.
- Al utilizar datos de área que representan conteos, ¿se compararán las áreas por cantidad de conteos observados similares, sin tener en cuenta la población expuesta? O, ¿se compararán ajustando por la exposición? Por ejemplo, en el caso de enfermedades la densidad de población en una metropolís no es igual a las áreas colindantes y es necesario saber el enfoque del estudio o se podría presentar un problema de área modificable.
- ¿Se desea explicar la variable de interés a través de una covariable? Por ejemplo, modelar el precio medio de la vivienda en función de la accesibilidad al transporte público.
- ¿Se desea obtener mapas suavizados? Es decir, se desea utilizar métodos de suavizado en los mapas, y si es así identificar qué métodos son adecuados para los datos, cómo funcionan, qué ventajas o desventajas aportan.

Habitualmente, en el análisis de datos agregados por áreas se suelen seguir los siguientes bloques metodológicos (Moraga, 2023):

1. Realizar análisis exploratorio de datos espaciales (ESDA) y diagnóstico espacial.
2. Utilizar modelado de referencia (no espacial) para efectos comparativos.
3. Utilizar modelos espaciales areales: incorporar la componente espacial para una representación más precisa de la estructura.
4. Validar los modelos creados mediante la evaluación de medidas de sensibilidad y precisión.
5. Comunicar los resultados.

2.3. Preparación espacial

En el contexto de este trabajo, que pretende aplicar técnicas de análisis de datos espaciales y, en particular, de datos agregados por áreas, resulta fundamental que los datos a utilizar estén disponibles en el formato adecuado. En otras palabras, para trabajar con datos agregados por área es imprescindible asegurar que tanto los datos como la geometría de los polígonos estén correctamente definidos e integrados, y que la estructura de proximidad entre áreas esté formalmente definida. Esta etapa determina la validez de todos los pasos posteriores.

2.3.1. Definición de la unidad espacial

Así como es necesario definir el fenómeno o la variable que se quiere estudiar, es necesario definir la unidad de análisis o la escala. Es decir, determinar qué unidad de área es más apropiada para representar el proceso de estudio, por ejemplo municipios, secciones censales, rejillas determinadas, etc. En muchos casos, esta decisión está impuesta por la fuente estadística, pero es importante considerar el MAUP ya que los resultados pueden variar según la unidad elegida. Del mismo modo, datos ligados a formatos predeterminados por la fuente pueden no siempre coincidir en unidad espacial, i.e. la renta bruta del hogar por sector censal cuando los datos de contaminación están en una rejilla de $1km^2$. Aunque no es imposible trabajar con datos en diferentes unidades espaciales, existen consideraciones y metodologías que se deben aplicar para lograr armonizar los datos como lo muestra Hewitt et al. (2024) al armonizar rejillas con sectores censales.

2.3.2. Vecindario espacial

A diferencia de los datos geoestadísticos o datos puntuales, al agregar datos por área es útil definir el concepto del vecindario de un área para el análisis exploratorio de los datos. Esto se hace con el objetivo de determinar si áreas cercanas tienen valores similares o no ya que pueden compartir condiciones ambientales, socioeconómicas, culturales, entre otras, que pueden influenciar a la variable observada. El **vecindario espacial** N_i del área A_i es el conjunto de áreas que se consideran próximas o

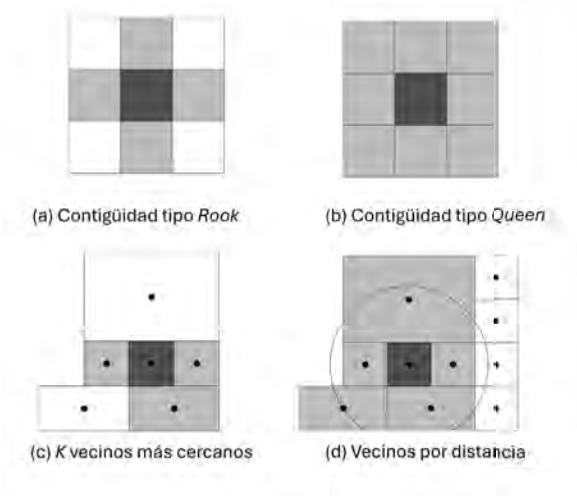


Figura 2.1: Criterios de vecindad para datos de área. El área de referencia A_i aparece en gris oscuro y sus vecinos N_i, en gris claro

Referencia: Matrices de vecindarios (Moraga, 2023)

conectadas a ella según algún criterio geográfico establecido por el analista. La condición mínima es que un área no sea vecina de sí misma: $A_i \notin N_i$. La definición del vecindario es una decisión metodológica fundamental: condiciona la estructura de dependencia asumida en el análisis y, por tanto, todos los resultados posteriores (Anselin, 1988). Por ello, la elección del criterio del vecindario debe considerar el proceso que se desea modelar y el territorio que se está analizando. Es decir, el territorio de estudio puede estar constituido por diferentes entidades geométricas, incluyendo islas (áreas sin conectividad a otras áreas) y, por otro lado, también por otros tipos de superficies que pueden no ser de interés. Estas consideraciones deben reflejarse en el vecindario a partir del criterio seleccionado. Los criterios más habituales son los siguientes (véase Figura 2.1):

Contigüidad *Rook*. Proveniente de la analogía del ajedrez en el que la torre (*Rook*) solo se mueve en linea recta de arriba a abajo y de izquierda a derecha, este criterio, considera vecinos a áreas que comparten un lado completo o más de un punto de contacto. Produce vecindarios relativamente restrictivos y es adecuado cuando la variable solo se propaga a través de fronteras continuas.

Contigüidad *Queen*. Este criterio hace referencia al movimiento de la reina (*Queen*) en ajedrez; dos áreas son vecinas si comparten al menos un punto de su frontera, incluyendo vértices. Genera vecindarios más amplios que la contigüidad *Rook* y es el criterio más empleado en la práctica con polígonos administrativos irregulares, pues reduce el riesgo de generar áreas sin vecinos (*islas*).

k **vecinos más cercanos (KNN).** El vecindario de A_i está formado por los k polígonos cuyos centroides se encuentran más próximos al centroide de A_i. Garantiza que todo polígono tenga exactamente k vecinos, lo que resulta especialmente útil en presencia de islas o geometrías muy

heterogéneas. Su principal limitación es que la relación puede ser asimétrica: $A_j \in N_i$ no implica necesariamente $A_i \in N_j$.

Distancia umbral. N_i está formado por todas las áreas cuyo centroide se encuentra dentro de un radio definido $d > 0$ alrededor del centroide de A_i. En este criterio el número de vecinos varía entre áreas. Además, si el umbral es demasiado pequeño pueden aparecer islas, mientras que si es demasiado grande la estructura de vecindad pierde resolución local (Pebesma y Bivand, 2023).

El Cuadro 2.1 resume y compara la propiedades principales de los cuatro criterios.

Criterio	**Vecinos garantizados**	**Número de vecinos**	**Asimetría posible**	**Más adecuado para**
Contigüidad *Rook*	No	Variable	No	Cuadrículas regulares
Contigüidad *Queen*	No	Variable	No	Polígonos irregulares, uso general
KNN	Sí	Fijo (k)	Sí	Heterogeneidad de tamaño, islas
Distancia	Depende del umbral	Variable	No	Procesos con escala conocida

Cuadro 2.1: Comparativa de criterios de definición de vecindario.
Fuentes: Anselin (1988), Bivand et al. (2013), LeSage y Pace (2009).

Un problema que surge al determinar el vecindario es el problema de las islas o unidades separadas debido a las condiciones geográficas, e.g. áreas separadas por ríos. Esto llega a ser un problema cuando se definen vecinos por contigüidad *Rook*, ya que áreas con un solo punto de contacto no tendrán vecinos. También, puede llegar a haber áreas sin vecinos cuando se delimitan vecinos por distancia (Pebesma y Bivand, 2023). De manera general, se recomienda:

- **Documentar:** registrar el número medio de vecinos y la distribución de ese número entre áreas para realizar un análisis de sensibilidad.
- **Comprobar conectividad:** verificar que todas las áreas tienen un vecindario, es decir, que no existen áreas aisladas. Las islas deben mitigarse explícitamente, por ejemplo, asignándoles el vecino más próximo, ya que algunos modelos areales requieren conectividad global (Bivand et al., 2013).

2.3.3. Matriz de pesos espaciales

Una vez definido el vecindario, la estructura de proximidad entre todas las áreas de la región se formaliza en la **matriz de pesos espaciales W**, de dimensión $n \times n$, cuyos elementos w_{ij} cuantifican la intensidad de la relación entre A_i y A_j. Por definición, $w_{ii} = 0$ para todo i: un área no puede ser vecina de sí misma.

El primer paso para construir **W** es obtener la **matriz de adyacencia A** $= (a_{ij})$, que recoge de forma binaria qué pares de áreas se consideran vecinas según el criterio elegido:

$$a_{ij} = \begin{cases} 1 & \text{si } A_j \in N_i, \\ 0 & \text{en caso contrario,} \end{cases} \qquad a_{ii} = 0. \tag{2.3}$$

El número de vecinos del área A_i es $n_i = \sum_{j=1}^{n} a_{ij}$. Cuando el criterio de vecindad es simétrico (contigüidad *Queen* o *Rook*, distancia umbral), la matriz de adyacencia es también simétrica: $a_{ij} = a_{ji}$. Sin embargo, con el criterio KNN esta propiedad puede no cumplirse.

A partir de la matriz de adyacencia se construye la matriz de pesos espaciales $\mathbf{W}$, asignando un valor w_{ij} a cada par de áreas vecinas ($a_{ij} = 1$) y cero al resto. La elección de cómo asignar esos pesos da lugar a dos grandes familias.

Pesos binarios

En el caso más simple, los pesos coinciden con la matriz de adyacencia:

$$w_{ij} = a_{ij} = \begin{cases} 1 & \text{si } A_j \in N_i, \\ 0 & \text{en caso contrario.} \end{cases} \tag{2.4}$$

Esta especificación trata igual todas las relaciones del vecindario, independientemente de la longitud de la frontera compartida o de la distancia entre centroides. Aunque no aporta más información respecto de la matriz de adyacencia, es la opción por defecto en la mayoría de los análisis con datos administrativos y es la que se utiliza en la formulación estándar de algunos modelos espaciales.

Pesos ponderados

Esta opción es adecuada para procesos de difusión continua como la dispersión de contaminantes atmosféricos. Cuando la intensidad de la relación entre áreas varía con la distancia o con alguna propiedad geométrica de los polígonos, puede ser más apropiado asignar pesos continuos. Una forma de asignarlos es con la **inversa de la distancia**: $w_{ij} = d_{ij}^{-\alpha}$, donde d_{ij} es la distancia entre centroides y $\alpha > 0$ controla la velocidad de decaimiento con la distancia. A mayor α, la influencia entre áreas decrece más rápidamente.

Normalización por filas

Independientemente del tipo de pesos (binarios o ponderados), en ocasiones se suele aplicar una **normalización por filas** de modo que los pesos de cada fila sumen la unidad:

$$w_{ij}^{*} = \frac{w_{ij}}{\sum_{k=1}^{n} w_{ik}}, \qquad \text{siempre que } \sum_{k=1}^{n} w_{ik} > 0. \tag{2.5}$$

Esta normalización es útil porque resuelve el problema concreto de comparabilidad. Sin ella, el **retardo espacial** $\sum_j w_{ij}\, z_j$ depende del número de vecinos n_i de cada área: un área con muchos vecinos acumula una suma con más términos que un área con pocos, aunque los valores individuales de sus vecinos sean parecidos. Esto impide comparar el retardo espacial entre áreas con vecindarios de tamaño muy distinto, situación habitual en regiones con unidades administrativas heterogéneas como la que se analiza en la Parte II de este trabajo. En otras palabras, con esta normalización, el **retardo espacial** $\sum_j w_{ij}^{*}\, z_j$ se interpreta como la media ponderada de los valores de los vecinos de A_i, facilitando la comparación entre áreas con distinto número de vecinos.

2.4. Análisis Exploratorio de Datos Espaciales

El *exploratory spatial data analysis (ESDA)* tiene como objetivos: describir la distribución de la variable de interés, con mapas, y diagnosticar la presencia de autocorrelación espacial antes de ajustar cualquier modelo.

2.4.1. Mapas coropléticos

Para describir la distribución de la variable de manera sencilla es común utilizar mapas temáticos. Los mapas coropléticos representan el valor de la variable de interés en una escala de color aplicada a polígonos de una región geográfica. Por su practicidad, es la representación estándar de los datos agregados por área. Sin embargo, deben tenerse en cuenta las siguientes consideraciones que pueden afectar la percepción visual del patrón:

- **Método de corte:** La elección de los cortes de clasificación en la construcción de un mapa coroplético es muy influyente, ya que dicta cómo se agrupan los valores. Realizar cortes por cuantiles (mismo número de unidades en cada clase) garantiza que todas las clases aparezcan en el mapa formando mapas visualmente equilibrados. Sin embargo, si la distribución de los datos es asimétrica, se pueden agrupar en la misma clase valores muy distintos. Por otro lado, realizar cortes según intervalos iguales permite interpretar fácilmente el mapa, pero puede producir mapas desequilibrados en distribuciones asimétricas, concentrando muchas observaciones en pocas clases y pocas observaciones en clases extremas. En práctica, suele ser útil explorar el mapa con distintos métodos de clasificación y, después, justificar la elección.

- **Tasas frente a recuentos:** representar recuentos brutos sin ajustar por exposición puede ser engañoso, las áreas más pobladas tenderán siempre a mostrar valores más altos. Para prevenir esta condición deben representarse las tasas o proporciones relativas a la población en riesgo.

- **Problema de unidades pequeñas:** como se mencionó en la sección 2.1.1, las áreas pequeñas pueden ocasionar ruido en el mapa en forma de valores extremos debido a su baja densidad poblacional.

2.4.2. Autocorrelación Espacial Global

Una medida de explorar la presencia de dependencia espacial, es decir, si áreas vecinas presentan valores similares es calcular la **autocorrelación espacial global**. La medida de autocorrelación espacial global puede evaluarse con índices que indican el grado en el que observaciones similares ocurren cerca una de otra dentro de la región de estudio. El índice más utilizado para evaluar la autocorrelación espacial en datos areales es el **índice I de Moran**, aunque también se puede utilizar la **C de Geary**.

(Moraga, 2023). Dado el vector de valores observados $\mathbf{z} = (z_1, \ldots, z_n)^\top$ y una matriz de pesos $\mathbf{W}$, el índice I se define como:

$$I = \frac{n}{S_0} \cdot \frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\displaystyle\sum_{i=1}^{n}(z_i - \bar{z})^2} = \frac{n}{S_0} \cdot \frac{\mathbf{z}_c^\top \mathbf{W}\mathbf{z}_c}{\mathbf{z}_c^\top \mathbf{z}_c}, \tag{2.6}$$

donde $\bar{z} = \dfrac{1}{n}\sum_{i=1}^{n} z_i$ es la media muestral, $\mathbf{z}_c = (z_1 - \bar{z}, \ldots, z_n - \bar{z})^\top$ es el vector centrado y $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ es la suma total de pesos. Este índice puede interpretarse, de manera aproximada, como una medida similar a una correlación, dado que vincula la desviación de cada área respecto de la media con las desviaciones que muestran sus áreas vecinas. Su valor permite sintetizar si, en términos globales, la variable tiende o no a presentar un patrón de agrupamiento espacial. La interpretación general es la siguiente:

- $I > 0$: autocorrelación positiva, áreas vecinas tienden a presentar valores similares (agrupamiento o clúster).

- $I < 0$: autocorrelación negativa, áreas vecinas tienden a presentar valores distintos (dispersión).

- $I \approx \frac{-1}{(n-1)}$: valores (positivos o negativos) próximos a este umbral indican ausencia de patrón espacial claro, independientemente del signo. Este valor representará el valor esperado bajo H_0: ausencia de autocorrelación espacial.

Gráficamente, el índice global I puede representarse con el valor de la variable, z_i, en el eje-x y el retardo espacial, $\sum_j w_{ij} z_j$, en el eje-y para crear un **diagrama de dispersión de Moran** (véase la Figura 2.2) en el que se muestran las observaciones como una nube de puntos y con una recta cuya pendiente indica la correlación positiva o negativa. Es importante mencionar, que cuando los valores de la variable están estandarizados y **W** está normalizada por filas, la pendiente del diagrama de dispersión corresponde al valor de I global.

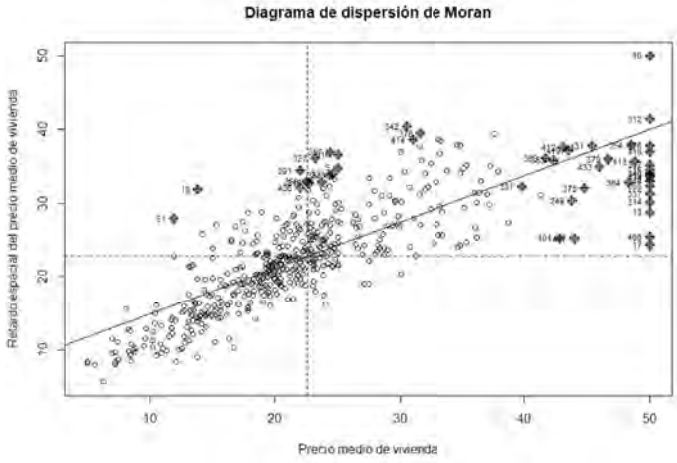


Figura 2.2: Diagrama de dispersión de Moran para la variable precio medio de vivienda según las secciones censales de Boston, USA en 1987, obtenidas del paquete **spData**. El diagrama muestra las observaciones frente a los valores de retardo espacial, calculado a partir de una matriz tipo *Queen*.

Referencia: Diagrama de dispersión de Moran (Moraga, 2023)

Por otro lado, la significación estadística se evalúa mediante un contraste de hipótesis donde:

$$H_0 : \text{ausencia de autocorrelación espacial global}$$

$$H_1 : \text{presencia de autocorrelación espacial global}$$

En donde el valor esperado del índice bajo H_0 es:

$$E_0[I] = -\frac{1}{n-1}$$

con varianza denotada $Var_0[I]$. Con ello, en este contraste se construye el estadístico tipificado:

$$Z_I = \frac{(I - E_0[I])}{\sqrt{Var_0[I]}}$$

y se compara bajo el supuesto de normalidad o permutaciones aleatorias (simulaciones Monte Carlo), que proporcionan un p-valor empírico sin depender de aproximaciones asintóticas, o, incluso en algunos casos se pueden utilizar técnicas de remuestreo (bootstrap paramétrico) (Bivand et al., 2013).

De manera alternativa, la autocorrelación espacial puede medirse mediante la C de Geary:

$$C = \frac{(n-1)}{2S_0} \cdot \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(z_i - z_j)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}. \tag{2.7}$$

A diferencia del índice de Moran, la C de Geary muestra una mayor sensibilidad a las variaciones locales entre áreas vecinas y una menor atención a la estructura global. En este trabajo se opta por el índice de Moran, dado que permite una interpretación directa como coeficiente de correlación espacial, mientras que la C de Geary se reserva como posible herramienta para futuros análisis de sensibilidad.

2.4.3. Autocorrelación espacial local

El índice global resume el patrón espacial del conjunto completo de áreas mediante un único valor y no permite localizar dónde se concentra la asociación espacial. Para ello se utilizan los **indicadores locales de asociación espacial** (LISA, *Local Indicators of Spatial Association*), que contrastan el valor observado en A_i con los valores de sus vecinos según $\mathbf{W}$. Esto permite localizar las zonas donde la asociación espacial es más intensa y distinguir las áreas que contribuyen al patrón general de aquellas cuyo comportamiento resulta diferente respecto a sus vecinas (Anselin, 1995).

El **índice local de Moran** es uno de los LISAs más populares, el cual, para el área A_i, se define como:

$$I_i = \frac{(z_i - \bar{z})}{s^2} \sum_{j=1}^{n} w_{ij}(z_j - \bar{z}), \qquad s^2 = \frac{1}{n} \sum_{i=1}^{n} (z_i - \bar{z})^2. \tag{2.8}$$

Esta expresión compara la desviación del área A_i respecto a la media con una combinación ponderada de las desviaciones de sus vecinos. De esta manera, el índice global es la suma reescalada de los índices locales: $I = S_0^{-1} \sum_{i=1}^{n} I_i$.

Los valores de los LISAs suelen ser graficados en mapas para representar visualmente las ubicaciones de áreas con asociación local alta o baja. Un valor positivo I_i indica que A_i está rodeada de áreas con valores similares (alto con alto o bajo con bajos), mientras que, un valor negativo de I indica que A_i está rodeada de áreas con valores distintos (alto con bajo o bajo con alto).

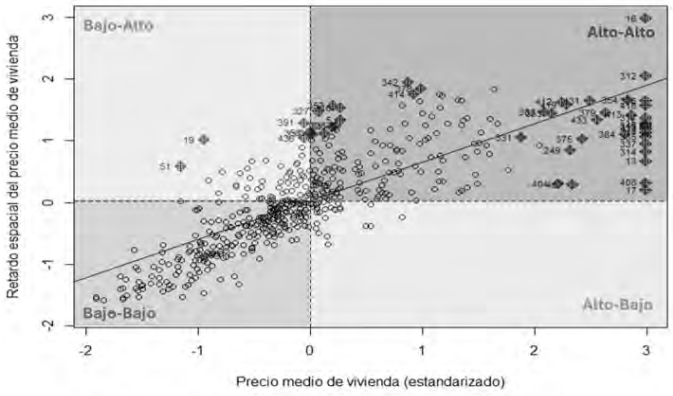
Para evaluar la significación de cada I_i es necesario valorar si ese valor es compatible o no con la H_0: ausencia de correlación espacial local. Para ello, se puede evaluar mediante un estadístico tipificado o por permutaciones, similar al índice global, obteniendo p-valores para cada área. Los resultados deben interpretarse con cautela: al realizar simultáneamente n contrastes, la tasa de falsos positivos aumenta y los p-valores locales deben entenderse principalmente como una herramienta exploratoria para identificar configuraciones de interés, no como evidencia concluyente área por área (Anselin, 1995).

Clústeres LISA

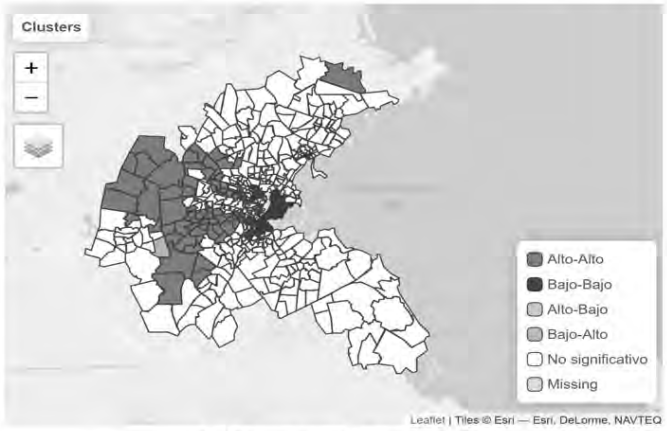
Obteniendo los valores significativos del índice I de Moran local, cada área se clasifica en uno de cuatro cuadrantes del **diagrama de dispersión de Moran** para identificar el tipo de agrupaciones o clústeres. Específicamente, se grafica el valor estandarizado de la variable observada frente a su retardo espacial (véase ejemplo en la Figura 2.3) para identificar, según cuadrantes, las siguientes agrupaciones:

- **Alto-Alto (AA)**: área con valor alto rodeada de vecinos con valores altos ($I_i > 0$), es decir, se observa como un clúster de valores elevados.
- **Bajo-Bajo (BB)**: área con valor bajo rodeada de vecinos con valores bajos ($I_i > 0$), da lugar a un clúster de valores reducidos.
- **Alto-Bajo (AB)**: valor alto rodeado de vecinos con valores bajos ($I_i < 0$), es un outlier espacial positivo.
- **Bajo-Alto (BA)**: valor bajo rodeado de vecinos con valores altos ($I_i < 0$), es un outlier espacial negativo.

Posteriormente, se puede graficar geográficamente la clasificación de clústeres mediante una escala de colores para visualizar la localización de clústeres locales, dejando sin colorear aquellas áreas con índices no significativos como se observa en la Figura 2.3(b). Esta visualización permite destacar las áreas con índices locales significativos y evidenciar la presencia o no de autocorrelación y, con ello, la presencia o no de dependencia espacial.



(a) Diagrama de dispersión de Moran identificando cuadrantes



(b) Mapa con clústeres LISA

Figura 2.3: Clústeres LISA para la variable precio medio de vivienda (estandarizada) según las secciones censales de Boston, USA en 1987, obtenidas del paquete **spData**. (a) Diagrama de dispersión de Moran local, muestra las observaciones estandarizadas frente a sus valores de retardo espacial, calculado a partir de una matriz tipo *Queen*. (b) Mapa de clústeres obtenidos a partir del diagrama de dispersión de Moran local.

Referencia: Clústeres (Moraga, 2023)

2.4.4. Análisis Bivariante

El análisis univariado de autocorrelación espacial, mediante la I de Moran y los indicadores LISA permite caracterizar la estructura espacial de cada variable por separado. Sin embargo, en algunos casos se trabaja fenómenos bivariantes en los cuales es necesario contemplar la asociación espacial de dos variables. Para ello, se recurre al índice de Moran bivariado y al análisis BiLISA (*Bivariate Local Indicators of Spatial Association*)(Anselin, 2019).

El índice de Moran bivariado global cuantifica el grado de asociación espacial entre dos variables distintas X (principal) e Y (secundaria) de la manera:

$$I^{BV} = \frac{\sum_i (x_i \sum_j w_{ij} y_j)}{\sum_i x_i^2} \tag{2.9}$$

donde x_i e y_i son los valores estandarizados de las variables en el área i, w_{ij} son los elementos de la

matriz de pesos espaciales $\mathbf{W}$ normalizada por filas, y n es el número de áreas. En el caso bivariante, un valor $I^{BV} > 0$ indica que las áreas con valores altos de la primera variable tienden a estar rodeados de áreas con valores altos de la segunda variable, y un valor $I^{BV} < 0$ indica el patrón contrario.

Es importante destacar que, a diferencia del índice univariado, el índice bivariado no es simétrico, es decir $I^{BV}(\mathbf{X}, \mathbf{Y}) \neq I^{BV}(\mathbf{Y}, \mathbf{X})$ ya que el retardo espacial se aplica a una de las variables. Aún así, al tratar con variables distintas con unidades diferentes, es importante estandarizar las variables para el análisis bivariante.

Por otro lado, el análisis BiLISA descompone el índice de Moran bivariado global en contribuciones locales, permitiendo identificar las áreas que presentan asociaciones espaciales bivariadas estadísticamente significativas (Anselin, 2019). De esta forma, el estadístico local para el área i es:

$$I_i^{BV} = x_i \sum_j w_{ij} y_j = x_i \cdot (\mathbf{WY})_i \tag{2.10}$$

donde $(\mathbf{WY})_i$ es el retardo espacial de la variable y en el área i, es decir, el promedio ponderado de los valores de $\mathbf{Y}$ en las áreas vecinas. Al igual que en el LISA univariado, en el BiLISA es posible clasificar las áreas en uno de los cuatro cuadrantes, pero ahora tomando en el eje-x el valor de la variable $\mathbf{X}$ y en el eje-y el valor del retardo espacial de la variable secundaria denotado por $(\mathbf{WY})_i$. Ahora bien, considerando únicamente las áreas con estadístico local estadísticamente significativo es posible crear clústeres de la siguiente manera:

- **Alto-Alto (AA)**: área con valor positivo $x_i > 0$ rodeada de retardos espaciales positivos $(\mathbf{WY})_i > 0$.
- **Alto-Bajo (AB)**: área con valor positivo $x_i > 0$ rodeada de retardos espaciales negativos $(\mathbf{WY})_i < 0$.
- **Bajo-Bajo (BB)**: área con valor negativo $x_i < 0$ rodeada de retardos espaciales negativos $(\mathbf{WY})_i < 0$.
- **Bajo-Alto (BA)**: área con valor negativo $x_i < 0$ rodeada de retardos espaciales positivos $(\mathbf{WY})_i > 0$.

Gráficamente, estos clústeres permiten identificar áreas de interés dentro del ESDA, ya que la asimetría permite detectar patrones de covariación espacial entre las dos variables que no serían visibles en los análisis univariados por separado. En síntesis, realizar el ESDA contemplando estos pasos permite establecer si el proceso observado presenta o no correlación. En consecuencia, este resultado ofrece una base para sustentar futuras modelizaciones del proceso.

2.5. Modelos de referencia no espaciales

Aunque el ESDA permita identificar si existe dependencia espacial, el alcance del análisis puede ser modelizar el comportamiento de la variable de interés y con ello, incorporar la dependencia espacial en el modelado. Sin embargo, antes de introducir efectos espaciales es recomendable ajustar modelos de referencia o *baseline*, que sirvan de comparación y permitan aislar la contribución de la estructura espacial. Estos modelos varían según el tipo de variable respuesta que se desea modelar.

2.5.1. Respuesta continua

Cuando la variable respuesta es continua y puede asumirse normal, el modelo de referencia es la **regresión lineal**. Ya sea con regresión simple o múltiple, la regresión lineal permite crear un modelo lineal que describe la relación de la variable respuesta Z y una o más variables explicativas o predictores X_j.

De manera general, para p distintos predictores el modelo de regresión lineal toma la siguiente forma:

$$Z_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2). \tag{2.11}$$

donde β_0 representa el coeficiente del intercepto, X_j representa j-ésimo predictor y β_j cuantifica la asociación entre la j-ésima variable y la respuesta (James et al., 2013).

Recordando que Z_i representa la variable respuesta en el área i y $Z_i \equiv Z(A_i)$. También, en el contexto de datos agregados por área las covariables se reescriben como:

$$X_{ij} \equiv X_j(A_i)$$

y

$$\varepsilon_i \equiv \varepsilon(A_i)$$

donde X_{ij} representa el valor de la j-ésima variable y ε_i el valor del error, ambos para el área i.

Adicionalmente, si la distribución de los errores es asimétrica, puede ser útil aplicar una transformación logarítmica a la respuesta.

2.5.2. Limitación fundamental

La limitación clave de los modelos no espaciales es que, si existe dependencia espacial residual, ésta queda sin capturar, lo que puede esconder asociaciones con covariables de interés. En un escenario ideal, se puede aplicar el test de Moran a los residuos del modelo; si este resulta significativo, será necesario pasar a los modelos espaciales descritos en la sección siguiente.

2.6. Modelos espaciales areales

La existencia de autocorrelación espacial en los residuos indica que el modelo lineal no logra explicar una parte de la variabilidad de los datos, la cual no puede atribuirse únicamente a las covariables y está vinculada con la estructura espacial del territorio. Si se ignora esta dependencia espacial, una fracción de la variación asociada a factores geográficos compartidos entre áreas vecinas es absorbida por los coeficientes de las covariables, generando estimaciones sesgadas (Bivand et al., 2013; Moraga, 2023). En tales circunstancias, lo más adecuado es representar explícitamente la dependencia espacial incorporando al predictor lineal una **componente latente** con estructura espacial. En el marco bayesiano, los modelos más habituales para datos agregados por áreas son el modelo ICAR (*Intrinsic Conditional Autoregressive Model*) y el modelo BYM (*Besag-York-Mollié*). La estimación se lleva a cabo mediante la aproximación de Laplace anidada integrada (INLA, *Integrated Nested Laplace Approximation*), que resulta considerablemente más eficiente computacionalmente que implementar Cadenas de Markov-Monte Carlo (MCMC) (Moraga, 2023).

2.6.1. Modelos CAR e ICAR

Los modelos **CAR** (*Conditional Autoregressive Models*) introducen la dependencia espacial especificando directamente un vector latente $\mathbf{U}$, representando el proceso areal con un modelo de la forma:

$$\mathbf{Z} = \mu + \mathbf{U} \tag{2.12}$$

donde $\mu = (\mu_1, \ldots, \mu_n)^\top$ y $\mathbf{U} = (U_1, \ldots, U_n)^\top$. Los modelos **CAR** consideran las distribuciones condicionales completas de cada componente latente (U) dado el resto, de la siguiente forma:

$$U_i \mid \mathbf{U}_{-i} \sim \mathcal{N}\left(\sum_{j=1}^{n} b_{ij}\, u_j,\ \sigma_i^2 \right), \qquad b_{ii} = 0, \tag{2.13}$$

donde los coeficientes b_{ij} se obtienen a partir de la matriz de pesos $\mathbf{W}$ y expresan cómo el valor u_j del área A_j contribuye a la media condicional de U_i, mientras que σ_i^2 denota la varianza condicional de U_i. La clave del CAR es que, si $b_{ij} = 0$, entonces U_i y U_j son *condicionalmente independientes* dado el resto: la dependencia condicional es local y solo actúa a través del vecino inmediato. Un vector aleatorio gaussiano con esta propiedad recibe el nombre de **campo aleatorio gaussiano de Markov** (GMRF, *Gaussian Markov Random Field*).

La versión más empleada de los modelos CAR es el **modelo ICAR** que introduce el vector latente $\mathbf{U}$ como un efecto aleatorio espacialmente estructurado que captura la variabilidad compartida entre áreas vecinas. Se fundamenta en la Primera Ley de Geografía de Tobler (Tobler, 1970) y plantea que el componente espacial de cada área se aproxima al promedio del componente espacial de sus áreas vecinas. Con ello, la distribución condicional de U_i, para cada área A_i, dado el resto de efectos $\mathbf{U}_{-i}$ es:

$$U_i \mid \mathbf{U}_{-i} \sim \mathcal{N}\left(\frac{\sum_j w_{ij}\, u_j}{\sum_k w_{ik}},\ \frac{\sigma^2}{\sum_k w_{ik}} \right). \tag{2.14}$$

donde w_{ij} denota los pesos espaciales de modo que el peso asignado a un área respecto de sí misma es cero, es decir, $w_{ii} = 0$.

En lo que respecta a las propiedades de los modelos ICAR, la media condicional de U_i corresponde a la media ponderada de las áreas vecinas, mientras que la varianza condicional disminuye a medida que el área está más conectada, lo que refleja que un entorno con mayor número de vecinos proporciona más información. En el caso Gaussiano, la estructura de dependencia condicional puede describirse mediante la **matriz de precisión $\mathbf{Q} = \mathbf{\Sigma^{-1}}$**. En el caso particular de los modelos ICAR, la matriz de precisión asociada viene dada por:

$$\mathbf{Q} = \frac{1}{\sigma^2}(\mathbf{D} - \mathbf{W}), \qquad \mathbf{D} = \operatorname{diag}\left(\sum_j w_{1j}, \ldots, \sum_j w_{nj} \right). \tag{2.15}$$

Es relevante señalar que, en los modelos espaciales, se suele operar con matrices binarias *sin* normalizar. La normalización por filas se utiliza principalmente en el análisis exploratorio y en los modelos con retardo espacial.

Propiedad de suavizado ICAR

La densidad conjunta (impropia) de $\mathbf{U}$ cuando la matriz de pesos $\mathbf{W}$ es simétrica se escribe como:

$$f(\mathbf{u} \mid \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}\, \mathbf{u}^\top (\mathbf{D} - \mathbf{W})\mathbf{u} \right) = \exp\left(-\frac{1}{4\sigma^2} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(u_i - u_j)^2 \right), \tag{2.16}$$

donde la segunda igualdad muestra que la densidad depende de la suma de las diferencias al cuadrado entre áreas vecinas, $(u_i - u_j)^2$, con signo negativo en la función exponencial: cuanto mayor es esa diferencia, menor es la densidad asociada a esa configuración. En consecuencia, el modelo penaliza las configuraciones en las que áreas vecinas toman valores muy distintos y favorece aquellas en las que los vecinos presentan valores similares, induciendo así **suavizado espacial**.

2.6.2. Modelo BYM

El modelo **BYM** (*Besag-York-Mollié*) (Besag et al., 1991) extiende el ICAR combinando la componente espacial estructurada U con una componente no estructurada V. Es decir, descompone la variabilidad no explicada por los efectos fijos en dos componentes independientes:

$$Z_i = \mu_i + U_i + V_i, \qquad U_i \mid \mathbf{U}_{-i} \sim \text{ICAR}, \qquad V_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2), \tag{2.17}$$

con **U** y **V** independientes entre sí. Con covariables, la media toma la forma:

$$Z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + U_i + V_i. \tag{2.18}$$

O, en forma matricial, de manera análoga a la ecuación 2.12, se expresa como:

$$\mathbf{Z} = \mathbf{X}\beta + \mathbf{U} + \mathbf{V} \tag{2.19}$$

De esta forma, la variabilidad observada entre áreas puede descomponerse en tres partes:

1. **Componente sistemática**: la parte explicada por las covariables, $\mathbf{X}\boldsymbol{\beta}$.
2. **Componente espacial estructurada** (**U**): variación entre áreas vecinas que sigue un patrón espacial suave, no explicada por las covariables.
3. **Componente no estructurada** (**V**): heterogeneidad específica de cada área sin pauta espacial, equivalente a un término de error independiente entre áreas.

La descomposición permite separar la variabilidad compartida entre áreas vecinas (recogida por U_i) de la heterogeneidad específica de cada unidad no explicada por las covariables (recogida por V_i). El BYM es el estándar en **cartografía de enfermedad** y en estudios ambientales por su flexibilidad para capturar tanto patrones espaciales suaves como anomalías locales, y por la disponibilidad de implementaciones eficientes mediante INLA.

2.6.3. Resultados que se obtienen en la práctica

Los modelos espaciales producen cuatro tipos de salidas de interés:

- **Mapas suavizados**: estimaciones del riesgo o la intensidad ajustada por covariables, más estables que las tasas brutas en áreas pequeñas.
- **Mapas de la componente espacial $\hat{\mathbf{U}}$**: el patrón residual compartido entre vecinos, no explicado por las covariables.
- **Incertidumbre espacialmente informada**: intervalos de credibilidad por área, más estrechos donde hay más información y más amplios en áreas pequeñas.
- **Hiperparámetros $\hat{\phi}^2$, $\hat{\sigma}^2_U$ y $\hat{\sigma}^2_V$**: su magnitud relativa indica qué proporción de la variabilidad residual tiene estructura espacial.

2.6.4. Cautela interpretativa

La presencia de un efecto espacial significativo no implica causalidad. La componente U_i recoge variación compartida entre vecinos que puede deberse a factores no medidos con estructura espacial, a errores de medición correlacionados o a artefactos de la unidad espacial elegida. Cualquier interpretación sustantiva debe ir acompañada de una discusión sobre las posibles fuentes de esa correlación residual.

2.7. Comparación de modelos

Una vez ajustados tanto el modelo de referencia como el modelo espacial BYM, es imprescindible contar con criterios formales que permitan elegir la especificación más adecuada. En el marco bayesiano implementado con `R-INLA`, los criterios más empleados son el DIC y el WAIC, ambos calculados directamente a partir de las aproximaciones de Laplace anidadas que INLA utiliza para estimar la distribución posterior, sin necesidad de recurrir a cadenas de Markov Monte Carlo (Rue et al., 2009).

2.7.1. Criterio de información de la desviancia (DIC)

El (*Deviance Information Criterion*, DIC) (Spiegelhalter et al., 2002) es una generalización bayesiana del criterio de información de Akaike (AIC) pensada para modelos jerárquicos con efectos aleatorios, como el BYM. Se define a partir de la **desviancia**

$$D(\boldsymbol{\theta}) = -2 \log p(\mathbf{z} \mid \boldsymbol{\theta}) + C, \tag{2.20}$$

donde $p(\mathbf{z} \mid \boldsymbol{\theta})$ es la verosimilitud del modelo y C una constante que no depende de $\boldsymbol{\theta}$ y que se cancela al comparar modelos. El DIC combina la bondad de ajuste, cuantificada mediante la desviancia posterior media $\overline{D(\boldsymbol{\theta})}$, con un término de penalización p_D que refleja la complejidad efectiva del modelo:

$$\mathrm{DIC} = \overline{D(\boldsymbol{\theta})} + p_D, \qquad p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}), \tag{2.21}$$

donde $\bar{\boldsymbol{\theta}}$ es la media posterior de los parámetros. El término p_D mide cuánto se *ajustan* los datos en promedio en comparación con el ajuste obtenido al sustituir los parámetros por su valor puntual posterior, y actúa como el número efectivo de parámetros: en un modelo BYM, por ejemplo, p_D suele ser sustancialmente menor que el número nominal de efectos aleatorios U_i, porque estos están fuertemente regularizados por la estructura de vecindad impuesta por el ICAR.

Al comparar dos modelos, se favorece aquel con menor DIC. Como regla práctica, Spiegelhalter et al. (2002) sugieren que diferencias de DIC inferiores a 5 puntos no son concluyentes, mientras que diferencias superiores a 10 puntos constituyen evidencia sustancial a favor del modelo con menor valor. No obstante, el DIC puede comportarse de forma inestable cuando la distribución posterior está lejos de la normalidad. Lo que motiva al uso complementario del WAIC.

2.7.2. Criterio de información de Watanabe-Akaike (WAIC)

El (*Widely Applicable Information Criterion*, WAIC) (Watanabe, 2010) evita algunas de las limitaciones del DIC al construirse enteramente a partir de la distribución predictiva posterior, en lugar de depender de una estimación puntual de $\boldsymbol{\theta}$. Se define como:

$$\mathrm{WAIC} = -2 \left(\mathrm{lppd} - p_{WAIC} \right), \tag{2.22}$$

donde lppd (*log pointwise predictive density*) es la verosimilitud logarítmica predictiva puntual esperada bajo la distribución posterior,

$$\mathrm{lppd} = \sum_{i=1}^{n} \log \Big(\mathrm{E}_{\mathrm{post}} \big[p(z_i \mid \boldsymbol{\theta}) \big] \Big), \tag{2.23}$$

y p_{WAIC} es el número efectivo de parámetros. El WAIC es asintóticamente equivalente a la validación cruzada *leave-one-out* bayesiana y, al promediar sobre toda la distribución posterior en lugar de utilizar un único punto $\bar{\boldsymbol{\theta}}$, tiende a mostrar un comportamiento más estable que el DIC en modelos con estructuras jerárquicas complejas, como el BYM, donde la posterior de los efectos aleatorios U_i puede ser marcadamente asimétrica en municipios con pocos vecinos. Al igual que con el DIC, valores más reducidos de WAIC indican un ajuste superior una vez penalizada la complejidad.

En la práctica, lo ideal es reportar e interpretar el DIC y el WAIC de forma conjunta, en lugar de basar la selección de modelos en un único criterio. Dado que ambos penalizan la complejidad de manera distinta, el DIC a través de p_D(calculado en torno a la media posterior) y el WAIC a través de p_{WAIC} (calculado a partir de la varianza posterior punto a punto), la concordancia entre ambos en la dirección y en la magnitud aproximada de la mejora de un modelo respecto a otro constituye una forma sencilla de robustecer la conclusión de la comparación.

2.8. Presentación de resultados

Los resultados derivados del análisis de datos agregados por área suelen presentarse mediante una combinación de:

- **Figuras:** como **mapas coropléticos** y **diagramas de dispersión espacial**. Los mapas constituyen el recurso principal para reflejar la dimensión geográfica de los resultados, ya que permiten identificar visualmente clústeres, valores atípicos espaciales y patrones de distribución que no se aprecian en los estadísticos globales. Por su parte, los diagramas de dispersión de Moran representan la relación entre el valor de cada área y el retardo espacial de sus vecinas, mostrando la pendiente asociada al índice I de Moran global y, en consecuencia, poniendo de manifiesto la dependencia espacial.
- **Tablas:** contienen las estimaciones puntuales, los intervalos de confianza y los criterios de comparación de modelos, lo que facilita valorar de forma sistemática la significación y la magnitud de los efectos.

2.9. Herramientas de software

Entre las herramientas de software que pueden emplearse para llevar a cabo la metodología descrita se encuentran:

- **Python:** con el ecosistema `PySAL` y `geopandas` es una alternativa potente para el análisis espacial (Rey y Anselin, 2010).
- **GeoDa:** el software de escritorio, especialmente indicado para el ESDA interactivo, sin embargo no implementa modelos BYM o INLA (Anselin et al., 2006)
- **Stan o JAGS:** lenguajes probabilísticos orientados a la estimación de modelos bayesianos mediante MCMC.

Sin embargo, R constituye la opción más completa para este estudio: dispone de implementaciones consolidadas de todos los métodos utilizados (ESDA, LISA, BiLISA y modelos BYM mediante INLA), ofrece integración nativa con datos espaciales a través de `sf` y `terra`, y asegura una reproducibilidad plena gracias al uso de scripts y el establecimiento de una semilla fija.

2.9.1. Análisis con R

En lo que respecta a la elaboración de un script reproducible, en R deben contemplarse las siguientes fases (Moraga, 2023; Bivand, 2022):

1. Lectura e integración de los datos geoespaciales mediante `sf` y `terra`;
2. Construcción del vecindario y de las matrices de pesos con `spdep`;
3. Análisis exploratorio de datos espaciales (ESDA), que incluya mapas coropléticos, el índice I de Moran global y local (LISA y BiLISA);
4. Ajuste de modelos de referencia con `lm` y
5. Ajuste de modelos espaciales de tipo BYM con `R-INLA`.

Por último, las funciones que se evalúan mediante permutaciones deben ejecutarse con una semilla fija (`set.seed()`) para asegurar la reproducibilidad de los resultados. A modo de síntesis, el siguiente cuadro muestra los paquetes de R junto con su función principal correspondiente:

Cuadro 2.2: Paquetes de R a utilizar en el análisis.

Paquete	**Función principal**
`sf`	Manejo de datos vectoriales (polígonos)
`terra`	Lectura y procesado de rasters NetCDF
`exactextractr`	Interpolación areal raster a polígono
`mapSpain`	Límites administrativos de España
`spdep`	Vecindarios, matriz$\mathbf{W}$, I de Moran, LISA
`bispdep`	Índice de Moran bivariado y BiLISA
`INLA`	Modelos BYM mediante aproximación INLA
`ggplot2`	Visualización y mapas temáticos

Parte II

Caso de estudio

Capítulo 3

Preparación Espacial

En este capítulo se busca aplicar los conceptos y fundamentos teóricos explicados en el capítulo anterior para el preprocesado de datos de concentración de contaminantes y renta bruta per cápita para los municipios de Galicia. Como se mencionó en la introducción del trabajo, la contaminación del aire representa uno de los principales riesgos ambientales para la salud en Europa. En el contexto de Galicia, comunidad autónoma caracterizada por una marcada dualidad entre areas metropolitanas dinámicas (Vigo, A Coruña, Santiago de Compostela, Ourense) y extensas zonas rurales, la distribución espacial de los contaminantes atmosféricos puede reflejar patrones de desigualdad ambiental de notable relevancia para las políticas públicas.

A lo largo de este capítulo se presenta una descripción minuciosa de la región de estudio, así como del proceso de obtención, depuración y preprocesamiento de los datos y, por último, de la preparación espacial mediante la definición del vecindario y la construcción de la matriz de pesos espaciales. La base de datos resultante, el código empleado para la limpieza de los datos y el código en `R` utilizado para el ESDA que se explica en el capítulo siguiente se encuentran disponibles para su consulta en el repositorio de `GitHub` Areal-Data.

3.1. Preprocesado de datos

Como se mencionó, el objetivo de este proyecto es aplicar herramientas de estadística espacial para datos agregados por área con el fin de caracterizar la distribucion espacial de NO_2 y $PM_{2,5}$ en los municipios gallegos, detectar clústeres espaciales significativos, y evaluar la relación espacial entre concentración de contaminantes y renta bruta per cápita municipal.

3.1.1. Descripción de la región espacial: Galicia

Para realizar un ESDA, es imprescindible conocer previamente la región geográfica con la que se va a trabajar. Entender la geografía de Galicia y su contexto socioeconómico resultará útil para definir el tipo de relación entre los municipios y, en consecuencia, para construir el vecindario, las matrices de adyacencia y pesos que constituyen elementos clave para el análisis.

Situación geográfica

Galicia es una comunidad autónoma de España situada en el extremo noroeste de la península con una superficie de 29 577 km^2 y una población de 2.7 millones de habitantes aproximadamente (INE, 2025a). Al norte y al oeste limita con el océano Atlántico, al sur con Portugal, y al este con las comunidades autónomas de Asturias y Castilla y León. Su territorio se divide administrativamente en: cuatro provincias (A Coruña, Lugo, Ourense y Pontevedra) y a su vez, en 313 municipios (véase Figura

3.1), que constituyen la unidad de análisis del presente trabajo. Su capital es Sasntiago de Compostela, sin embargo, el municipio con mayor población es Vigo.

Figura 3.1: Comunidad Autónoma de Galicia con sus delimitaciones municipales, elaborado en R con paquetes *leaflet* y *mapSpain*.

Contexto socioeconómico

Geográficamente, Galicia presenta una marcada heterogeneidad espacial. Las provincias atlánticas de A Coruña y Pontevedra concentran las principales áreas metropolitanas de la comunidad con ciudades como Vigo, A Coruña, Santiago de Compostela y Ferrol, ciudades caracterizadas por su elevada densidad poblacional, actividad industrial y/o portuaria, e intensa movilidad de vehículos de motor. Por su parte, las provincias del interior, Lugo y Ourense, muestran un modelo territorial caracterizado por municipios rurales de reducido tamaño, con menores densidades de población, economías basadas en la agricultura y la ganadería, y un intenso proceso de envejecimiento demográfico (en el caso de Ourense) (INE, 2025b; INE, 2025a).

Esta dualidad territorial se refleja igualmente en los indicadores socioeconómicos. En términos de renta, Galicia se sitúa sistemáticamente por debajo de la media nacional española según los datos del Atlas de Distribución de Renta de los Hogares del Instituto Nacional de Estadística INE (2023). La renta media por municipio muestra una distribución espacialmente desigual, con niveles más altos en los entornos metropolitanos de Vigo y A Coruña y valores claramente más bajos en los municipios del interior de Lugo y Ourense, como se observa en la Figura 3.2. Aunque en dicha figura el eje de rentas no parte de cero, lo que acentúa visualmente la diferencia entre niveles de renta, la disparidad entre medias existe y refleja las diferencias socioeconómicas existentes dentro de la comunidad autónoma. Esta heterogeneidad territorial hace de Galicia un caso de estudio especialmente adecuado para el análisis de justicia ambiental, entendida como la distribución desigual de las cargas ambientales (en este caso la exposición a contaminantes atmosféricos) entre grupos socioeconómicamente diferenciados (Walker, 2012).

3.1.2. Dominio y Unidad de Análisis

Como se mencionó, la elección de Galicia como dominio de análisis corresponde con su heterogeneidad territorial. No obstante, también depende de la disponibilidad de información necesaria para aplicar las técnicas de análisis de datos agregados por áreas. Por este motivo, el presente análisis se

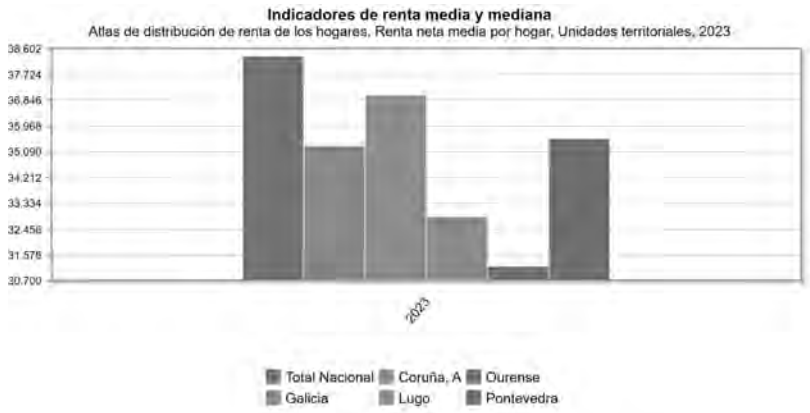


Figura 3.2: Renta media por hogar 2023: Comparativa del Total Nacional, C. A. de Galicia y sus provincias.

Referencia: Atlas de Distribución de Renta de los Hogares (INE, 2023)

ha podido realizar gracias a la disponibilidad de los datos socioeconómicos publicados por el Instituto Galego de Estatística. Los municipios, como unidad de análisis, representan la escala administrativa más desagregada para la cual el IGE publica datos de renta. Sin embargo, trabajar con datos agregados por área supone considerar que el valor de cada variable es uniforme dentro de cada municipio, lo que, especialmente en municipios de gran extensión, puede generar problemas inherentes como el MAUP. Por ello, es fundamental tomar en consideración estas limitaciones, tal como se indicó en la sección 2.1.1.

3.1.3. Descripción de los datos

Límites administrativos

Los polígonos de los municipios de Galicia se obtuvieron mediante el paquete `mapSpain` (Hernangómez, 2024) para el software estadístico R, que proporciona acceso programático a la cartografía oficial del Instituto Geográfico Nacional (IGN) y del servicio GISCO de Eurostat. Los límites se seleccionaron por el código NUTS2 de comunidad autónoma de Galicia ("$ES11$"), aunque también puede filtrarse con el código (`codauto == "12"`), y se transformaron al sistema de referencia de coordenadas WGS84 para su compatibilidad con los datos rasterizados de calidad del aire. Para el cálculo de la matriz de pesos espaciales, los límites se reproyectaron al sistema ETRS89 / UTM zona 29N (EPSG:25829), que es el sistema de referencia oficial para Galicia y expresa las coordenadas en metros, permitiendo el cálculo correcto de distancias euclídeas entre centroides (Pebesma y Bivand, 2023).

Datos de contaminación atmosférica

Las concentraciones de dióxido de nitrógeno (NO_2) y partículas finas ($PM_{2,5}$) se obtuvieron del *Copernicus Atmosphere Monitoring Service* (CAMS), específicamente del producto de reanálisis de calidad del aire europeo (*European Air Quality Reanalyses*) con el sistema de asimilación español MONARCH (Copernicus Atmosphere Monitoring Service, 2023). En la página web propia del CAMS European air quality reanalysis se seleccionaron los datos de la siguiente manera:

1. la o las **variable(s)**: en este caso (NO_2) y ($PM_{2,5}$),
2. **modelo de asimilación:** en este caso MONARCH,

3. **nivel:** altura en metros sobre la superficie, en este caso 100 m,
4. **tipo:** tipo de análisis, como se mencionó se escogió el reanálisis validado,
5. **año:** se seleccionó el 2023, que es la fecha más reciente que coincide con los datos socioeconómicos,
6. **mes:** se seleccionaron todos los meses del año.

Esta web proporciona campos tridimensionales de concentración de contaminantes en formato NetCDF (`.nc`), con una resolución espacial de $0{,}15° \times 0{,}15°$ ($203km^2$) y resolución temporal horaria.

Para el presente análisis se descargaron los archivos correspondientes a todos los meses del año 2023, desde enero hasta diciembre, dado que son los datos más recientes disponibles y coinciden con los datos disponibles de la variable socioeconómica. A partir de los datos horarios de cada mes se calculó la mediana mensual de todas las capas temporales del archivo. La mediana anual de cada contaminante se obtuvo posteriormente como promedio simple de las doce medianas mensuales resultando finalmente en **las variables continuas**: $pm25\_median$ y $no2\_median$. Los valores de concentración se expresan en microgramos por metro cúbico ($\mu g/m^3$) y se comparan con los valores de directrices anuales establecidos por la Organización Mundial de la Salud: 10 $\mu g/m^3$ para el NO_2 y 5 $\mu g/m^3$ para el $PM_{2,5}$ (OMS, 2021). Es importante mencionar, que la extracción de los valores rasterizados a nivel municipal se realizó mediante interpolación areal exacta con el paquete `exactextractr` (Baston, 2026), que pondera cada celda del raster por la fracción de su área que intersecta con el polígono municipal. La resolución espacial del producto CAMS-MONARCH de $203km^2$ es superior a la extensión de numerosos municipios del interior de Galicia, lo que constituye una limitación del análisis que debe ser considerada en la interpretación de los resultados.

Datos socioeconómicos: renta bruta per cápita en Galicia

La variable socioeconómica utilizada en el análisis es la *renta bruta per cápita* por municipio para el año 2023, procedente del Instituto Gallego de Estadística (Instituto Galego de Estatística, 2023). La renta bruta per cápita mide la suma de todos los ingresos brutos: rentas del trabajo y del capital, así como transferencias sociales como pensiones y prestaciones por desempleo; dividida entre el número de habitantes del municipio, por lo que constituye un indicador próximo al nivel de vida real de la población residente. Los datos se descargaron en formato CSV directamente desde el portal estadístico del IGE.

Al proceder de un organismo autonómico, los topónimos municipales están recogidos en su forma oficial en lengua gallega lo que reduce las inconsistencias de nomenclatura respecto a las fuentes nacionales del INE y facilita la correspondencia con los identificadores del paquete `mapSpain` (Hernangómez, 2024), que también incorporan los nombres en lengua oficial gallega. No obstante, se aplicó igualmente una etapa de emparejamiento: eliminación de tildes y conversión a minúsculas, para garantizar la robustez del *join* por el nombre en común. La variable continua obtenida se nombró *renta_pc*, sin embargo, al presentar valores en magnitud elevada se procedió a realizar una transformación logarítmica resultando en la variable *renta_log* la cual será utilizada en el análisis del siguiente capítulo.

3.2. Preparación Espacial

3.2.1. Construcción del vecindario

Siguiendo los criterios descritos en la sección 2.3.2, se optó por el criterio de **contigüidad *Queen*** para definir el vecindario entre los 313 municipios gallegos. Esta elección se fundamenta en dos motivos complementarios. En primer lugar, la contigüidad *Queen* es el criterio más adecuado para polígonos administrativos irregulares, ya que considera vecinos a todas las áreas que comparten al menos un punto de su frontera, incluyendo vértices, reduciendo así el riesgo de generar áreas aisladas. En segundo lugar, la contigüidad *Queen* refleja mejor la naturaleza del fenómeno estudiado: la dispersión de contaminantes

atmosféricos no se limita a las fronteras compartidas en línea recta sino que se produce en todas las direcciones (Bivand et al., 2013).

La inspección del vecindario resultante identificó un municipio aislado **La Isla de Arousa**, una pequeña isla cuya conexión con los municipios continentales se produce únicamente a través de una carretera que no genera contigüidad poligonal en los datos cartográficos. Para garantizar la conectividad global del grafo de vecindad, condición necesaria para los modelos ICAR/BYM (Bivand et al., 2013), se asignó a La Isla de Arousa el vecino más próximo por distancia euclídea entre centroides mediante la función `addlinks1()` del paquete `spdep`. El municipio resultó conectado con Cambados, en primer lugar, y Vilanova de Arousa en segundo lo que resulta geográfica y contextualmente apropiado dado que ambos comparten el mismo ámbito socioeconómico del seno de Arousa. Aunque Vilanova de Arousa parezca estar más cercano a la isla, debido a la ubicación del centroide en Vilanova de Arousa se define a Cambados como el primer vecino más cercano.

Como alternativa exploratoria se construyó también un vecindario basado en los **5 vecinos más cercanos** (KNN, $k = 5$). Este tipo de vecindarios resulta útil para el cálculo de la matriz de pesos por inversa de la distancia (IDW), sin embargo, para este trabajo se realizó con propósito meramente comparativo. Ambas alternativas se pueden observar en la Figura 3.3. No obstante, el vecindario *Queen* ajustado se utilizó como estructura de referencia para todos los análisis de autocorrelación espacial y modelización, por su mayor coherencia con la contigüidad administrativa.

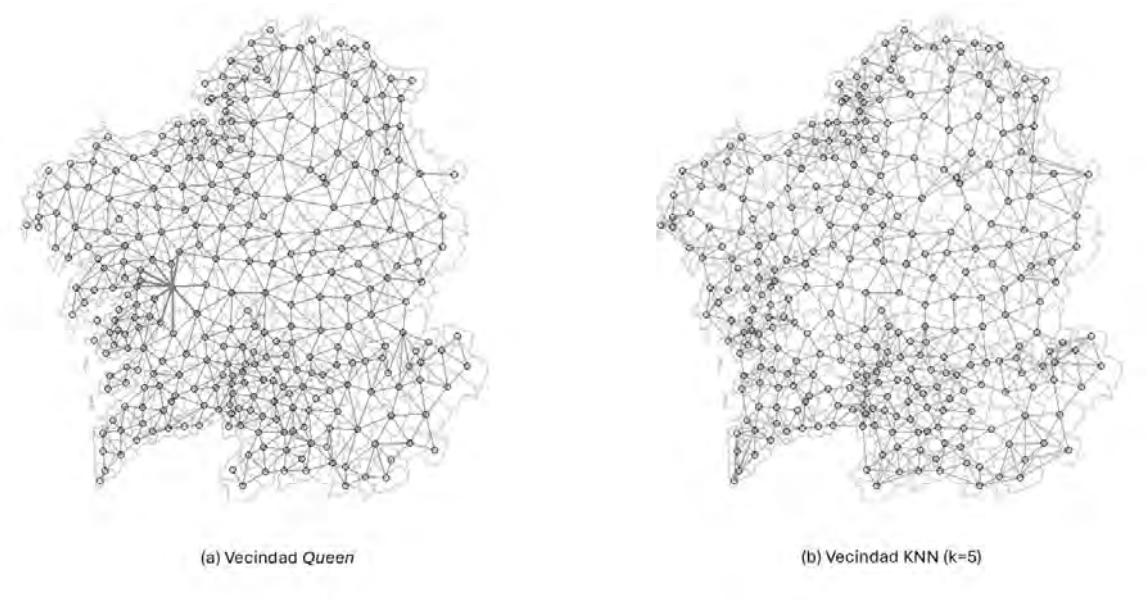


Figura 3.3: Comparativa de grafo creado por vecindad tipo *Queen* (a) mostrando con líneas rojas el municipio con mayor cantidad de vecinos y k=5 vecinos más cercanos (b).

3.2.2. Matriz de pesos espaciales

A partir del vecindario *Queen* ajustado se construyeron las siguientes matrices de pesos:

- **Matriz normalizada por filas** ($\mathbf{W_q}$, estilo W): utilizada en el ESDA y en los modelos de retardo espacial. Cada fila suma la unidad, de modo que el retardo espacial $\sum_j w_{ij}^* z_j$ se interpreta como la media ponderada de los vecinos de i.
- **Matriz binaria no normalizada** ($\mathbf{W_b}$, estilo B): utilizada en la formulación de los modelos ICAR/BYM, donde la estructura de vecindad se incorpora directamente a través de la matriz de

precisión $\mathbf{Q} = \frac{1}{\sigma^2}(\mathbf{D} - \mathbf{W_b})$ (Besag et al., 1991). Para su uso en `R-INLA` se convirtió al formato matricial disperso `CsparseMatrix`.

El número medio de vecinos en el vecindario *Queen* ajustado fue de 5.28 (mínimo: 1, máximo: 11): siete municipios con sólo un vecino y un municipio, La Estrada, con 11 vecinos como se observa en la Figura 3.3; lo que refleja la heterogeneidad de tamaño y forma de los municipios gallegos.

Como alternativa conceptual, dado que la dispersión de contaminantes atmosféricos es un proceso físicamente continuo, se exploró también una matriz de pesos por inversa de la distancia (IDW). Sin embargo, esta especificación no es directamente compatible con la estructura dispersa y simétrica que requiere el componente ICAR del modelo BYM (Sección 2.6.1), y su uso exigiría truncar la matriz en un grafo local, lo que equivaldría a un criterio de distancia umbral o KNN. Además, la variable renta no es un proceso de difusión físico, es decir, no se dispersa espacialmente como un contaminante (decaimiento geográfico) sino que responde a mercados laborales, infraestructura, densidad poblacional, es decir, fenómenos delimitados por fronteras administrativas. Adicionalmente, dado que la resolución del producto CAMS-MONARCH ($\sim$203 km^2) ya introduce un suavizado espacial explícito comparable a la extensión de muchos municipios gallegos, la contigüidad *Queen* se consideró la especificación metodológicamente coherente más entre el ESDA y la modelización posterior.

Capítulo 4

Análisis

En este capítulo se lleva a cabo la aplicación de los métodos expuestos en la Parte I al conjunto de datos correspondiente a los 313 municipios de Galicia, empleando el software `RStudio` como herramienta para el análisis estadístico y la elaboración de gráficos, junto con las librerías indicadas en la Sección 2.9.1. El análisis se estructura en tres bloques: el ESDA con los mapas coropléticos, la autocorrelación espacial y el análisis bivariado; los modelos de referencia no espaciales; y los modelos espaciales BYM con diagnóstico y comparación.

4.1. Análisis Exploratorio de Datos Espaciales

4.1.1. Distribución espacial de las variables

La Figura 4.1 muestra la distribución espacial de las medianas anuales de $PM_{2,5}$ y NO_2, y de la renta bruta per cápita en los 313 municipios gallegos para el año 2023. Los estadísticos descriptivos básicos se recogen en el diagrama de la Figura 4.2 y detalladamente en el Cuadro 1 .

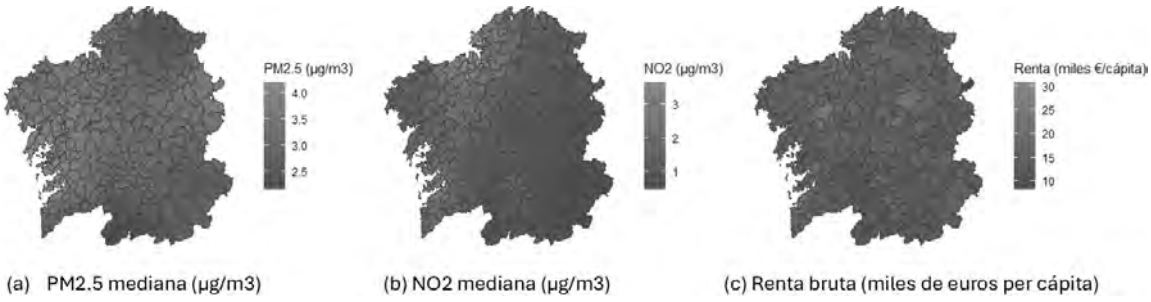


Figura 4.1: Distribución espacial de las variables (a) *pm25_median*, (b) *no2_median* y (c) *renta_pc* mostrados en sus respectivas unidades. Datos agregados por municipio.

Los mapas coropléticos revelan patrones espaciales diferenciados para los tres contaminantes. Las concentraciones más elevadas de $PM_{2,5}$ se observan principalmente, en las regiones costeras cerca de Vigo, Santiago de Compostela y La Coruña, mientras que los valores más bajos se concentran en los municipios del interior de las provincias de Ourense y de Lugo (al noreste). El patrón del NO_2 presenta una distribución más concentrada en los entornos urbanos y periurbanos de las principales ciudades gallegas, incluyendo Ourense (ciudad), coherente con su origen predominantemente vinculado al tráfico motorizado. Por su parte, la renta bruta per cápita muestra una distribución con un patrón menos marcado, mostrando solo valores altos puntuales en las ciudades más habitadas.

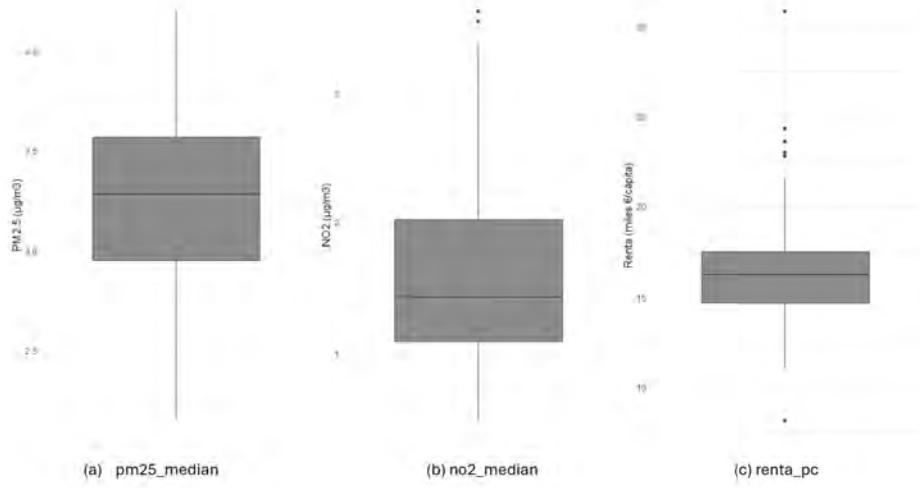


Figura 4.2: Gráfico de caja y bigotes de las variables *pm*25*_median*, *no*2*_median* y *renta_pc* mostrados en sus respectivas unidades.

Aunque el nivel de exposición no es homogéneo en el territorio gallego, es importante mencionar, que las concentraciones de ambos contaminantes se sitúan por debajo de los valores límite anuales establecidos por la Organización Mundial de la Salud (10 µg/m^3 para NO_2 y 5 µg/m^3 para $PM_{2,5}$) (OMS, 2021).

4.1.2. Correlación no espacial

Como parte del análisis exploratorio es importante conocer cómo se relacionan las variables de estudio. Para ello, se utiliza la función `ggpairs` de la librería `GGally`.

Como se puede observar en la Figura 4.3 las correlaciones cruzadas de las variables son todas positivas, indicando que al incrementar una variable incrementa la otra, para todas las variables. Es importante remarcar que existe una correlación positiva entre ambos contaminantes. Esto se debe a que el NO_2 es un precursor de las $PM_{2,5}$ (Ministerio para la Transición Ecológica y el Reto Demográfico, s.f.). Sin embargo, las directrices establecidas por la OMS (OMS, 2021) y la carga ambiental, así como enfermedades atribuidas (Soares et al., 2024) se presentan para cada contaminante, por lo que se considerarán ambas de forma separada.

Por otro lado, considerando que el objetivo del caso de estudio es identificar si áreas socioeconómicamente vulnerables pueden hacer frente a la carga atribuida a los contaminantes aéreos, para propósitos de este estudio, no se considera la correlación entre contaminantes, sino que se evalúan por separado en función de la variable de renta como se verá más adelante en la modelización.

4.1.3. Autocorrelación Espacial Global

El índice I de Moran se calculó para las tres variables bajo la matriz de pesos $\mathbf{W_q}$ (contigüidad *Queen* normalizada por filas) y se evaluó su significación mediante dos métodos: la aproximación normal bajo el supuesto de aleatorización y la inferencia por permutación con 999 simulaciones Monte Carlo. Los resultados se resumen en el Cuadro 4.1.

Para las tres variables se rechaza la hipótesis nula de ausencia de autocorrelación espacial global ($H_0 : I = \mathrm{E}_0[I]$) con niveles de significación inferiores a $\alpha = 0{,}05$, tanto bajo la aproximación normal como mediante inferencia por permutaciones. Los valores positivos del índice indican que para las tres

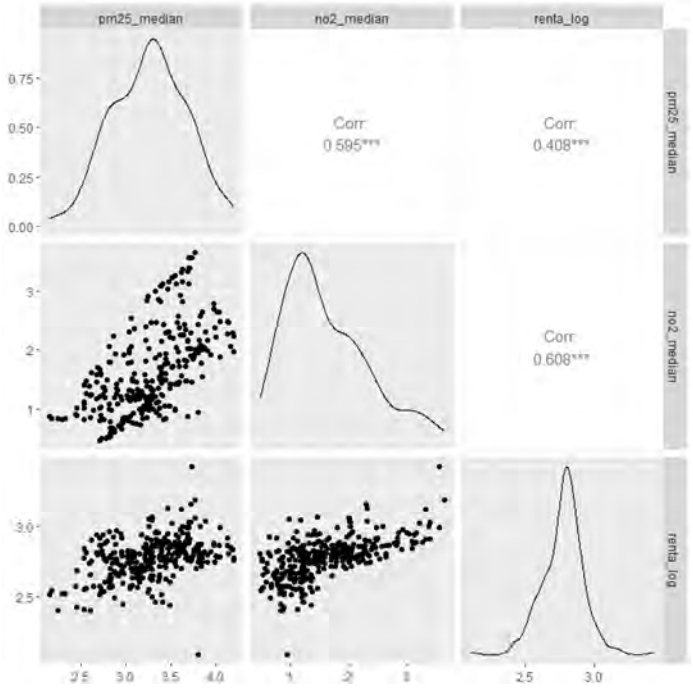


Figura 4.3: Correlación no espacial para las variables *pm*25*_median*, *no*2*_median* y *renta_log*. En los páneles superiores derechos se incluyen los coeficientes de correlación cruzados, en la diagonal del gráfico las densidades de cada variable, y en los páneles inferiores izquierdos los diagramas de dispersión cruzados.

variables, los municipios presentan valores similares entre sí. Es decir, municipios con concentraciones elevadas o bajas de contaminantes tienden a estar rodeados de municipios con valores parecidos, y es el mismo caso para la renta. Esto confirma la presencia de un patrón de agrupamiento espacial acorde con la Primera Ley de Geografía (Tobler, 1970). Este comportamiento se aprecia en los diagramas de dispersión de Moran de la Figura 1, donde se observa una tendencia positiva en los tres casos.

4.1.4. Autocorrelación espacial local: LISA

El índice local de Moran I_i se calculó para las tres variables mediante permutación con 999 simulaciones (`localmoran_perm()`), y los municipios con $p - valor < 0.05$ se clasificaron en los cuatro cuadrantes definidos en la Sección 2.4.3. Es importante mencionar que para asegurar la reproducibilidad de los resultados, se fijó la semilla (`set.seed(27750018)`) antes de cada simulación de datos. Los resultados de la clasificación se recogen en el Cuadro 4.2 y se representan geográficamente en la Figura 4.4.

El clúster Alto-Alto de $PM_{2,5}$ se localiza principalmente en áreas urbanas como Vigo, Santiago de Compostela y La Coruña, así como en algunos municipios con menor densidad de población, como Noia. En cambio, los clústeres Bajo-Bajo se observan sobre todo en las provincias de Lugo y Ourense. En el caso del NO_2, el patrón refleja una mayor presencia de clústeres Alto-Alto de forma puntual en núcleos urbanos, coherente con la distribución del tráfico motorizado en ciudades como Vigo, Santiago de Compostela, La Coruña, Ferrol y sus áreas próximas. Por su parte, el clúster Bajo-Bajo se agrupa en municipios del interior situados en el sureste de la comunidad. Asimismo, la ausencia de outliers espaciales (cuadrantes Alto-Bajo y Bajo-Alto), que identificarían municipios con niveles de contaminación discordantes respecto a su entorno inmediato, indica que no se detectan fuentes o sumideros

Cuadro 4.1: Índice I de Moran global para $PM_{2,5}$, NO_2 y logarítmo de renta. Vecindario *Queen*, $n = 313$, 999 permutaciones.

Variable	I	p**-valor (MC)**	**Conclusión**
*pm*25_*median*	0.9311	0.0000	Autocorrelación positiva
*no*2_*median*	0.9469	0.0000	Autocorrelación positiva
renta_log	0.5164	0.0000	Autocorrelación positiva

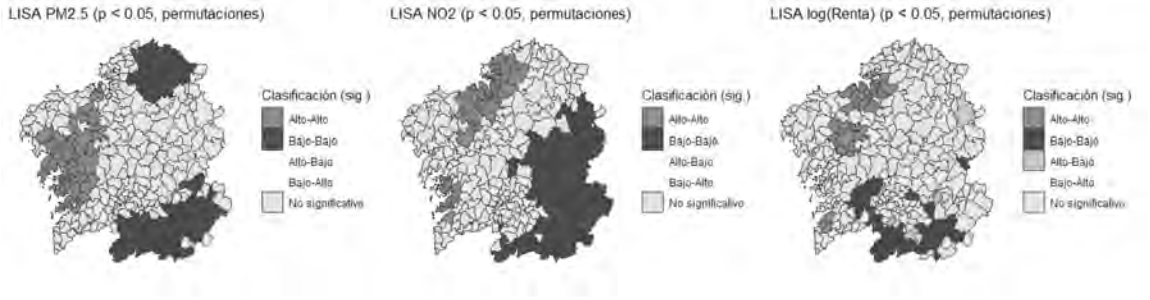


Figura 4.4: Mapas de clústeres LISA significativos ($p - valor < 0.05$) para $PM_{2,5}$, NO_2, y logaritmo de renta, respectivamente.

locales de emisión.

Por su parte, en la renta, sí es posible identificar outliers espaciales. Los municipios de La Fonsagrada (111), El Carballiño (179), Castro Caldelas (183), Celanova (184) y Xinzo de Limia (192) destacan como municipios con niveles de renta elevados rodeados por municipios de renta baja. A su vez, los tres clústeres Alto-Alto se localizan principalmente en la provincia de La Coruña, junto con un pequeño clúster en la provincia de Pontevedra. Por otro lado, los clústeres de valores bajos se concentran sobre todo en el sur de Galicia.

4.1.5. Análisis bivariado: BiLISA

Para contrastar directamente la hipótesis de justicia ambiental se realizó el análisis BiLISA entre los contaminantes ($PM_{2,5}$ y NO_2) como variables principales y la transformación logarítmica de la renta como variable secundaria. Es decir, se analizaron los valores de las variables de contaminación frente a los retardos de la variable *renta_log*, siguiendo la formulación de la Sección 2.4.4. Los índices locales bivariados se evaluaron mediante 999 permutaciones con la función `localmoran.bi()` del paquete `bispdep` para identificar cuáles de ellos resultan significativos.

En este caso de estudio, **si** $I^{BV} > 0$, un índice bivariado positivo indicará que los municipios con alta concentración de contaminante tienden a estar rodeados de municipios con renta elevada, y viceversa, los municipios con contaminación baja rodeados de municipios con renta baja. Este resultado indicaría que la carga ambiental derivada de la concentración de contaminantes se corresponde con el nivel socioeconómico de los municipios (alto con alto, bajo con bajo) y refleja un escenario ideal en el que las condiciones socioeconómicas permiten afrontar adecuadamente dicha carga ambiental.

Por otro lado, **si** $I^{BV} < 0$, índice bivariado negativo, indicaría que existen municipios con contaminación alta y renta baja (evidencia de injusticia ambiental) o contaminación baja y renta alta (situación favorable). Señalando una distribución desigual de la carga ambiental, ya sea en detrimento

Cuadro 4.2: Formación de Clústeres: Clasificación LISA significativa ($p-valor < 0.05$, permutaciones) en municipios para $PM_{2,5}$, NO_2, y logaritmo de renta.

Cuadrante	**$PM_{2,5}$**	**NO_2**	**Renta (log)**	**Interpretación**
Alto-Alto	1 (58 municipios)	2 (50 municipios)	3 (32 municipios)	Clúster de valores elevados
Bajo-Bajo	2 (64 municipios)	1 (72 municipios)	2 (36 municipios)	Clúster de valores bajos
Alto-Bajo	0	0	5 (5 municipios)	Outlier espacial positivo
Bajo-Alto	0	0	0	Outlier espacial negativo

de las poblaciones más vulnerables o favorable para las poblaciones con un nivel socioeconómico alto.

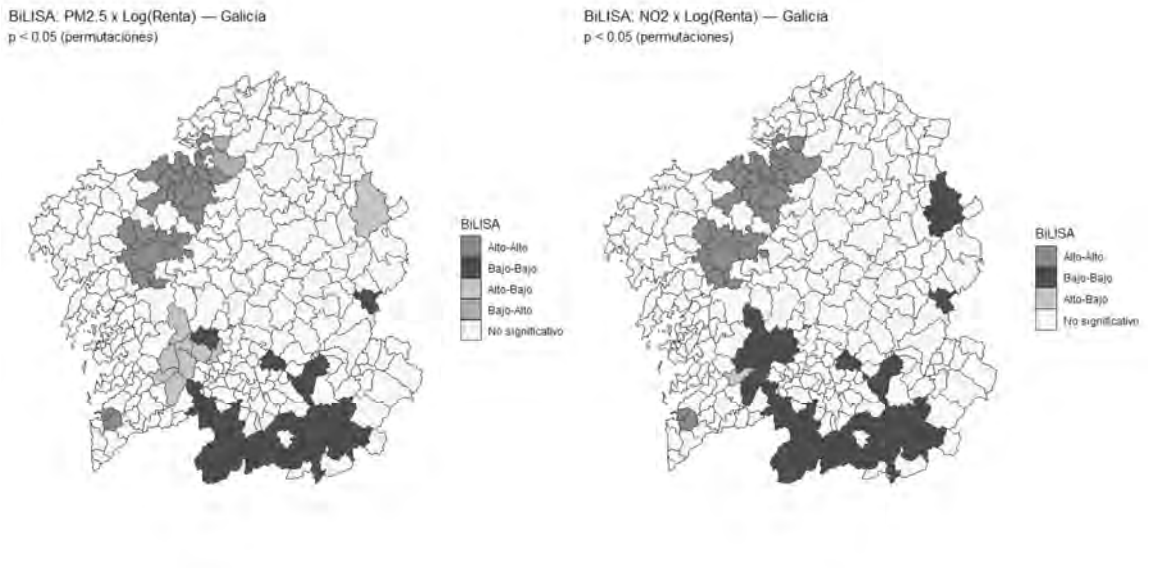


Figura 4.5: Mapas de clústeres BiLISA significativos ($p - valor < 0.05$) para $PM_{2,5}$, NO_2, frente a retardos del logaritmo de renta.

Los mapas BiLISA (Figura 4.5) ayudan a identificar los municipios con asociaciones espaciales bivariadas significativas ($p < 0{,}05$).

El cuadrante de mayor interés para la hipótesis de justicia ambiental es el **Alto-Bajo**: municipios con contaminación propia por encima de la media, rodeados de vecinos con renta por debajo de la media. Es importante mencionar que para el análisis BiLISA se utilizaron las variables de forma estandarizada ya que la variable *renta_log* y las variables de contaminantes tienen unidades diferentes. Para la contaminación por $PM_{2,5}$, los municipios clasificados en este cuadrante fueron nueve (véase Cuadro 4.3): La Fonsagrada, Avión, Beariz, Boborás, El Carballiño, Covelo, Forcarei, Fornelos de Montes, y La Lama. Para NO_2 solo Fornelos de Montes está dentro de esta clasificación. Estos municipios pertenecen principalmente a las provincias de Ourense y Pontevedra, y un municipio en Lugo, véase el Cuadro 2 para una lista detallada. Estos municipios no pertenecen a las principales ciudades de Galicia, tienen una renta baja y contaminación alta respecto a los valores de la comunidad. Aunque los valores de contaminación no sobrepasan las directrices señaladas por la OMS (OMS, 2021) y no resultan significativos, véase Figura 4.4, vale la pena resaltar que estos municipios caen dentro de un

Cuadro 4.3: Formación de Clústeres: Clasificación BiLISA significativa ($p - valor < 0.05$, permutaciones) en municipios para contaminantes $PM_{2,5}$ y NO_2 como variables primarias, y *renta_log* como variable secundaria.

Cuadrante	**$PM_{2,5}$**	**NO_2**	**Interpretación**
Alto-Alto	3 (29 municipios)	3 (32 municipios)	Clúster de valores elevados
Bajo-Bajo	5 (34 municipios)	5 (42 municipios)	Clúster de valores bajos
Alto-Bajo	2 (9 municipios)	1(1 municipio)	Distribución desigual: injusticia ambiental
Bajo-Alto	1 (2 municipios)	0	Distribución desigual: favorable

sector socioeconómicamente vulnerable al tener valores LISA significativamente bajos para *renta_log*. En este caso, es de alta importancia evitar generar una carga ambiental elevada, principalmente de $PM_{2,5}$, en estos municipios para evitar caer en un caso de injusticia ambiental.

El cuadrante **Bajo-Alto**, que identifica la situación ambientalmente favorable, agrupa los municipios de Monfero y La Capela respecto a los niveles de $PM_{2,5}$; para NO_2 no se identifican municipios significativos en esta clasificación.

4.2. Modelización

4.2.1. Modelos de referencia no espaciales

Como modelo de referencia se ajustó un modelo lineal mediante mínimos cuadrados ordinarios (OLS) para cada contaminante. Se utilizó la variable logarítmo de la renta (*renta_log*) como variable explicativa y como respuesta cada contaminante: *no2_median* y *pm25_median*. Este modelo lineal equivalente al modelo gaussiano de `R-INLA` sin componente espacial:

$$pm25\_median_i = \beta_0 + \beta_1 renta\_log_i + \varepsilon_i, \tag{4.1}$$

$$no2\_median_i = \beta_0 + \beta_1 renta\_log_i + \varepsilon_i. \tag{4.2}$$

Los resultados se resumen en el Cuadro 4.4.

Cuadro 4.4: Resultados de los modelos lineales de referencia para *pm25_median* y *no2_median* en función de *renta_log*.

	pm25_median		*no2_median*	
Parámetro	**Estimación**	**p-valor**	**Estimación**	**p-valor**
$\hat{\beta}_0$ (intercepto)	0.0633	0.876	-6.4081	$<2\text{x}10^{-16}$
$\hat{\beta}_1$ (*renta_log*)	1.1531	$5.27\text{x}10^{-14}$	2.8839	$<2\text{x}10^{-16}$
R^2 ajustado	0.164		0.3681	

En ambos modelos el coeficiente de *renta_log*, $\hat{\beta}_1$, resultó positivo y estadísticamente significativo, indicando que, en promedio, los municipios con mayor renta per cápita presentan concentraciones

más altas de contaminantes. Para $PM_{2,5}$, el coeficiente estimado fue $\hat{\beta}_1 = 1.1531$, mientras que para el NO_2 fue $\hat{\beta}_1 = 2.8839$. Este resultado, aparentemente contraintuitivo desde la perspectiva de la justicia ambiental, es el comportamiento esperado y se explica por la naturaleza de los contaminantes estudiados: tanto el NO_2 como el $PM_{2,5}$ en Galicia están asociados principalmente a la actividad industrial, el tráfico y la densidad poblacional, que son más elevados en los municipios metropolitanos de mayor renta (Hewitt et al., 2024). Adicionalmente, se observa que los coeficientes de determinación ajustados son muy bajos en ambos modelos indicando un ajuste pobre y que existe variabilidad que no está siendo explicada con los modelos ajustados. En este punto se aconseja evaluar los residuos de los modelos para identificar si existe autocorrelación espacial que pueda proporcionar información sobre la variabilidad no explicada.

Análisis de residuos

Al utilizar el test (`moran.mc()`) de Moran por permutación para evaluar la autocorrelación en los residuos del modelo lineal se reveló persistencia de autocorrelación espacial significativa como se observa en el Cuadro 4.5, lo que invalida los supuestos de independencia del modelo y justifica el paso a los modelos espaciales. Es importante mencionar que para realizar el test de Moran se especifica la estructura de vecindad a través de la matriz de pesos calculada tipo *Queen* ($\mathbf{W_q}$).

Cuadro 4.5: Test de Moran sobre residuos del modelo lineal. Simulación Monte Carlo con 999 permutaciones.

Modelo	I **(residuos)**	p**-valor**	**Conclusión**
LM $PM_{2,5}$	0.8360	$<2.2\text{x}10^{-16}$	Se rechaza H_0
LM NO_2	0.7220	$<2.2\text{x}10^{-16}$	Se rechaza H_0

Los valores del índice I sobre los residuos indican autocorrelación positiva para ambos modelos lineales (LM) y con $p-$valores significativos se puede rechazar la hipótesis nula de ausencia de autocorrelación espacial. De esta forma, se comprueba que el modelo lineal no consigue capturar la dependencia espacial presente en los datos. Esta dependencia se puede visualizar al graficar los residuos en el mapa como se observa en la Figura 4.6 en donde se aprecian, claramente, zonas de áreas que con residuos positivos y otras negativos en lugar de mostrar un mapa con residuos aleatorios.

4.2.2. Modelos espaciales areales

Una vez identificada la dependencia espacial en los residuos se ajustaron modelos BYM para cada contaminante mediante `R-INLA` con la especificación `besagproper`, que garantiza un ajuste estable respecto de la componente espacial estructurada $\mathbf{U}$ (Rue et al., 2009) y utilizando una matriz tipo *Queen* binaria esparcida utilizando `CsparseMatrix`. La fórmula del modelo para ambos contaminantes resulta de la forma:

$$Z_i = \beta_0 + \beta_1 renta\_log_i + U_i + V_i, \quad U_i \mid \mathbf{U}_{-i} \sim \text{ICAR}, \quad V_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_V^2). \tag{4.3}$$

Los coeficientes estimados por los modelos BYM se recogen en el Cuadro 4.6, junto con los valores de los modelos lineales. Debido a que los criterios de información DIC y WAIC exigen un marco de inferencia bayesiano y `lm()` no los proporciona, el modelo de referencia lineal se reajustó también en `R-INLA` con la misma estructura de media pero sin componentes espaciales, empleando la misma familia gaussiana:

$$Z_i \mid \beta_0, \beta_1, \sigma^2 \sim \mathcal{N}(\mu_i, \sigma^2), \qquad \mu_i = \beta_0 + \beta_1\, renta\_log_i, \tag{4.4}$$

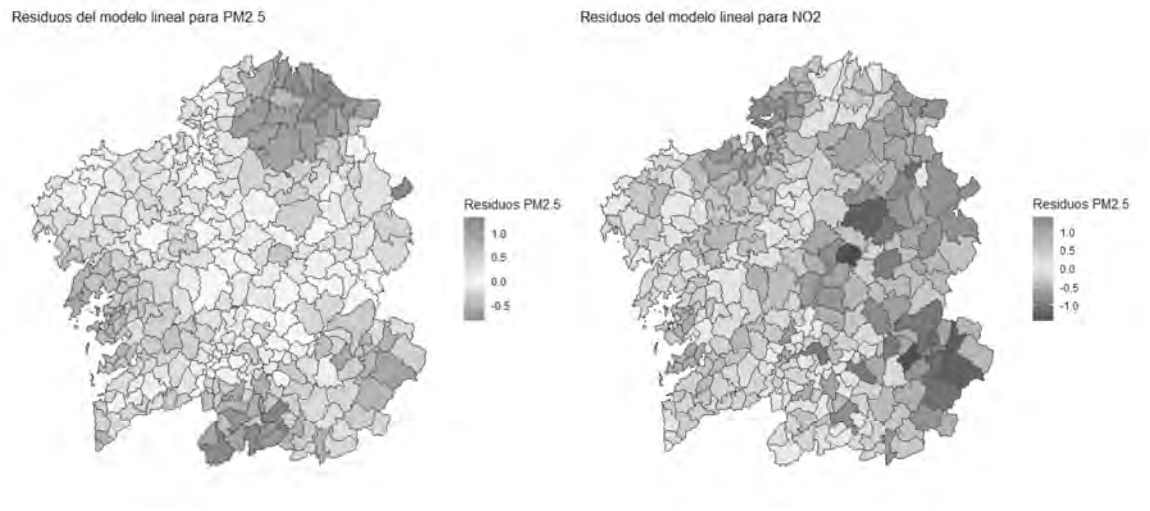


Figura 4.6: Mapas de residuos para modelos lineales de variable respuesta $PM_{2,5}$ y NO_2 según la variable explicativa *renta_log*.

Bajo este modelo, la media posterior de β_0 y β_1 coincide numéricamente con el estimador de mínimos cuadrados ordinarios (Cuadro 4.4). Adicionalmente, el ajuste bayesiano permite calcular el DIC y el WAIC bajo el mismo marco de verosimilitud utilizado para el modelo BYM (ecuación 4.3), garantizando así una comparación de modelos coherente. En el código, este modelo se especifica añadiendo `control.compute = list(dic = TRUE, waic = TRUE)` a la llamada de `inla()`, sin incluir el término espacial `f(id, model = "besagproper", graph = W_inla)` presente en la ecuación 4.3.

Cuadro 4.6: Comparación de coeficientes entre modelos lineales y modelos BYM (INLA) para $PM_{2,5}$ y NO_2. Intervalos de confianza al 95 % para BYM.

	*pm*25_*median*		*no*2_*median*	
Parámetro	**LM-INLA**	**BYM (IC 95 %)**	**LM-INLA**	**BYM (IC 95 %)**
$\hat{\beta}_0$	0.063	2.885 (1.999, 3.776)	-6.408	0.529 (-0.911, 1.981)
$\hat{\beta}_1$ (*renta_log*)	1.153	0.136 (0.024, 0.249)	2.884	0.384 (0.209, 0.559)

La incorporación de las componentes espaciales en el modelo BYM produjo una reducción sustancial del coeficiente de la variable *renta_log* en ambos modelos: de 1.153 a 0.136 para las partículas finas $PM_{2,5}$ y de 2.884 a 0.384 para el NO_2. Sin embargo, ambos coeficientes permanecen estadísticamente significativos (el intervalo de confianza al 95 % no incluye el cero), lo que indica que la asociación positiva entre renta y concentración de contaminantes es real aunque más modesta de lo que sugerían los modelos lineales. Esta reducción se explica porque la componente estructurada $\hat{U}_i$ absorbe la variabilidad espacial compartida entre municipios vecinos que en el modelo lineal era atribuida a la renta.

4.3. Validación y diagnóstico

Análisis de componentes

Adicionalmente, para entender cómo se comportan las componentes U y V en los modelos BYM es posible descomponer la varianza residual total en la proporción atribuible a la componente espacial estructurada frente a la componente no estructurada, mediante el parámetro:

$$\phi = \frac{\sigma_U^2}{\sigma_U^2 + \sigma_V^2}, \tag{4.5}$$

donde un valor de ϕ próximo a uno indica que casi toda la variabilidad residual responde a un patrón con estructura espacial, mientras que un valor de ϕ cercano a 0 señalaría que prevalece el ruido no estructurado, independiente entre municipios. Los valores estimados de ϕ para los dos modelos ajustados se presentan en el Cuadro 4.7.

Cuadro 4.7: Proporción de varianza espacial estructurada (ϕ) en los modelos BYM.

Modelo	σ_U^2	σ_V^2	$\hat{\phi}$	**IC 95 %**
BYM $PM_{2,5}$	0.0516	5.9589x10^{-5}	0.9988	(0.9958, 0.9998)
BYM NO_2	0.1253	8.1546 x10^{-5}	0.9993	(0.9972, 0.9999)

Como se puede observar, un valor de $\hat{\phi}$ cercano a uno en ambos contaminantes confirma que la variabilidad residual no explicada por la renta es, fundamentalmente, de carácter espacial y no ruido independiente por municipio. Este resultado es esperado debido a que la contaminación se comporta principalmente como un proceso de difusión regional, por la meteorología compartida, el flujo de tráfico, industrias, entre otros. Con excepción de que existan fuentes puntuales aisladas en municipios concretos, o bien, ruido causado por la fuente de medición o agregación. No obstante, para verificar que la componente espacial estructurada U captura adecuadamente la dependencia espacial y que V recoge únicamente el ruido restante, se somete la componente no estructurada V a un contraste `moran.mc`, cuyos resultados se presentan en el Cuadro 4.8.

Cuadro 4.8: Test de Moran sobre componente espacial no estructurada V. Simulación Monte Carlo con 999 permutaciones.

Modelo	I **(V)**	p**-valor**	**Conclusión**
BYM $PM_{2,5}$	0.0272	0,386	No se rechaza H_0
BYM NO_2	0.0458	0,208	No se rechaza H_0

A partir de los p-valores indicados en el test de Moran para contrastar la H_0: ausencia de autocorrelación espacial, se observa que, en ambos casos, el componente espacial no estructurado V deja de ser significativo lo que lleva a no rechazar la hipótesis nula de ausencia de autocorrelación espacial indicando que U, la componente espacial estructurada, recoge correctamente el patrón espacial. Este comportamiento se puede observar claramente en la Figura 4.7 en la cual los mapas de la componente espacial estructurada muestran la variación compartida entre municipios vecinos, resultando en un patrón suave, con gradiente regional, mientras que, en los mapas para la componente no estructurada se muestra la variación específica de cada municipio. En este caso, dado que el valor de $\hat{\phi}$ es tan cercano a uno, indicando que la variabilidad residual es casi completamente espacialmente estructurada

(vecinos se parecen a vecinos), y una vez que U recoge el patrón espacial, lo restante, V solo recoge el ruido local, es decir, las diferencias locales que existen entre municipios.

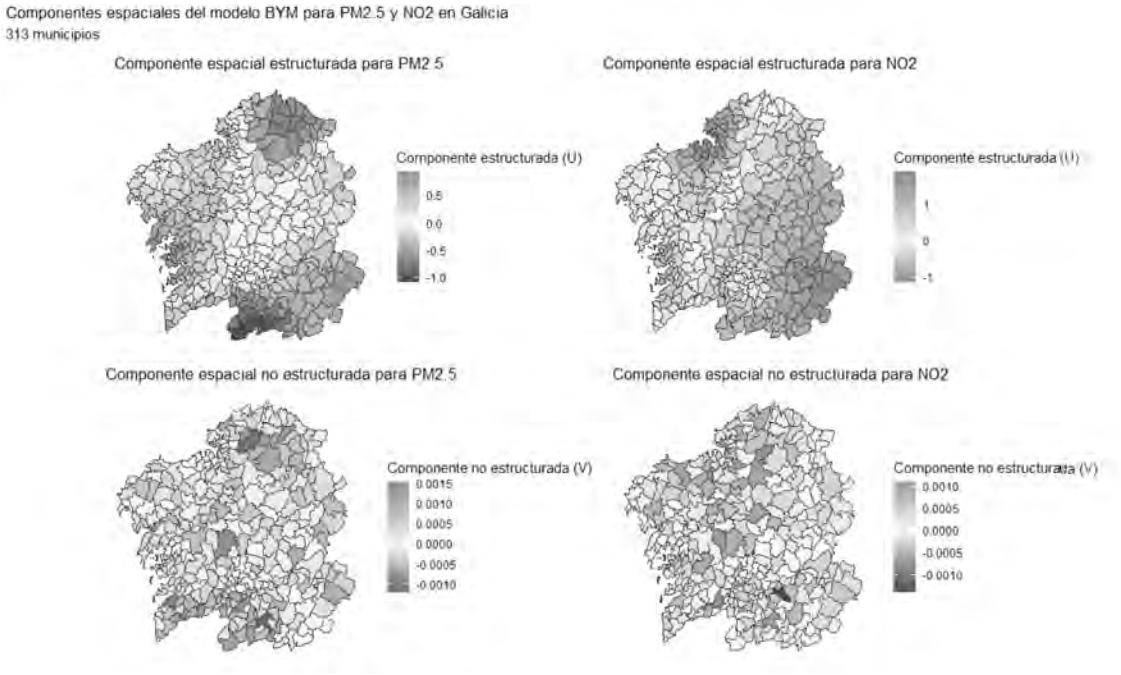


Figura 4.7: Mapas de las componentes espaciales estructuradas (arriba) y no estructuras (abajo) para los modelos BYM $PM_{2,5}$ (derecha) y NO_2 (izquierda) según la variable explicativa *renta_log*.

La combinación de un $\hat{\phi}$ cercano a uno con V sin autocorrelación confirma que el reparto entre componentes es coherente y que el modelo BYM está correctamente especificado para los datos. Es decir, otras combinaciones con $\hat{\phi}$ y V podrían indicar que U y V compiten por la varianza. En ese caso incluso valdría la pena investigar e incluir variables adicionales en el modelo, en lugar de solo contemplar la dependencia espacial.

4.3.1. Análisis de sensibilidad al criterio de vecindario

Dado que la definición del vecindario condiciona la estructura de dependencia asumida (Sección 2.3.2), se evaluó la sensibilidad de los resultados principales frente al criterio alternativo de k **vecinos más cercanos** ($k = 5$), construido en la Sección 3.2.1 con fines comparativos. Para ello se reajustaron los modelos BYM de la ecuación 4.3 sustituyendo la matriz binaria $\mathbf{W_b}$ (*Queen*) por la matriz binaria correspondiente al vecindario KNN, manteniendo el resto de las especificaciones sin cambios.

Cuadro 4.9: Comparación del coeficiente $\hat{\beta}_1$ (*renta_log*) bajo los criterios de vecindario *Queen* y KNN ($k = 5$).

Modelo	$\hat{\beta}_1$ **(*Queen*)**	$\hat{\beta}_1$ **(KNN,** $k = 5$**)**
BYM $PM_{2,5}$	0.1364 (0.0242, 0.2487)	0.0863 (-0.0314, 0.2040)
BYM NO_2	0.3840 (0.2093, 0.5588)	0.3013 (0.1370, 0.4656)

Es importante mencionar que para realizar esta comparativa, `INLA` necesita una matriz binaria, y para poder compararlo frente a un modelo Gaussiano la matriz de pesos para el vecindario debe ser simétrica y, como se menciona en la Sección 2.3.2, los vecindarios por k-vecinos más cercanos no siempre son simétricos. Lo cual es el caso para este vecindario. Es decir, al tener que simetrizar el vecindario, los municipios dejan de tener 5 vecinos. Es por ello que se decidió anteriormente designar un vecindario por contigüidad *Queen*.

Por otro lado, los coeficientes estimados bajo el vecindario KNN resultaron cercanos en signo y magnitud a los obtenidos con contigüidad *Queen*, con solapamiento sustancial de los intervalos de confianza con ($\alpha = 5\,\%$). Este resultado sugiere que la asociación positiva entre renta y contaminación identificada en la Sección 2.6.2 **no es un resultado de la definición particular de vecindario elegida**, sino un patrón robusto frente al menos dos especificaciones razonables de proximidad espacial, lo que refuerza la validez del modelo frente al problema de zonificación del MAUP (Sección 2.1.1).

4.3.2. Validación de los modelos

Para comparar el modelo lineal no espacial y el modelo BYM bajo un mismo criterio bayesiano, ambos se ajustaron especificando `control.compute = list(dic = TRUE, waic = TRUE)` en `R-INLA`: el modelo llineal como el modelo gaussiano $LMINLA$ de la ecuación 4.4, y el modelo espacial como el BYM BYM de la ecuación 4.3.

Cuadro 4.10: Comparación de criterios de información entre modelos lineales y modelos BYM (INLA).

	*pm*25_*median*		*no*2_*median*	
Parámetro	**LM**	**BYM**	**LM**	**BYM**
DIC	287.61*	-1952.71	524.25*	-1887.54
WAIC	288.31*	-2024.46	524.57*	-1953.29

*Para estos valores de DIC y WAIC se utilizó el modelado lineal mediante un modelo gaussiano de `R-INLA` sin componente espacial.

Los resultados del Cuadro 4.10 muestran una reducción sustancial y consistente de ambos criterios al incorporar la componente espacial. Para $PM_{2,5}$, el DIC disminuye de 287.61 a -1952.71 y el WAIC de 288.31 a -2024.46. Para el NO_2, el DIC pasa de 524.25 a y el WAIC de 524.57 a -1953.29. En los cuatro casos la diferencia excede ampliamente el umbral convencional de 10 puntos que Spiegelhalter et al. (2002) consideran evidencia sustancial a favor del modelo con menor valor (Sección 2.7), y el hecho de que DIC y WAIC coincidan en la dirección y en la magnitud aproximada de la mejora, pese a penalizar la complejidad de forma distinta, refuerza la robustez de esta conclusión.

Conviene señalar que los valores negativos de DIC y WAIC del modelo BYM no son un error de cálculo, sino una consecuencia esperada de trabajar con una verosimilitud gaussiana continua: la densidad gaussiana puede tomar valores arbitrariamente grandes cuando la varianza residual es muy pequeña, lo que arrastra la desviancia, y en consecuencia el DIC y el WAIC, hacia valores negativos. Esto es coherente con el análisis de componentes de la Sección 4.3: al ser $\hat{\phi} \approx 0.999$ en ambos contaminantes, casi toda la varianza residual representada en el modelo lineal queda absorbida por la componente espacial estructurada U en el modelo BYM.

Adicionalmentet, esta mejora tan pronunciada debe romarse con la cautela interpretativa señalada mencionada anteriormente: una parte de la ganancia en ajuste refleja genuinamente la fuerte dependencia espacial de los contaminantes atmosféricos, compartida por municipios vecinos debido a la meteorología y el tráfico regional, pero otra parte puede estar magnificada por la resolución relativamente gruesa del producto CAMS-MONARCH ($\sim$203 km^2) frente al tamaño de muchos municipios,

que introduce de por sí un suavizado espacial en los datos de entrada antes incluso de ajustar el modelo. El test de Moran sobre la componente no estructurada V (Cuadro 4.8, no significativo en ambos casos) respalda que el reparto entre U y V es coherente y no un simple artefacto de sobreajuste, pero no permite descartar que parte de la señal capturada por U provenga de la propia interpolación de los datos de contaminación más que de un proceso poblacional subyacente.

En conjunto, tanto el DIC como el WAIC confirman de forma consistente que el modelo BYM ofrece un ajuste sustancialmente superior al modelo lineal sin componente espacial, por lo que se selecciona como especificación final para la interpretación de los resultados de este trabajo.

Capítulo 5

Conclusiones

Este trabajo ha aplicado técnicas de estadística espacial para datos agregados por área con el objetivo de caracterizar la distribución espacial de la contaminación atmosférica en los 313 municipios gallegos y evaluar su relación con la desigualdad socioeconómica municipal, en el marco de los Objetivos de Desarrollo Sostenible de la Agenda 2030.

5.1. Principales hallazgos

Existencia de autocorrelación espacial global

El índice I de Moran resultó positivo y estadísticamente significativo para las tres variables analizadas ($PM_{2,5}$, NO_2 y renta(log)), lo que confirma que la distribución espacial de la contaminación atmosférica en Galicia no es aleatoria. Los municipios con altas concentraciones de contaminantes tienden a estar rodeados de municipios igualmente contaminados, formando clústeres regionales que reflejan la influencia de factores geográficos, meteorológicos e industriales compartidos.

Identificación de clústeres locales con LISA

El análisis local reveló clústeres Alto-Alto de contaminación en áreas urbanas (ciudades con mayor densidad poblacional) y clústeres Bajo-Bajo en el interior rural de Lugo y Ourense. Aunque ninguno de los contaminantes sobrepasan las directrices de la OMS para concentraciones anuales, la identificación de estos patrones proporciona información geográficamente específica que puede orientar las políticas de vigilancia y reducción de la contaminación atmosférica en Galicia.

Asociación espacial entre contaminación y renta

Por su parte, el análisis BiLISA y los modelos BYM muestran una asociación positiva entre renta y concentración de contaminantes: los municipios que reportan mayor renta tienden a presentar mayor exposición a NO_2 y $PM_{2,5}$, y municipios con baja renta presentan baja exposición a contaminantes, lo que contrasta con la hipótesis de injusticia ambiental. Este resultado es coherente con la naturaleza de los contaminantes en el contexto gallego: el NO_2 y el $PM_{2,5}$ a escala municipal están asociados principalmente a la actividad económica, el tráfico y la densidad poblacional, que son más elevados en los municipios metropolitanos. No obstante, el análisis BiLISA identificó un conjunto de municipios en el cuadrante Alto-Bajo, municipios con alta contaminación y vecinos con baja renta, que sí presentan indicios de injusticia ambiental local y merecen atención en las políticas de equidad ambiental.

Necesidad de las componentes espaciales

Los modelos BYM demostraron que los modelos lineales sobreestiman sustancialmente el efecto de la renta sobre la contaminación: el coeficiente se redujo en más de un 85 % al incorporar la componente espacial. Este resultado subraya la importancia de utilizar modelos que incorporen explícitamente la dependencia espacial en el análisis de datos municipales, ya que los modelos no espaciales pueden producir conclusiones sesgadas.

Selección del modelo mediante criterios de información

La comparación formal mediante DIC y WAIC respaldó de forma inequívoca la elección del modelo BYM frente al modelo de referencia: para $PM_{2,5}$ y para NO_2. En los cuatro casos la mejora superó ampliamente el umbral de 10 puntos que se considera evidencia sustancial (Spiegelhalter et al., 2002), y la concordancia entre ambos criterios refuerza la fiabilidad de esta selección.

Robustez frente a la definición del vecindario

El análisis de sensibilidad realizado con un vecindario alternativo de $k = 5$ vecinos más cercanos produjo coeficientes de renta muy próximos, en signo y magnitud, a los obtenidos con contigüidad *Queen* con solapamiento sustancial de los intervalos de credibilidad al 95 %. Esto indica que la asociación positiva entre renta y contaminación identificada en este trabajo no depende de la especificación particular de vecindario elegida, sino que constituye un patrón razonablemente robusto frente al problema de zonificación del MAUP.

5.2. Limitaciones

El análisis presenta varias limitaciones que deben tenerse en cuenta al interpretar los resultados. En primer lugar, la resolución espacial del producto CAMS-MONARCH (aproximadamente 203 km^2 por celda) es comparable o superior a la extensión de muchos municipios gallegos del interior, lo que reduce la precisión de las estimaciones municipales de concentración.

En segundo lugar, el modelo solo incluye una covariable (*renta_log*), por lo que la componente espacial estructurada U absorbe la variabilidad asociada a factores no medidos como la densidad de tráfico, la proximidad a fuentes industriales o las condiciones meteorológicas locales.

En tercer lugar, por la disposición y alto volumen de datos el análisis se limita al año 2023, lo que limita la interpretación a solamente un año en el que se pueden presentar comportamientos atípicos correspondientes a fenómenos propios de ese año.

Por otro lado, el MAUP implica que los patrones observados a nivel municipal pueden diferir de los que se observarían a escalas más finas, si bien el análisis de sensibilidad realizado mitiga parcialmente la preocupación asociada al efecto de zonificación, no aborda el efecto de escala, que exigiría datos socioeconómicos y de contaminación disponibles a una resolución más fina de la que permite actualmente la fuente utilizada.

Finalmente, conviene subrayar que la asociación espacial identificada entre renta y contaminación, tanto en el análisis BiLISA como en los modelos BYM, es de naturaleza descriptiva y correlacional, no causal. La componente espacial estructurada U_i recoge variación compartida entre municipios vecinos que puede deberse a múltiples factores no medidos con estructura espacial similar (tráfico, industria, meteorología), por lo que la explicación ofrecida en este trabajo, mayor renta asociada a mayor actividad económica y tráfico, es la interpretación más plausible a la luz de la literatura, pero no queda demostrada de forma causal por el diseño observacional y transversal de este estudio.

5.3. Contribución a los Objetivos de Desarrollo Sostenible

Más allá de los hallazgos sustantivos, este trabajo aporta una base metodológica reproducible para el seguimiento de los ODS 10 (reducción de las desigualdades) y ODS 11 (ciudades y comunidades sostenibles) en el contexto gallego. El flujo de trabajo desarrollado, desde la construcción del vecindario y la matriz de pesos hasta el ajuste de modelos BYM en `R-INLA`, es directamente replicable con datos de años posteriores o con covariables adicionales, y permite identificar de forma sistemática los municipios donde coinciden vulnerabilidad económica y carga ambiental elevada, información que no resulta evidente a partir de estadísticos agregados a escala autonómica. Si bien los niveles de contaminación observados en Galicia en 2023 se mantienen por debajo de los límites máximos establecidos por la OMS, y por tanto no se derivan de este trabajo implicaciones inmediatas para políticas correctivas de equidad ambiental, la metodología aquí presentada ofrece una herramienta de vigilancia temprana: permite detectar, antes de que se conviertan en un problema de salud pública, los municipios en los que una futura intensificación de fuentes contaminantes podría generar una carga ambiental desproporcionada sobre población ya vulnerable.

5.4. Investigaciones futuras

A partir de las limitaciones identificadas, se proponen las siguientes líneas de trabajo:

- Incorporar covariables adicionales al modelo BYM: densidad de tráfico, distancia a polígonos industriales, índice de urbanización y variables meteorológicas, para separar mejor el efecto de la renta de otros determinantes de la contaminación.
- Extender el análisis a series temporales (2018-2023) para evaluar si los patrones de desigualdad ambiental han aumentado o disminuido en un período específico, e.g. el periodo posterior a la pandemia.
- Comparar los resultados con otras comunidades autónomas españolas para evaluar si los patrones observados en Galicia son específicos de su estructura territorial o comunes en contextos urbano-rurales similares.
- Explorar técnicas de *downscaling* o desagregación dasimétrica para armonizar la resolución del producto CAMS-MONARCH con la unidad municipal, mitigando así el riesgo de que la dependencia espacial estimada por el modelo BYM esté parcialmente inflada por la interpolación de los datos de entrada, en línea con la metodología propuesta por Hewitt et al. (2024) para la armonización de rejillas con sectores censales.

En síntesis, aunque este trabajo no encuentra evidencia de una distribución desigual de la carga ambiental en Galicia bajo los límites establecidos por la OMS, sí identifica un conjunto concreto y localizable de municipios donde vulnerabilidad económica y exposición a contaminantes coinciden, y que merecen seguimiento prioritario. Más allá del caso gallego, el principal valor de este trabajo reside en la metodología: un flujo de trabajo reproducible de estadística espacial para datos agregados por área que, aplicado de forma sistemática y periódica, puede transformar el compromiso abstracto con los ODS en una herramienta concreta de vigilancia y toma de decisiones en materia de justicia ambiental.

Apéndice

1. Código

Para referencia del código utilizado en este trabajo véase el repositorio de GitHub Areal-Data. En este repositorio se encuentra:

- **Carpeta de datos** `data`: con los datos brutos de la renta per cápita extraídos de la página del Instituto Galego de Estatística en formato *CSV* y los datos procesados para el análisis en formato *CSV* y *GPKG*.
- **Código de procesado de datos** `Limpieza_bbdd.R`: el código de `R` implementado para la limpieza y el preprocesado de los datos brutos en una base de datos analizable.
- C**ódigo de análisis** `ESDA.R`: el código en `R` correspondiente a todo el ESDA, el modelado espacial y no espacial y la evaluación de los mismos.

1.1. Resultados

Estadísticos descriptivos

En el Cuadro 1 se incluyen los valores específicos de las medidas de localización de las variables.

Cuadro 1: Estadísticos descriptivos de las variables de análisis para datos de agregados por área en 313 municipios de Galicia, 2023.

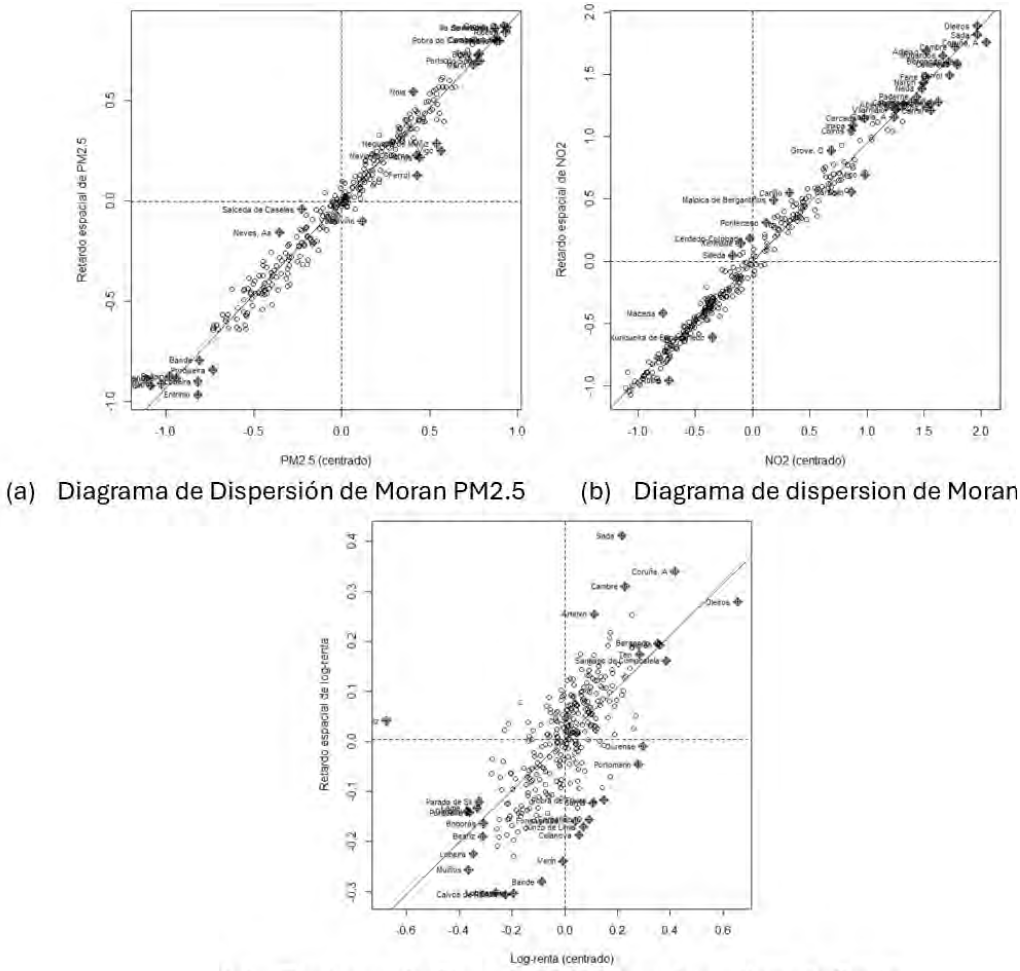
Variable	Mínimo	Primer Cuartil	Mediana	Media	Tercer Cuartil	Máximo
$pm25\_median$ (µg/m^3)	2.150	2.956	3.280	3.263	3.568	4.203
$no2\_median$(µg/m^3)	0.4794	1.0859	1.4292	1.5943	2.0187	3.6376
renta_pc (miles €/pc)	8.152	14.687	16.238	16.209	17.503	30.879

Diagramas de dispersión de Moran

Debido a su tamaño, los diagramas de dispersión de Moran individuales se incluyen en este apartado en la Figura 1 para su visualización clara.

Resultados BiLISA

En el Cuadro 2 se enlistan los municipios correspondientes a significativamente altos niveles de contaminantes con retardo de *renta_log* bajos, es decir en el cuadrante Alto-Bajo, con su provincia correspondiente, separados por contaminante.



(a) Diagrama de Dispersión de Moran PM2.5 (b) Diagrama de dispersion de Moran NO2

(c) Diagrama de Dispersión de Moran Logaritmo de Renta

Figura 1: Diagramas de dispersión de Moran para $PM_{2,5}$, NO_2, y logaritmo de renta.

Municipio	Provincia	$PM_{2,5}$	NO_2
Fonsagrada, A	Lugo	✓	—
Avión	Ourense	✓	—
Beariz	Ourense	✓	—
Boborás	Ourense	✓	—
Carballiño, O	Ourense	✓	—
Covelo	Pontevedra	✓	—
Forcarei	Pontevedra	✓	—
Fornelos de Montes	Pontevedra	✓	✓
Lama, A	Pontevedra	✓	—

Cuadro 2: Municipios en el cuadrante BiLISA Alto-Bajo

Bibliografía

Anselin, Luc (2019). «A local indicator of multivariate spatial association: extending Geary's C». En: *Geographical Analysis* 51.2, págs. 133-150.

Anselin, Luc (1995). «Local indicators of spatial association—LISA». En: *Geographical Analysis* 27.2, págs. 93-115.

Anselin, Luc (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers.

Anselin, Luc, Ibnu Syabri y Youngihn Kho (2006). «GeoDa: An introduction to spatial data analysis». En: *Geographical Analysis* 38.1, págs. 5-22. DOI: 10.1111/j.0016-7363.2005.00671.x.

Baston, Daniel (2026). ***exactextractr**: Fast Extraction from Raster Datasets using Polygons*. R package. URL: https://isciences.gitlab.io/exactextractr/.

Besag, Julian, Jeremy York y Annie Mollié (1991). «Bayesian image restoration, with two applications in spatial statistics». En: *Annals of the Institute of Statistical Mathematics* 43.1, págs. 1-20. DOI: 10.1007/BF00116466.

Bivand, Roger (2022). «R packages for analyzing spatial data: A comparative case study with areal data». En: *Geographical Analysis* 54.3, págs. 488-518.

Bivand, Roger S., Edzer Pebesma y Virgilio Gómez-Rubio (2013). *Applied Spatial Data Analysis with R*. 2nd. Springer.

Copernicus Atmosphere Monitoring Service (2023). *CAMS European Air Quality Reanalyses (EAC4)*. European Centre for Medium-Range Weather Forecasts (ECMWF). Consultado en 2026. URL: https://ads.atmosphere.copernicus.eu/datasets/cams-europe-air-quality-reanalyses?tab=download.

Cressie, Noel (1993). *Statistics for spatial data*. John Wiley & Sons.

Desarrollo-Sostenible, ONU (2023). *Los 17 Objetivos*. URL: https://sdgs.un.org/es/goals (visitado 21-04-2026).

Hernangómez, Diego (2024). ***mapSpain**: Administrative Boundaries of Spain*. R package. URL: https://ropenspain.github.io/mapSpain/.

Hewitt, Richard J, Eduardo Caramés y Rafael Borge (2024). «Is air pollution exposure linked to household income? Spatial analysis of Community Multiscale Air Quality Model results for Madrid». En: *Heliyon* 10.5.

INE (2023). *Atlas de Distribución de Renta de los Hogares 2023*. Instituto Nacional de Estadística. Consultado en 2026. URL: https://www.ine.es/jaxiT3/Tabla.htm?t=53689&L=0.

INE (2025a). *Censo anual de población 2021-2025*. Instituto Nacional de Estadística. Consultado en 2026. URL: https://www.ine.es/jaxiT3/Datos.htm?t=67988#_tabs-tabla.

INE (2025b). *Índice de envejecimiento por municipio*. Instituto Nacional de Estadística. Consultado en 2026. URL: https://www.ine.es/jaxiT3/Tabla.htm?t=30702&L=0.

Instituto Galego de Estatística (2023). *Renta bruta per cápita por municipio, 2023*. Instituto Galego de Estatística. Consultado en 2026. URL: https://www.ige.gal/igebdt/esq.jsp?c=0307007004&ruta=verTabla.jsp?COD=9642&R=9915%5Bn0:n3%5D&C=0%5B2023%5D&F=1:0&OP=1.

James, Gareth, Daniela Witten, Trevor Hastie, Robert Tibshirani et al. (2013). *An introduction to statistical learning: with applications in R*. Vol. 103. Springer.

LeSage, James P. y Robert Kelley Pace (2009). *Introduction to Spatial Econometrics*. CRC Press.

Ministerio para la Transición Ecológica y el Reto Demográfico (s.f.). *Partículas*. Glosario de contaminantes, Calidad y Evaluación Ambiental. Consultado en 2026. URL: https://www.miteco.gob.es/en/calidad-y-evaluacion-ambiental/temas/atmosfera-y-calidad-del-aire/glosario-de-terminos/glosario-contaminantes/particulas.html.

Moraga, Paula (2023). *Spatial statistics for data science: theory and practice with R*. Chapman y Hall/CRC.

Moraga, Paula, Susanna M Cramb, Kerrie L Mengersen y Marcello Pagano (2017). «A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE». En: *Spatial Statistics* 21, págs. 27-41.

Moreno-Jimenez, Antonio, Rosa Cañada-Torrecilla, María Jesús Vidal-Domínguez, Antonio Palacios-Garcia y Pedro Martinez-Suarez (2016). «Assessing environmental justice through potential exposure to air pollution: a socio-spatial analysis in Madrid and Barcelona, Spain». En: *Geoforum* 69, págs. 117-131.

ONU (1987). *Report of the World Commission on Environment and Development : note / by the Secretary-General*. URL: https://digitallibrary.un.org/record/139811?v=pdf (visitado 21-04-2026).

Organización Mundial de la Salud (2021). *Directrices mundiales de la OMS sobre la calidad del aire: partículas en suspensión ($PM_{2,5}$ y PM_{10}), ozono, dióxido de nitrógeno, dióxido de azufre y monóxido de carbono. Resumen ejecutivo*. Licencia: CC BY-NC-SA 3.0 IGO. Ginebra: Organización Mundial de la Salud. ISBN: 978-92-4-003546-1. URL: https://apps.who.int/iris/handle/10665/345329.

Pebesma, Edzer y Roger Bivand (2023). *Spatial data science: with applications in R*. Chapman y Hall/CRC.

Rey, Sergio J. y Luc Anselin (2010). «PySAL: A Python library of spatial analytical methods». En: *The Review of Regional Studies* 37.1, págs. 5-27.

Rue, Håvard, Sara Martino y Nicolas Chopin (2009). «Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations». En: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71.2, págs. 319-392.

Soares, Joana, Dietrich Plass, Sarah Kienzler, Alberto González Ortiz, Artur Gsella y Jan Horálek (dic. de 2024). *Assessing the environmental burden of disease related to air pollution in Europe in 2022*. Inf. téc. European Topic Centre on Human Health y the Environment. URL: https://www.eionet.europa.eu/etcs/etc-he/products/etc-he-products/etc-he-reports/etc-he-report-2024-6-assessing-the-environmental-burden-of-disease-related-to-air-pollution-in-europe-in-2022.

Spiegelhalter, David J, Nicola G Best, Bradley P Carlin y Angelika Van Der Linde (2002). «Bayesian measures of model complexity and fit». En: *Journal of the royal statistical society: Series b (statistical methodology)* 64.4, págs. 583-639.

Tobler, Waldo R. (1970). «A computer movie simulating urban growth in the Detroit region». En: *Economic Geography* 46, págs. 234-240.

Walker, Gordon (2012). *Environmental justice: Concepts, evidence and politics*. Routledge.

Watanabe, Sumio (2010). «Equations of states in singular statistical estimation». En: *Neural Networks* 23.1, págs. 20-34.