USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Análise de datos reticulares

Jose Fuentes Rodríguez

2018/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Análise de datos reticulares

Jose Fuentes Rodríguez

Xulio 2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Análise de datos reticulares
Breve descrición do contido
Neste traballo revisaranse os principais métodos asociados a datos reticulares espaciais (en inglés, Lattice Data). Os datos reticulares prodúcense cando temos datos espaciais que resumen os valores dunha zona e se quere analizar a sua dependencia coas zonas cercanas (veciños). A configuración global da veciñanza das zonas chamase retícula e permítenos estimacións adaptadas a esa dependencia.
Recomendacións
Soltura co linguaxe de programación R
Outras observacións

Índice general

Resumen

La estadística espacial es la rama de la estadística que trata datos teniendo en cuenta su situación y distribución en el espacio. La estadística de datos reticulares divide el espacio en regiones o áreas discretas y estudia el efecto de variables aleatorias sobre sí mismas (autorregresión) como consecuencia de su realización en áreas próximas.

En este documento se introduce una construcción teórica de los modelos espaciales más utilizados para los datos reticulares centrándose en el caso gaussiano, como los modelos simultáneamente y condicionalmente especificados. Se continúa con una relación de las estructuras espaciales discretas utilizadas comúnmente para describir datos reticulares. En el tercer capítulo se discuten las distintas pruebas para evaluar la existencia de dependencia espacial. Finalmente, en el último capítulo se incluye un ejemplo práctico en el que se construyen modelos espaciales sobre la mortalidad por cáncer de pulmón en municipios de Galicia incluyendo los efectos de varias covariables.

Abstract

Spatial statistics is the branch of statistics that processes data considering their location and distribution in space. Lattice data statistics divides space into regions or discrete areas and studies the effect of random variables on themselves (autoregression) as a consequence of their realization in nearby areas .

In this document, a theoretical construction is introduced for the most usual spatial models for lattice data, such as simultaneously and conditionally specified models, focusing on the Gaussian case. It continues with an enumeration of the discrete spatial structures commonly used to describe lattice data. The third chapter discusses the different tests used to evaluate the existence of spatial dependence. Finally, the last chapter includes a practical example in which spatial models on lung cancer mortality in municipalities of Galicia are fitted taking into account the effects of several covariates.

Introducción

La estadística espacial puede definirse como el conjunto de técnicas estadísticas usadas para estudiar datos en base a sus características geográficas, geométricas o topológicas. Tiene sus orígenes en la elaboración de mapas de datos, estudios botánicos y agrarios y censos, sin embargo, su verdadero desarrollo no llega hasta el siglo XX, cuando se empiezan a elaborar los primeros modelos estadísticos que tienen en cuenta la localización de los datos en el espacio, y que permiten resumir, analizar y predecir resultados. Una de las principales diferencias de la estadística espacial con respecto a otras ramas es la falta de necesidad de datos independientes e idénticamente distribuidos, pues por la propia naturaleza espacial de los datos que son objeto de estudio, cada elemento de una muestra dependerá del valor de los elementos cercanos. Esta correlación entre distintos elementos de una muestra se denomina autocorrelación y es la misma idea detrás del estudio de datos temporales, como las series de tiempo.

Los datos y variables con los que trabaja la estadística espacial pueden tener formas muy distintas. Pueden ser variables discretas o continuas, su distribución en el espacio puede ser también continua o referida a un conjunto discreto de localizaciones y, asimismo, estas localizaciones pueden tener diferente grado de regularidad.

En cualquier caso, se puede presentar [1] un modelo general para cualquier proceso con características espaciales. Sea $s \in \mathbb{R}^d$ un punto en el espacio euclídeo d-dimensional y se define $Z(s)$ como el posible dato en ese punto de una cierta cantidad aleatoria. Si s varía en un conjunto de índices D, se obtiene un campo aleatorio multivariante. $\{Z(s) : s \in D\}$. Una realización de esta "población en cada punto" se denota como $\{z(s) : s \in D\}$. D se considera como un *conjunto aleatorio*, es decir, una aplicación medible del espacio de probabilidades a un espacio de medida de subconjuntos de $\mathbb{R}^d$. Esto significa que D, es decir, el conjunto de localizaciones, puede o no variar entre distintas muestras del mismo proceso aleatorio.

Esta definición permite abordar multitud de problemas incluidos en la estadística espacial. Si D es fijado pero conteniendo subconjuntos de medida positiva, se tienen problemas de *datos geoestadísticos*; si D es una colección contable de puntos fijados, se trata de *datos reticulares*; si, en cambio, D no es fijado, sino un proceso puntual en $\mathbb{R}^d$, son los llamados

problemas de *patrones de puntos*. En este trabajo se tratarán únicamente los resultados, modelos y casos relativos a los datos reticulares.

Originalmente, el análisis de datos basados en retículas aparece ligado al tratamiento de problemas agrarios[2] [16].En estos, un área de cultivo se divide en parcelas rectangulares, de forma que para cada propiedad o característica se recoge un dato por cuadrícula. Este planteamiento no se restringe al ámbito agrícola, por ejemplo, es similar al ideado para el tratamiento de imágenes y la teledetección vía satélite[1, Section 7.4], donde una información que se obtiene por diversos sensores es digitalizada, pasando a ser representada por una cantidad discreta de píxeles y valores. Estos píxeles cuadrados pasan a ser las parcelas de los ejemplos anteriores.

Además de estos ejemplos, en los que la estructura cuadriculada permitía construir una matriz regular, los datos reticulares tienen una gran presencia en estudios econométricos y médicos. Resulta muy sencillo tener datos recogidos en relación a países, provincias o regiones y estas, a su vez, tienen una clara estructura reticular, ya que las divisiones administrativas dividen áreas geográficas en conjuntos discretos. Ejemplos de esta clase de datos se pueden encontrar en el estudio del Síndrome de Muerte Súbita del Lactante (SIDS), realizado por Cressie y Chan [12], o los datos de incidencia de leucemia en condados del estado de Nueva York discutidos en [10]. En estos casos, al no necesitar construir una retícula para cada experimento, es mucho más fácil obtener estos datos. Por esta razón, este trabajo se centra más en datos con retícula irregular; sin embargo, el planteamiento siguiente será generalizado y aplicable a los problemas con mayor regularidad.

Este texto presenta inicialmente el marco teórico para poder estudiar, modelizar y predecir datos reticulares y, a continuación, se incluyen distintas formas en las que se pueden aplicar estos modelos así como técnicas para valorar el grado de dependencia espacial entre distintas áreas. Finalmente, se incluyen ejemplos sobre datos reales de la aplicación de las técnicas y modelos discutidos en este texto.

Capítulo 1

Datos reticulares y modelos espaciales

En este capítulo se resumen ideas y resultados necesarios para desarrollar la estadística espacial de datos reticulares. Se definen cómo son estas estructuras, así como los modelos posibles y los teoremas y propiedades que establecen las hipótesis necesarias para poder realizarse.

1.1. Retículas

Las retículas son colecciones de regiones contiguas en el espacio separadas unas de otras por sus fronteras de forma que, para cualquier proceso o suceso, no importa el punto donde ocurra, solamente la región. Las regiones, por lo tanto, pueden ser identificadas por un punto del espacio que se denomina centro y puede escogerse de distintas formas [6].Como las regiones son esencialmente polígonos, se suele tomar el centroide de estas. Sin embargo, en casos donde se conozcan puntos más representativos de cada región, también se pueden escoger como centro; por ejemplo: capitales de provincia, centros de alta densidad de población o desarrollo económico, etc. Una forma más formal de definir una estructura reticular a partir de los centros [1, pp. 384, 385] puede ser:

Definición 1.1. Una retícula es una colección contable de puntos en el espacio, donde ocurren sucesos de forma que no pueden tener lugar en ningún punto no incluido en la retícula. Es decir $D = \{(i; s_i) : i \in I, s_i \in R^d\}$

Comúnmente se trabaja sobre mapas geográficos, por lo que $d = 2$ y $s_i = (s_i^{(1)}, s_i^{(2)})$; de aquí en adelante se trabaja con retículas 2-dimensionales, a no ser que se especifique lo contrario.

Los sucesos que ocurren en las retículas pueden o no ser afectados por las regiones "próximas". Esta relación de proximidad se denomina relación de vecindad, y cómo se define (o se estima) en cada caso forma parte de las discusiones de este capítulo y del siguiente. Esta definición permite entender las retículas como un grafo de la siguiente forma.

$$D = \{(s_i) : i \in I, s_i \in \mathbb{R}^d\}$$
$$N_i = \{k \in I : k \text{ es vecino de } i\}$$
$$D_N = \{(i, N_i) : i \in I\}$$

Cuando se considera un proceso espacial $Z(s) : s \in D$ en una retícula se asume que este solo puede realizarse en las regiones determinadas; es decir, se trata como si las únicas localizaciones posibles fueran los centros de las áreas. Si se considera además la retícula como una parte de una retícula infinita, se puede pensar en los modelos de datos reticulares como en el análogo espacial de las series de tiempo. En lo siguiente, se sigue de forma sintética el planteamiento de Cressie [1], de forma que se sigue también la siguiente notación:

Notación 1.2. La probabilidad condicionada de un proceso $Pr(Z(t_1) = z(t_1), \dots, Z(t_R) = z(t_R) \mid Z(t_{R+1}) = z(t_{R+1}), \dots, Z(t_{R+S}) = z(t_{R+S}))$ se notará por

$$Pr(z(t_1), \dots, z(t_R) \mid z(t_{R+1}), \dots, z(t_{R+S})) \tag{1.1}$$

Además, se tratará Z como variable discreta. En caso de que fuera continua bastaría con substituir $\Pr(Z)$ por la función de distribución de la variable $f(Z)$.

1.2. Autocorrelación en series de tiempo y su análogo espacial

Definición 1.3. Una serie de tiempos es un proceso aleatorio indexado en un índice continuo (generalmente el tiempo). Este índice se supone observado en incrementos discretos, es decir $\{Z(t) : t = 0, 1, \dots\}$

En los datos de series de tiempo es normal que los datos presenten dependencia entre ellos, es decir, el proceso depende de los estados por los que haya pasado anteriormente. La forma más sencilla en la que aparece dependencia son las llamadas cadenas de Markov, donde los datos solo dependen del valor inmediatamente anterior temporalmente. Se pueden dar dos definiciones de una cadena de Markov: una teniendo en cuenta la probabilidad o distribución conjunta de Z, y otra trabajando con las probabilidades condicionadas. Estos dos enfoques marcarán las pautas para las distintas formas de modelos de regresión espacial en datos reticulares.

Definición 1.4 (Enfoque de probabilidad conjunta)**.** Un proceso temporal $\{Z(t) : t = 0, 1, \dots\}$ es una cadena de Markov si:

$$\Pr(z(1), \dots, z(i) \mid z(0)) = \prod_{t=1}^{i} Q_t(z(t); z(t-1)) \quad \forall\, i \geq 1 \tag{1.2}$$

Donde Q_t es una función de $z(t)$ y de $z(t-1)$. En el caso de independencia, Q solo depende de $z(t)$.

Definición 1.5 (Enfoque de probabilidad condicionada)**.** Un proceso temporal $\{Z(t) : t = 0, 1, \dots\}$ es una cadena de Markov si:

$$\Pr(z(i) \mid z(0), \dots, z(i-1)) = \Pr(z(i) \mid z(i-1) \quad \forall\, i \geq 1 \tag{1.3}$$

La equivalencia entre ambas definiciones se puede encontrar en [1, p. 403]

En el caso espacial, esta equivalencia, generalmente, no se cumple. Por ejemplo, se supone como retícula la cuadrícula en el plano $D = s = (u, v)' : u, v \in \mathbb{Z}$ y que a cada región solo le afectan las casillas directamente norte, sur, este y oeste de ella (Véase relación de torre en el capítulo siguiente). En este caso el análogo a 1.4 es

$$\Pr(z) = \prod_{(u,v)\in D} Q_{uv}(z(u, v); z(u-1, v), z(u+1, v), z(u, v-1), z(u, v+1)) \tag{1.4}$$

Mientras que el caso condicional 1.5 se puede escribir como

$$\begin{aligned} &\Pr(z(u, v)) \mid z(k, l) : (k, l) \neq (u, v)) = \\ &\qquad \Pr(z(u, v)) \mid z(u-1, v), z(u+1, v), z(u, v-1), z(u, v+1)) \quad \forall (u, v) \in D \end{aligned} \tag{1.5}$$

Cada uno de estos dos enfoques da lugar a la construcción de dos tipos distintos de modelos, los modelos especificados simultáneamente y los especificados condicionalmente. Se introduce ahora una idea de la construcción de estos modelos en el caso gaussiano, que se completará más adelante, tras dar la base teórica que hace posible su construcción.

1.3. El modelo espacial simultáneamente especificado

Los modelos simultáneamente especificados se construyen sobre variables gaussianas independientes en la retícula, que son modificados por el efecto de las regiones cercanas.

Sea $\epsilon(s) \sim \mathcal{N}(0, \Lambda)$ donde $\Lambda = \sigma^2 I$. Se define una matriz $B = (b_{ij})$ como la matríz que aplica dependencia espacial entre los valores de Z, es decir, b_{ij} es el efecto de Z_j sobre Z_i. Si $(I - B)^{-1}$ existe, entonces $Z = (Z(s_1), \dots, Z(s_n)$ se puede definir como

$$(I - B)(Z - \mu) = \epsilon \tag{1.6}$$

Equivalentemente $(Z-\mu) = \epsilon(I-B)^{-1}$ por lo que se tiene

$$\mathbb{E}(Z-\mu) = \mathbb{E}(\epsilon(I-B)^{-1}) = 0 \text{ y por tanto } \mathbb{E}(Z) = \mu$$
$$var(Z) = var(Z-\mu) = var(\epsilon(I-B)^{-1}) = (I-B)^{-1}\Lambda(I-B')^{-1}$$

Es decir $Z(s) \sim \mathcal{N}(\mu, (I-B)^{-1}\Lambda(I-B')^{-1})$

Con esto es posible establecer un modelo de autorregresión para Z, reescribiendo 1.6 como:

$$Z(s_i) = \mu_i + \sum_{j=1}^{n} b_{ij}(Z(s_j) - \mu_j) + \epsilon_i \tag{1.7}$$

Estimando por máxima verosimilitud μ, σ, y b_{ij} para $i \neq j$, pues un dato no se puede influenciar así mismo.

1.4. El modelo espacial condicionalmente especificado

Construir modelos basados en probabilidad condicionada, gaussiana o no, es más complicado y requiere una base teórica para definir en qué casos, y de qué forma, es posible hacerlo. Al igual que en el caso simultáneo, el objetivo es el mismo: obtener una distribución para Z en función de unos parámetros que son estimados mediante verosimilitud u otros métodos de inferencia. La diferencia radica en que en vez de considerar esta distribución a partir de otra conjunta, como es el caso de los ϵ en los modelos simultáneos, se parte de la función de distribución condicionada; si Z es normal toma la forma siguiente:

$$f(z(s_i) \mid \{z(s_j) : j \neq i)\}) = \frac{1}{\tau_i\sqrt{2\pi}} \exp(-(z(s_i) - \theta_i(\{z(s_j) : j \neq i\}))^2/2\tau_i^2) \qquad i = 1, \ldots, n \tag{1.8}$$

El objetivo será obtener μ, Σ de forma que

$$Z \sim \mathcal{N}(\mu, \Sigma)$$

1.4.1. Campos aleatorios de Markov

En esta parte se sentarán las condiciones necesarias sobre la retícula y la distribución para poder realizar modelos basados en la probabilidad condicionada. A partir de esto, se obtendrán resultados que servirán de guía para obtener la distribución conjunta. De aquí en adelante, se utiliza la notación para variables discretas por comodidad, puesto que todo lo expuesto es válido, no sólo para el caso gaussiano. Para variables continuas basta con cambiar la función de masas Pr por la de distribución f y las sumas por integrales.

Definición 1.6. En una cierta retícula con regiones $\{s_i : i = 1, \ldots, n\}$ se observa una variable Z. Si se define $\zeta = \{z : \Pr(z) = \Pr(Z = z) > 0\}$ y $\zeta_i = \{z : \Pr(z_i) = \Pr(Z_i = z_i) > 0\}$ entonces se dice que Z satisface la *condición de positividad* si $\zeta = \zeta_1 \times \cdots \times \zeta_n$

Definición 1.7. Una región k de una retícula se dice *vecina* de otra i si la distribución condicionada de $Z(s_i)$ condicionada a todos los valores $z(s_j)$ $j = 1, \ldots, n$ depende funcionalmente del valor $z(s_k)$ para $k \neq i$. Además, se define el conjunto de vecinos de i como

$$N_i = \{k : k \text{ es vecino de } i\}$$

Definición 1.8. Una *camarilla (clique)* se define como un conjunto de regiones (o una sola región) de la retícula, de forma que todos son vecinos unos de otros.

Definición 1.9. Toda medida de probabilidad cuyas distribuciones condicionadas definan una estructura de vecindades como la anterior, se denomina un *campo aleatorio de Markov, Markov random field (MRF)* en inglés.

Definición 1.10. Se puede suponer sin pérdida de generalidad que $\vartheta = (0, ..., 0) \in \zeta$ (si no puede tomarse otro valor). Se define la *función negpotencial* como

$$Q(z) = \log\left(\frac{\Pr(z)}{\Pr(\vartheta)}\right) \tag{1.9}$$

Proposición 1.11 (Propiedades de la función negpotencial)**.** *La función Q cumple las siguientes propiedades:*

1. $$\frac{\Pr(z(s_i) \mid \{z(s_j) j \neq i\}}{\Pr(0(s_i) \mid \{z(s_j) j \neq i\})} = \frac{\Pr(z)}{\Pr(z_{0;i})} = \exp(Q(z) - Q(z_i)) \tag{1.10}$$

   Donde $0(s_i)$ denota a $Z(s_i) = 0$ y $z_{0;i} = (z(s_1), \ldots, z(s_{i-1}), 0, z(s_{i+1}), \ldots, z(s_n))$

2. *Q puede expandirse unívocamente como*

$$\begin{aligned} Q(z) = &\sum_{1 \leq i \leq n} z(s_i) G_i(z(s_i)) + \sum_{1 \leq i < j \leq n} \sum z(s_i) z(s_j) G_{ij}(z(s_i), z(s_j)) \\ &+ \sum_{1 \leq i < j < k \leq n} \sum \sum \sum z(s_i) z(s_j) z(s_k) G_{ijk}(z(s_i), z(s_j), z(s_k)) + \ldots \\ &+ z(s_1) \ldots z(s_n) G_{1,\ldots,n}(z(s_1) \ldots z(s_n)) \end{aligned} \tag{1.11}$$

La demostración de esta proposición puede encontrarse en [1, p. 416]. Debido a la importancia de las funciones G en la expresión anterior, es de utilidad encontrar propiedades sobre ellas. La más importante es el *Teorema de Hammersley-Clifford*[1, p. 417]

Teorema 1.12 (Hammersley-Clifford)**.** *Sea Z un vector aleatorio en una retícula que constituye una MRF y tal que ζ cumple la condición de positividad. Si las áreas $i, j, \ldots, s$ no forman una camarilla, entonces $G_{ij\ldots s}$ es idénticamente nula.*

Demostración. Consideremos sin pérdida de generalidad la primera región. Sea $z_{0;1} = (0, z(s_2), \ldots, z(s_n))$. Aplicando la segunda propiedad de la función Q 1.11 se tiene la descomposición

$$Q(z_{0;1}) = \sum_{2\leq i\leq n} z(s_i)G_i(z(s_i)) + \sum_{2\leq i<j\leq n}\sum z(s_i)z(s_j)G_{ij}(z(s_i), z(s_j)) + \sum_{1\leq i<j<k\leq n}\sum\sum z(s_i)z(s_j)z(s_k)G_{ijk}(z(s_i), z(s_j), z(s_k)) + \cdots + 0$$

Pues $z(s_1) = 0$. Por tanto se tiene la siguiente expresión

$$Q(z) - Q(z_{0;1}) = z(s_1)\Bigg(G_1(z(s_1)) + \sum_{2\leq j\leq n} z(s_j)G_{1j}(z(s_1), z(s_j)) + \sum_{1\leq j<k\leq n}\sum z(s_j)z(s_k)G_{1jk}(z(s_1), z(s_j), z(s_k)) + \ldots + z(s_2)\ldots z(s_n)G_{1,2,\ldots,n}(z(s_1)\ldots z(s_n))\Bigg)$$

De la proposición 1.10 se sabe que si el área l no es vecina del área 1 entonces $Q(z)-Q(z_{0;1})$ no depende de $z(s_l)$. Además, como la expansión anterior se ha de cumplir para toda realización de Z en la retícula, podemos tomar $z(s_i) = 0\ \forall i \notin \{1, l\}$ y se obtiene

$$Q(z) - Q(z_{0;1}) = z(s_1)G_1(z(s_1)) + z(S_1)z(s_l)G_{1l}(z(S_1), z(s_l))$$

Como G_{1l} ha de depender explícitamente de $z(s_l)$, la única opción para que $Q(z)-Q(z_{0;1})$ no dependa de $z(s_l)$ es que G_{1l} sea identicamente 0.

Si se repite este desarrollo con el resto de regiones no vecinas, se llega al resultado. □

El teorema de Hammersley-Clifford presenta un resultado de gran importancia como corolario que relaciona la función de distribución conjunta con la función negpotencial.

Corolario 1.13. *Si ζ es un conjunto contable en $\mathbb{R}^n$ (o es un conjunto Lebesgue-medible en el caso continuo) y se pueden obtener G-funciones bien definidas respecto a una estructura MRF, entonces si se cumple la condición de sumabilidad*

$$\sum_{z\in\zeta}\exp(Q(z)) < \infty \left[\text{O en el caso continuo} \int_\zeta \exp(Q(z))dz < \infty\right]$$

Entonces Q define una única distribución conjunta dada (en el caso discreto) por

$$\Pr(z) = \exp(Q(z))/\sum_{y\in\zeta}\exp(Q(y)) \;\forall\, z \in \zeta$$

La demostración puede verse en [1, p. 419]

Tan solo son necesarios dos teoremas más para completar el marco teórico que permite construir la distribución de los modelos condicionadamente especificados.

El primero, da una simplificación de la distribución condicionada para distribuciones en la familia exponencial. Esta distribución condicionada, en el caso de variable discreta, viene dada por [1, p. 419]

$$\Pr(z(s_i) \mid \{z(s_j) : i \neq j\}) = \exp\Big(A_i(\{z(s_j) : i \neq j\})B_i(z(s_i)) + C_i(z(s_i)) + D_i(\{z(s_j) : i \neq j\}\Big)\; i = 1, \ldots, n \quad (1.12)$$

Teorema 1.14 (Teorema de Besag)**.** *Se dice que un proceso tiene pairwise-only dependence si todas la G-funciones G_A son idénticamente nulas si A tiene más de dos elementos. Esto no quiere decir que se aplique el teorema de Hammersley-Clifford, es decir, no es necesario que los cliques tengan como máximo dos elementos, sino que algunas de las G-funciones que no tenían porqué ser cero, lo son.*

En caso de que se dé este tipo de dependencia

$$A_i(\{z(s_j) : i \neq j\}) = \alpha_i + \sum_{j=1}^{n}\theta_{ij}B_j(z(s_j)) \quad (1.13)$$

Con $\theta_{ij} = \theta_{ji}$, $\theta_{ii} = 0$ y $\theta_{ik} = 0$ para todo k no vecino de i.

La prueba de este teorema se puede encontrar en [1, p. 420]

Por otro lado, es necesario también el teoorema de Factorización, que da un vínculo entre a probabilidad conjunta y la condicionada.

Teorema 1.15 (Teorema de factorización)**.** *Supóngase que $\{Z(s_i) : i = 1, \ldots, n\}$ tienen una función de masa conjunta* Pr *(o función de densidad f en el caso continuo) y que cumple la condición de positividad. En ese caso*

$$\frac{\Pr(z)}{\Pr(y)} = \prod_{i=1}^{n}\frac{\Pr(z(s_i) \mid z(s_1), \ldots, z(s_{i-1}), y(s_{i+1}, \ldots, y(s_n))}{\Pr(y(s_i) \mid z(s_1), \ldots, z(s_{i-1}), y(s_{i+1}, \ldots, y(s_n))} \;\forall\, z, y \in \zeta \quad (1.14)$$

Su demostración se ve en [1, p.413]

1.4.2. De la probabilidad condicionada al modelo condicionalmente especificado

Con las definiciones y teoremas anteriores, ya es posible dibujar el paso de 1.8 a 1.4 y por lo tanto definir rigurosamente los modelos condicionalmente especificados.

Teorema 1.16. *Para $\{Z(s) : s \in D\}$ proceso gaussiano sobre una retícula que sea un campo aleatorio de Markov que cumpla la condición de positividad y con pairwise-only dependence entre regiones, se tiene que Z tiene como distribución conjunta:*

$$Z \sim \mathcal{N}(\mu, (I - C)^{-1}M) \tag{1.15}$$

Cumpliendo

- $Z = (Z(s_1), \ldots, Z(s_n))$
- *μ es el vector de medias $\mu = (\mu_1, \ldots, \mu_n)$ con $\mu_i = \mathbb{E}(Z(s_i))$*
- *$(I - C)$ es invertible*
- *$(I - C)^{-1}M$ es simétrica y definida positiva*
- *$C = (c_{ij})$ es una matriz $n \times n$*
- *M es una matriz diagonal $n \times n$ $M = diag(\tau_1^2, \ldots, \tau_n^2)$*
- *$c_{ii} = 0$, $c_{ik} = 0$ si i, k no son vecinos*

Demostración. Para comenzar, $Z(s_i)$ tiene una distribución normal, la distribución condicionada $f(z(s_i) \mid \{z(s_j) : j \neq i)\}$ viene dada por 1.8. Sin embargo, como la normal forma parte de las distribuciones exponenciales, 1.12 da otra forma de calcular la distribución condicionada.

Juntando ambas definiciones teniendo en cuenta que el exponente de 1.8.

$$\begin{aligned} f(z(s_i) \mid \{z(s_j) : j \neq i)\}) = \\ &= \frac{1}{\tau_i\sqrt{2\pi}} \exp(-(z(s_i) - \theta_i(\{z(s_j) : j \neq i\}))^2/2\tau_i^2) \\ &= \exp\left(A_i(\{z(s_j) : i \neq j\})B_i(z(s_i)) + C_i(z(s_i)) + D_i(\{z(s_j) : i \neq j\}\right) \\ &= \exp\left(\theta_i(\{z(s_j) : i \neq j\})\tau_i^{-2}z(s_i) + C_i(z(s_i)) + D_i(\{z(s_j) : i \neq j\})\right) \\ & \qquad\qquad i = 1, \ldots, n \end{aligned} \tag{1.16}$$

Pues si se reescribe 1.8 como exponencial:

$$f(z(s_i) \mid \{z(s_j) : j \neq i)\}) = \frac{1}{\tau_i\sqrt{2\pi}} \exp(-(z(s_i) - \theta_i(\{z(s_j) : j \neq i\}))^2/2\tau_i^2)$$
$$= \exp\left(\log\left(\frac{1}{\tau_i\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{z(s_i)}{\tau_i} - \frac{\theta_i(\{z(s_j) : j \neq i\})}{\tau_i}\right)^2 \right) \quad (1.17)$$

Es posible expresarla como 1.16 para ciertas funciones D_i, C_i

Ahora, como $\{Z(s) : s \in D\}$ es un campo aleatorio de Markov con la condición de positividad y con *pairwise-only dependence* entre regiones, entonces por el teorema de Besag 1.14.

$$\theta_i(\{z(s_j) : j \neq i\}) = \mu_i + \sum_{j=1}^{n} c_{ij}(z(s_j) - \mu_j) \quad (1.18)$$

Con $c_{ij}\tau_j^2 = c_{ij}\tau_i^2$, $c_{ii} = 0$, $c_{i,k} = 0$ si i, k no son vecinos, y $\mu_i = \mathbb{E}(Z(s_i))$.

Por tanto

$$Z(s_i) \mid \{z(s_j) : i \neq j\} \sim \mathcal{N}\left(\mu_i + \sum_{j=1}^{n} c_{ij}(z(s_j) - \mu_j), \tau_i^2\right) \quad (1.19)$$

Ahora, aplicando el teorema de factorización 1.15 para $y = \mu$.

$$\log\left(\frac{f(z)}{f(\mu)}\right)$$
$$= \log\left(\prod_{i=1}^{n} \frac{f(z(s_i) \mid z(s_1), \ldots, z(s_{i-1}), \mu(s_{i+1}, \ldots, \mu(s_n))}{f(\mu(s_i) \mid z(s_1), \ldots, z(s_{i-1}), \mu(s_{i+1}, \ldots, \mu(s_n))}\right)$$
$$= \sum_{i=1}^{n} \Big(\log\left(f(z(s_i) \mid z(s_1), \ldots, z(s_{i-1}), \mu(s_{i+1}, \ldots, \mu(s_n))\right)$$
$$- \log\left(f(\mu(s_i) \mid z(s_1), \ldots, z(s_{i-1}), \mu(s_{i+1}, \ldots, \mu(s_n))\right) \Big)$$

$$= \sum_{i=1}^{n} \left(\log\left(\frac{1}{\tau_i\sqrt{2\pi}} \exp\left(-(z(s_i) - \theta_i(\{z(s_i) \mid z(s_1), \ldots, z(s_{i-1}), \mu(s_{i+1}, \ldots, \mu(s_n)\}))^2/2\tau_i^2\right)\right) \right.$$
$$\left. -\log\left(\frac{1}{\tau_i\sqrt{2\pi}} \exp\left(-(\mu(s_i)) - \theta_i(\{\mu(s_i) \mid z(s_1), \ldots, z(s_{i-1}), \mu(s_{i+1}, \ldots, \mu(s_n)\}))^2/2\tau_i^2\right)\right) \right)$$

Aplicando la definición de las medias condicionadas 1.18

$$\begin{aligned}
&\log\left(\frac{f(z)}{f(\mu)}\right) \\
&\quad = \sum_{i=1}^{n}\Bigg(-(1/2)(z(s_i)-\theta_i(\{z(s_i)\mid z(s_1),\ldots,z(s_{i-1}),\mu(s_{i+1},\ldots,\mu(s_n)\}))^2/\tau_i^2 \\
&\qquad +(1/2)(\mu(s_i))-\theta_i(\{\mu(s_i)\mid z(s_1),\ldots,z(s_{i-1}),\mu(s_{i+1},\ldots,\mu(s_n)\}))^2/\tau_i^2\Bigg) \\
&= -(1/2)\sum_{i=1}^{n}\left((z(s_i)-\mu_i-\sum_{j=1}^{i-1}c_{ij}(z(s_j)-\mu_j))^2/\tau_i^2\right)+(1/2)\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{i-1}c_{ij}(z(s_j)-\mu_j)\right)^2/\tau_i^2\right) \\
&\quad = -(1/2)\sum_{i=1}^{n}\left((z(s_i)-\mu_i)^2/\tau_i^2\right)+(1/2)\sum_{i=1}^{n}\left(\sum_{j=1}^{i-1}c_{ij}(z(s_j)-\mu(s_i))/\tau_i^2\right) \\
&\quad = -(1/2)(z-\mu)'M^{-1}(I-C)(z-\mu)
\end{aligned}$$

Por tanto

$$\begin{aligned}
f(z) &= f(\mu)\exp\left(-(1/2)(z-\mu)'M^{-1}(I-C)(z-\mu)\right) \\
&= \frac{1}{\det((I-C)^{-1}M)(2\pi)^{n/2}}\exp\left(-(1/2)(z-\mu)'M^{-1}(I-C)(z-\mu)\right)
\end{aligned}$$

Es decir

$$Z\sim\mathcal{N}\left(\mu,(I-C)^{-1}M\right) \tag{1.20}$$

□

Es posible, por tanto, construir modelos de autorregresión análogos a los simultáneamente especificados de forma similar a 1.7, estimando por máxima verosimilitud los parámetros μ,M, C:

$$Z(s_i)=\mu_i+\sum_{j=1}^{n}c_{ij}(Z(s_j)-\mu_j)+\nu_i \tag{1.21}$$

Con los pseudoerrores definidos como $\nu=(I-C)(Z-\mu)$. Como $\mathbb{E}(\nu Z')=M$, ν_i es independiente de los valores de Z en áreas distintas al área i, a diferencia de lo que ocurría en el modelo simultáneamente especificado.

1.5. Comparación entre los modelos

En [1, pp. 408-410] se puede encontrar una discusión más detallada de las diferencias entre ambos planteamientos. Además de la diferencia de dependencia de los errores del

apartado anterior, resulta muy importante destacar que a diferencia del caso temporal, solo pueden ser equivalentes ambos modelos si las matrices de covarianzas $(I-C)^{-1}M$ y $(I-B)^{-1}\Lambda(I-B')^{-1}$ son iguales. Es posible también comprobar que todo modelo simultáneamente especificado puede reescribirse como condicionalmente especificado, pero no al revés. Esto se puede ver en [1, pp. 409], cómo en los modelos de retícula regular un modelo simultáneamente especificado que tiene en cuenta solo los vecinos más cercanos (N,S,E,O) tiene una equivalencia en un modelo condicionalmente especificado que tiene en cuenta los vecinos hasta un tercer grado de proximidad y, por lo tanto, modelos condicionales de primer o de segundo orden no tienen equivalencia en modelos simultáneos.

1.6. Modelos de regresión con efectos de otras variables explicativas

Cuando se realizan estudios sobre datos reales normalmente los procesos en retículas $Z(s) : s \in D$ no solo dependen de la estructura espacial si no de otras variables independientes. Se trata por tanto de unir las ideas de dependencia espacial y autorregresión con los modelos más sencillos de regresión lineal. Así se pueden separar los efectos sobre Z en dos [1, pp. 136 137] la llamada "variación a gran escala" que se corresponde con la parte que el modelo explica sobre la media de Z y "variación a pequeña escala" de la matriz de covarianzas que explica la dependencia espacial.

1.6.1. Modelos autorregresivos condicionales (CAR)

Se supone que Z depende de q variables explicativas que pueden depender o no del área en la retícula. En este caso se plantea un modelo de regresión general de forma que la media de Z se explica por

$$\mathbb{E}(Z) = \mu = X\beta$$

Con X una matrix $n \times q$ en a que las columnas son las variables explicativas y β es el vector de q parámetros a estimar. Equivalentemente, para cada localización

$$\mathbb{E}(Z(s_i)) = \sum_{j=1}^{q} x_j(s_i)\beta_j$$

Ahora bien, teniendo en cuenta la especificación condicional

$$\mathbb{E}(Z) = \mu = X\beta + \delta$$

Con $\delta \sim \mathcal{N}(0, (I-C)^{-1}M)$. Si $C = 0$ y $M = \tau^2 I$ el modelo CAR se reduce a un modelo lineal. Para C, M conocidas el estimador de mínimos cuadrados generalizados pasa a ser

[1, p. 436]

$$\beta^* = (X'M^{-1}(I-C)X)^{-1}X'M^{-1}(I-C)Z$$

Para obtener los valores de C es común expresar $C = \lambda H$ con H una de las matrices de pesos definidas por los métodos del capítulo siguiente.

Test de influencia espacial

Para comprobar que existe un efecto espacial, en modelos sencillos se puede implementar el siguiente test de hipótesis [1, p. 438].

Para $C = \lambda H$, $M = \tau^2 I$ y modelando bajo CAR,

$$\mathbf{H}_0 : \lambda = \lambda_0$$
$$\mathbf{H}_a : \lambda \neq \lambda_0$$

Rechazando $\mathbf{H}_0$ en caso de que el estadístico de contraste:

$$O = \frac{(Z - X\widehat{\beta})'(I - \lambda_0 H)H(Z - X\widehat{\beta})}{(Z - X\widehat{\beta})'(I - \lambda_0 H)(Z - X\widehat{\beta})} \tag{1.22}$$

Con $\widehat{\beta}$ es el estimador usual de regresión lineal $\beta = (X'X)^{-1}X'Z$. O es asintóticamente normal lo que permite construir intervalos de confianza.

1.6.2. Modelos autorregresivos simultáneos (SAR)

Siguiendo los mismos pasos que el caso anterior es posible a partir de 1.7 se puede escribir un modelo que también explique la media a partir de q variables explicativas [6, pp.293,294]

$$Z = X'\beta + B(Z - X'\beta) + \epsilon \tag{1.23}$$

con ϵ_i errores independientemente distribuidos, normales de media cero y covarianza diagonal (muchas veces con elementos iguales en la diagonal)

Al igual que en el modelo CAR, se suele expresar $B = \lambda W$ con W una matriz conocida (obtenida por los métodos del apartado siguiente), de forma que el modelo pasa a escribirse como

$$Z = X'\beta + \lambda W(Z - X'\beta) + \epsilon \tag{1.24}$$

Test de influencia espacial

Para comprobar que existe un efecto espacial en el caso del SAR con $B = \lambda W$ se puede emplear el test siguiente [1, p. 442].

$$\mathbf{H}_0 : \lambda = \lambda_0$$
$$\mathbf{H}_a : \lambda \neq \lambda_0$$

Escribiendo $e = Z - X\widehat{\beta}$ con $\widehat{\beta}$ el estimador usual de mínimos cuadrados, entonces el estadístico de contraste es:

$$I = \frac{e'We}{e'e} \tag{1.25}$$

Que coincide con la I de Moran que se tratará más en profundidad en 3.

1.6.3. Modelos para datos no gaussianos

En casos de que los datos no sean gaussianos es posible construir modelos similares al SAR y CAR, puesto que las discusiones anteriores a su construcción no dependían del carácter normal de Z[1]. Sin embargo, para muchas de estas distribuciones resulta complicado realizar inferencia sobre los estimadores por dificultad a la hora de calcular o maximizar la verosimilitud. Usualmente [6] es mejor idea transformar los datos si es posible para que presenten una distribución asintóticamente normal. Por ejemplo, para datos con distribución Poisson que aparecen a menudo en estudios de enfermedades se pueden utilizar ratios en vez de casos como se ve en el ejemplo del SIDS de Cressie en [1]. Debido al teorema central del límite es sencillo entender el porqué de dar tanta importancia al caso gaussiano.

Capítulo 2

Relación de vecindad y pesos espaciales

Como se estableció al final del capítulo anterior, al realizar modelos de regresión de datos reticulares, tanto simultáneamente especificados como condicionalmente especificados, con varias variables explicativas (variación a gran y pequeña escala), resulta imposible (o de escaso valor) estimar todos los efectos espaciales. Se requiere, por tanto, simplificar la influencia espacial. La forma más común [6] de hacer esto es, en lugar de estimar todas las componentes espaciales (b_{ij}, por ejemplo, en el caso del SAR), se den razonablemente fijadas y se estime el grado de influencia como un parámetro de escala que multiplique estos pesos espaciales prefijados. En este capítulo, se resumen las distintas formas de relacionar áreas y poder aplicar los modelos de regresión.

Establecer la relación existente entre regiones puede considerarse como un proceso doble. Por un lado, ha de determinarse para cada nodo qué regiones se consideran vecinas y, por otra parte, el grado de influencia que tiene cada vecino. Esta información toma forma de *pesos espaciales* que se pueden expresar matricialmente como la *matriz de pesos*.

Definición 2.1. Si llamamos n al número de regiones es decir $n = \#I$ se define la matriz de pesos $N \times N$ como

$$\mathbf{W} = (w_{i,j}) = \begin{pmatrix} 0 & w_{1,2} & \cdots & w_{1,j} & \cdots & w_{1,n} \\ w_{2,1} & 0 & \cdots & w_{2,j} & \cdots & w_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{i,1} & w_{i,2} & \cdots & 0 & \cdots & w_{i,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{N,1} & w_{n,2} & \cdots & w_{n,2} & \cdots & 0 \end{pmatrix} \tag{2.1}$$

donde para todo i, j los elementos $w_{i,j}$ son fijos (es decir, no aleatorios), no negativos y finitos[5, pp. 39]. Existen numerosas formas de establecer los pesos $w_{i,j}$, se suelen enumerar [5][8][4][3] varias categorías:

- Pesos basados en la distancia
- Pesos basados en la adyacencia
- Pesos basados en los centroides

2.1. Distancias

Es natural pensar que la distancia entre regiones tiene un efecto sobre los datos. Regiones próximas deberían tener valores similares. Esta idea lleva a construir una matriz de distancias entre los distintos nodos de D_N; es decir, la distancia entre los centros de cada región [5, pp. 34-37]. La distancia escogida es también objeto de discusión y depende de las propiedades de los datos. Además de la distancia euclídea, la distancia de Manhattan o distancia L^1 puede ser más útil para modelar áreas muy urbanizadas debido a la estructura cuadriculada de grandes ciudades o en casos donde sea de interés que la propiedad $d_{AC} = d_{AB} + d_{BC}$ se cumpla. En caso de que el proceso estudiado dependa del desplazamiento humano, puede diseñarse una distancia que sea función del tiempo de viaje y del coste asociado o del porcentaje de industrialización de las regiones [1, pp. 385].

Los datos de distancias entre centros se pueden expresar en forma de la matriz de distancias.

$$\mathbf{D} = (d_{i,j}) = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,j} & \cdots & d_{1,n} \\ d_{2,1} & 0 & \cdots & d_{2,j} & \cdots & d_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{i,1} & d_{i,2} & \cdots & 0 & \cdots & d_{i,n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ d_{n,1} & d_{n,2} & \cdots & d_{n,2} & \cdots & 0 \end{pmatrix} \tag{2.2}$$

Donde $d_{i,j}$ es la distancia escogida. Nótese que $\mathbf{D}$ es simétrica y su diagonal es cero, pues la distancia de un punto a sí mismo es nula.

2.2. Pesos basados en la distancia

2.2.1. Distancia fija

Posiblemente la manera más sencilla de construir vecindades teniendo en cuenta la distancia es considerar como vecinos de un una región a todas las regiones que estén por

debajo de una cierta distancia crítica $\overline{d}$, es decir

$$w_{i,j} = \begin{cases} 1 & d_{i,j} \leq \overline{d} \; \forall \, i,j = 1,\ldots,n \; ; \; i \neq j \\ 0 & \text{en otro caso} \end{cases} \tag{2.3}$$

Dependiendo de cómo estén repartidas las regiones, este método puede provocar que áreas con regiones muy pequeñas tengan muchos vecinos, mientras que en otras con una densidad menor de regiones se puede dar que nodos no tengan ningún vecino.

2.2.2. *k-nearest neighbor*

Para solucionar los problemas anteriores, se puede proponer que cada localización tenga una cantidad fija de vecinos, escogiendo las k regiones más cercanas.

$$w_{i,j} = \begin{cases} 1 & d_{i,j} \leq \overline{d}_{i(k)} \; \forall \, i,j = 1,\ldots,n \; ; \; i \neq j \\ 0 & \text{en otro caso} \end{cases} \tag{2.4}$$

Uno de los problemas de estas construcciones es que son relaciones binarias, pues los pesos son ceros o unos y no dan una idea de la magnitud del efecto de cada región. Parece lógico pensar que la influencia de los vecinos decrece con la distancia.

2.2.3. Pesos continuos basados en la distancia

De la idea anterior se origina la siguiente definición de pesos espaciales que dependen del inverso de la distancia:

$$w_{i,j} = \begin{cases} d_{i,j}^{-\theta} & \forall \, i,j = 1,\ldots,n \; ; \; i \neq j \\ 0 & \forall \, i = j \end{cases} \tag{2.5}$$

Donde el parámetro θ permite penalizar más o menos rápidamente la distancia entre nodos. Usualmente [5][4] θ toma el valor 1 o 2, pero pueden tomarse otros valores o puede dejarse desconocido y estimarse a partir de los datos.

Otra forma de implementar la caída de influencia por distancia es usando una exponencial negativa:

$$w_{i,j} = \begin{cases} e^{-d_{i,j}} & \forall \, i,j = 1,\ldots,n \; ; \; i \neq j \\ 0 & \forall \, i = j \end{cases} \tag{2.6}$$

De esta forma, se da una mayor importancia a los datos cercanos que con los diseños anteriores. Una vez más, θ puede ser un valor fijado o un estadístico a determinar.[3]

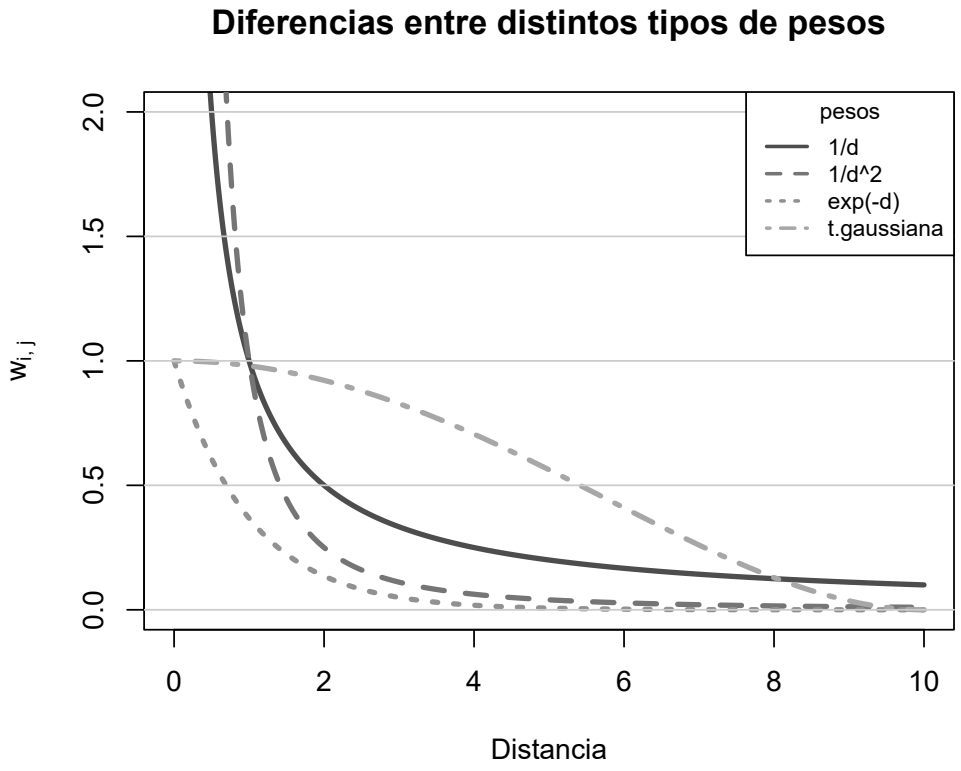


Figura 2.1: Comparación del efecto de cada transformación de la distancia para construir cada peso

Estas definiciones de pesos continuos pueden combinarse con 2.3 y con 2.4 sustituyendo el valor 1 por el calculado en casos continuos, por ejemplo, combinando 2.3 y 2.6 se obtiene la regla:

$$w_{i,j} = \begin{cases} e^{-\theta d_{i,j}} & d_{i,j} \leq \overline{d} \; \forall \, i,j = 1,\ldots,n \; ; \; i \neq j \\ 0 & \text{en otro caso} \end{cases} \tag{2.7}$$

2.2.4. Transformación gaussiana

Esta transformación está diseñada para que los pesos decrezcan con menor velocidad, sobre todo en los valores menores. Para su definición se requiere siempre dar una distancia límite que debe ser escogida. Los pesos vienen dados por:

$$w_{i,j} = \begin{cases} (1 - (\frac{d_{i,j}}{\overline{d}})^2)^2 & d_{i,j} \leq \overline{d} \; \forall \, i,j = 1,\ldots,n \; ; \; i \neq j \\ 0 & \text{en otro caso} \end{cases} \tag{2.8}$$

La variación de cada tipo de peso con respecto a la distancia entre regiones se puede apreciar en la figura 2.1

2.3. Pesos basados en la adyacencia

Esta clase de pesos relacionan las regiones si ambas son adyacentes, es decir, si comparten puntos de su frontera. Originalmente, estas relaciones fueron diseñadas para datos estructurados en retículas regulares, es decir, cuadrículas. Usados en este caso, estas vecinanzas toman una forma muy sencilla y por ello son muy populares en la literatura para explicar los procesos espaciales[1][9, pp. 78][3, pp. 8]. Primeramente, resaltaremos la definición de estos pesos en cuadrículas para luego extenderlo a la definición formal a retículas no necesariamente regulares. Se definen tres tipos distintos de pesos, en analogía con los movimientos de piezas de ajedrez:

- **Relación de torre**: en este caso, los vecinos de cada casilla son las casillas inmediatamente al norte, sur, este y oeste. De forma general se puede expresar

$$w_{i,j} = \begin{cases} 1 & \text{Las fronteras de } i \text{ y } j \text{ comparten una cantidad no contable de puntos} \\ 0 & \text{en otro caso} \end{cases} \tag{2.9}$$

- **Relación de dama**: los vecinos de cada casilla son las casillas que la rodean, es decir, las que e son vecinos por la relación de torre y, además, las ubicadas inmediatamente NE,NW,SE y SW. De forma general:

$$w_{i,j} = \begin{cases} 1 & \text{Las fronteras de } i \text{ y } j \text{ comparten al menos un punto} \\ 0 & \text{en otro caso} \end{cases} \tag{2.10}$$

Existe también una relación de alfil en a que los vecinos solo comparten no más que una cantidad finita de puntos pero no suele resultar útil.

Además, [4] propone una construción no binaria de la relación de torre mediante la relación

$$w_{i,j} = \begin{cases} l_{i,j}/l_i & \text{Las fronteras de } i \text{ y } j \text{ comparten al menos un punto} \\ 0 & \text{en otro caso} \end{cases} \tag{2.11}$$

con l_i el perímetro de la frontera de i y $l_{i,j}$ la longitud de frontera compartida.

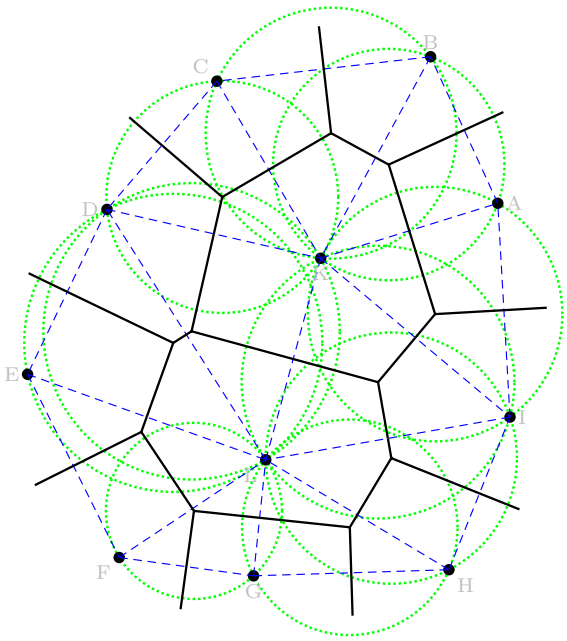


Figura 2.2: En negro, diagrama de Voronoi; en azul, triángulos de Delaunay asociados con sus circunferencias circunscritas en verde.

2.4. Pesos basados en los centroides

Estas estructuras se basan, igual que las basadas en distancia, en tratar las regiones como si solo fueran sus centros. En este caso las relaciones se establecen de forma que el grafo resultante presenta ciertas condiciones. Dos construcciones de este tipo que aparecen en [8] y en [9, pp. 80] son la triangulación de Delaunay y los gráficos de esfera de influencia (SoI).

2.4.1. Vecindades de Delaunay

Para construir pesos de esta forma se realiza el proceso siguiente: en primer lugar, se representan los centros de las regiones en su localización en e espacio; después se realiza una triangulación de Delaunay de los puntos, es decir, se dibujan aristas entre nodos de forma que para cualquier triángulo en el grafo formado, el circulo que lo circunscribe no incluye a ningún otro nodo; finamente se considera $w_{i,j} = 1$ si la triangulación anterior contiene una arista de i a j. La construcción de la triangulación es equivalente a separar el espacio en celdas de Dirichlet (también conocido como diagrama de Voronoi) y utilizar la relación de torre sobre estas regiones, más información sobre esto puede encontrarse en [7, sección 4.3].

2.4.2. Vecindades de esferas de influencia

Se define la esfera de influencia de un nodo como la circunferencia centrada en el nodo y con radio igual a la distancia al nodo más cercano. De esta forma $w_{i,j} = 1$ si las esferas de influencia de i y de j se cortan en dos puntos, y cero en otro caso.

Por último, sobre cualquiera de los pesos tratados en estas secciones se pueden añadir otros datos propios de cada nodo para construir nuevos pesos. Por ejemplo si las regiones y el experimento planteado tiene que ver con comportamiento humano o enfermedades, puede ser de interés tener en cuenta los valores de población estandarizada en cada región, es decir, $e_i = \frac{p_i - \overline{p}_i}{s}$ donde $\overline{p}_i, s$ son la media y desviación típica muestrales del número de habitantes en cada región. Una forma de computar estos *efectos de nodo* sobre los pesos es simplemente $\widehat{w}_{i,j} = e_i e_j w_{i,j}$, donde $w_{i,j}$ se ha calculado por cualquier otro método.

Capítulo 3

Medidas de autocorrelación espacial

El propósito de la construcción de modelos de regresión basados en datos espaciales o reticulares no es sólo con objetivo de obtener métodos de predicción; sino también información del efecto espacial sobre los datos. Esta es la idea detrás de la autocorrelación, es decir, la correlación entre las observaciones de una variable por la "proximidad" (definida por la matriz de pesos) entre ellas en el espacio. La forma de estudiar la autocorrelación es emplear estadísticos que son usados en un test de hipótesis sobre la relación espacial de los datos.

Existen dos enfoques, estudiar si existe o no correlación entre todos los datos (estadísticos globales) y estudiar qué datos o conjuntos de datos influyen en mayor medida (estadísticos locales).

3.1. Medidas globales de autocorrelación

El estadístico más utilizado para medir la autocorrelación global es la I de Moran. Otros estadísticos de autocorrelación global incluyen la C de Geary y el MEET de Tango [4]. La I de Moran de una muestra dada se calcula mediante la expresión siguiente [3]:

$$I = \frac{n}{W_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n}(z_i - \overline{z})^2} \tag{3.1}$$

Donde $\overline{z}$ es la media muestral de los datos y $W_o = \sum_{i=1}^{n} \sum_{j \neq i}^{n} w_{i,j}$

En el denominador de 3.1 se encuentra la varianza muestral de los z_i y en el numerador la matriz de pesos actúa como forma cuadrática sobre la desviación de los datos, por lo tanto, una interpretación de la I de Moran es cuánta de la varianza viene dada por los pesos. En caso de que los vecinos apenas influyeran en los datos, el numerador (y, por tanto, I) sería también muy próximo a 0. El factor n/W_0 es un factor de escala. Normalmente, se

suele trabajar con los pesos $w_{i,j}$ normalizados por filas, es decir, de forma que a cada nodo la suma de todos los pesos de sus aristas asociadas sumen uno. En ese caso $n/W_0 = 1$, a no ser que existan nodos sin vecinos.

Intuitivamente, valores altos en valor absoluto de I estarán ligados a poblaciones de mayor autocorrelación. Al igual que se realiza con otros estadísticos, resta considerar un test de hipótesis para poder dar resultados. El problema de la I de Moran a la hora de construir este test es que, a pesar de que su distribución es conocida [13], es difícil de calcular y es muy costosa de computar, pues depende de los autovalores de $\mathbf{W}$, con lo que con muchos datos, resulta inviable[18]. Usualmente, se realiza una aproximación, la cual puede hacerse varias maneras, se resaltan dos:

Aproximación normal

Si se asume que z_i son observaciones de variables aleatorias Z_i con distribución normal, entonces la distribución I de Moran se aproxima a una normal de media y varianza[3]

$$\mathbb{E}(I) = \frac{-1}{n-1} \tag{3.2}$$

$$var(I) = \frac{n^2(n-1)W_1 - n(n-1)W_2 - 2W_0}{(n+1)(n-1)^2 W_0^2} \tag{3.3}$$

Y por lo tanto se puede utilizar el z-valor $z = \frac{I - \mathbb{E}(I)}{\sqrt{var(I)}}$ para el test de hipótesis en una distribución normal estándar, rechazando la hipótesis nula cuando $|z| >> 0$

Aproximación aleatoria

Otra opción es asumir que todos las observaciones se realizan con igual probabilidad en cada región y, a continuación, "barajar" los datos entre las regiones, creando $N!$ posibles permutaciones de los datos entre las regiones. Usando estas permutaciones y calculando su correspondiente I se puede obtener una función empírica o bien aproximarla mediante un proceso de Monte Carlo de muestreos aleatorios de las $n!$ permutaciones[3]. En este caso, la hipótesis nula se rechaza si el valor de I se encuentra muy extremo en esta distribución. Como no se trata de una verdadera distribución, se utiliza un valor llamado pseudo p-valor para cuantificar el nivel. El pseudo p-valor viene dado por la fórmula[18]:

$$\widehat{p} = \frac{M+1}{R+1} \tag{3.4}$$

donde R es el número de permutaciones consideradas en la aproximación y M la cantidad de veces que en la distribución empírica aproximada aparecen valores igual o más extremos que el dado.

Cabe destacar que a diferencia de los p-valores, pseudo p-valores no pueden compararse. Si se han computado 99 Ies distintas y la I a probar es más extrema que todas las demás, su pseudo p-valor será 0,01. Si ocurre lo mismo con 999, el p-valor será 0,001, pero esto no tiene porqué significar que la influencia espacial sea mayor o más probable[19].

Interpretación de la I de Moran

Bajo hipótesis nula de ausencia de autocorrelación , la esperanza de I $\mathbb{E}(I) = \frac{-1}{n-1}$ no es cero pero tiende a cero conforme aumenta la cantidad de datos; por tanto, se rechazará si el valor de I se aleja mucho de cero. A diferencia de otros tests de hipótesis, el test sobre I da información a mayores del rechazo o no de la hipótesis nula. Si se obtiene un valor significativo y el z-valor o I es positivo se dice que hay autocorrelación positiva, es decir, datos parecidos aparecen en regiones próximas. En caso de ser negativo se dice que existe autocorrelación negativa, es decir, que regiones de valores altos se rodean de regiones con valores bajos.

Es importante señalar también que la definición de I requiere que la varianza de Z no dependa de la región i [18]. Esto no ocurre, por ejemplo, cuando se trabaja con ratios, es decir, variables de conteo que han sido transformadas para seguir una distribución normal, puesto que su varianza depende de la población en cada región. Para solucionar esto, se pueden estandarizar los ratios.

La I de Moran como pendiente de regresión. Gráfico de dispersión de Moran

Si se define el vector $e = (e_1, e_2, \ldots, e_n)$ con $e_i = (z_i - \overline{z})$ la expresión 3.1 puede darse de forma matricial como:

$$I = \frac{n}{W_0} \frac{e' \mathbf{W} e}{e' e} \tag{3.5}$$

El estimador de mínimos cuadrados de una regresión lineal simple $y = \alpha + \beta x$ viene dado por

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{3.6}$$

Por lo que I, o más correctamente $\frac{w_0}{N} I$, se puede considerar como la pendiente de regresión de $\mathbf{W}e = a + Ie$. Si se representan los puntos $(e_i, \sum_{j=1}^{n} w_{i,j} e_j$ se obtiene un gráfico que se denomina diagrama de dispersión de Moran. En este diagrama, los puntos que se encuentran en el primer cuadrante son aquellos datos elevados que se encuentran rodeados de datos elevados; en el tercer cuadrante son datos bajfos cuyos vecinos son también datos bajos. Los puntos en los cuadrantes II y IV contribuyen a reducir la autocorrelación [18]. Este diagrama da una idea por tanto, de la estructura local, sin embargo, para obtener información real con una sensibilidad, es necesario realizar tests sobre estadísticos locales.

3.2. Medidas locales de autocorrelación

Entre otras medidas de correlación local se incluyen los *Indicadores Locales de Asociación Espacial* (LISA) que se corresponden con descomposiciones de los estadísticos globales en el efecto de cada región y dato. Dado que en la sección anterior se trató en profundidad la I de Moran, se muestra aquí su indicador local asociado; sin embargo, existen también otros indicadores como la c local de Geary [3]. La I_i local de Moran viene dada para cada observación i por:

$$I_i = (z_i - \overline{z}_J) \sum_{j \in J}^{n} w_{i,j}(z_j - \overline{z}_J)^2 \tag{3.7}$$

Donde J es el conjunto de índices de vecinos, y $\overline{z}_J$ es el promedio de los z_j con $j \in J$.

Una vez más la distribución de este estadístico no es una cuestión trivial y, aunque es posible calcular varios momentos, no se conoce la distribución exacta [3, pp. 28]. La aproximación aleatoria para calcular la I global se puede utilizar aquí también. El cálculo de I_i presenta también problemas pues para áreas i, k que comparten vecinos, aparecerán correlacionadas. En estos casos es necesario aplicar correcciones sobre el nivel de significación. Así para obtener una significación α, de m comparaciones (test correlacionados) se debe exigir una significación individual α_i de α/m (Bonferroni) o de $1 - (1 - \alpha)^{\frac{1}{m}}$ (Sidák)[14, p. 96].

Capítulo 4

Aplicación sobre datos de mortalidad por cáncer de pulmón

4.1. Análisis inicial de los datos

En este capítulo se trabajará con las herramientas anteriores para obtener modelos y resultados en datos reales. Se utilizarán datos de mortalidad por cáncer de pulmón en el año 2014 obtenidos del Sistema de Información de Mortalidad del Cáncer de Galicia (SIMCA) de SERGAS [1], en los que se explorará su dependencia espacial así como su dependencia a otras covariantes obtenidas del Instituto Galego de Estadística [2], de Meteogalicia [3] y del Laboratorio de Radón de Galicia vinculado a la Universidad de Santiago de Compostela[4].

Los datos tomados del SERGAS son datos de razón de mortalidad estándar (RME) por municipio, obtenidos dividiendo la cantidad de muertes observadas entre la cantidad esperada, suponiendo que la región a estudiar tenga las mismas tasas especificas que una población de control. SERGAS realizó también un suavizado por medio de un modelo jerárquico bayesiano para poder comparar los datos entre sí. Estas medidas suavizadas serán consideradas como la variable dependiente Y de los modelos. Sin embargo, estos

[1]Datos obtenidos de la web de SIMCA (`https://www.sergas.es/Saude-publica/SIMCA-Pulmon-Distribuci%C3%B3n-xeogr%C3%A1fica-T%C3%A1boas-e-mapas`)

[2]Extraído de la web del banco de datos municipales del IGE (`https://www.ige.eu/igebdt/esq.jsp?paxina=002003003&c=-1&idioma=gl&ruta=navmunicipal.jsp%3FESP%3D`) del apartado "Poboación segundo sexo e grandes grupos de idade"

[3]Datos extraídos del informe acumulado de 2014 de meteogalicia (`https://www.meteogalicia.gal/Caire/informesCaire.action?request_locale=gl`)

[4]Datos tomados de la web del Laboratorio de Radón de Galicia (`http://www.usc.es/radongal/radon-en-galicia/tablas-de-medicions/`)

Los mapas incluídos son producidos gracias a la información geográfica obtenida del Centro Nacional de Información Geográfica, copyright , Instituto Geográfico Nacional de España.

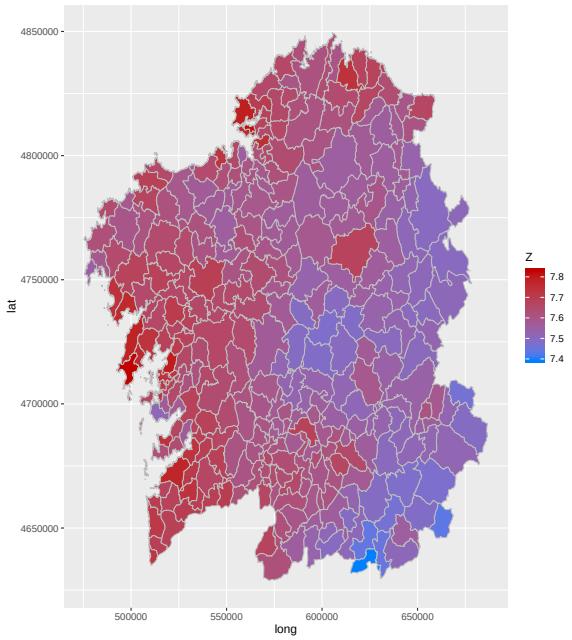


Figura 4.1: Datos de las RME modificadas

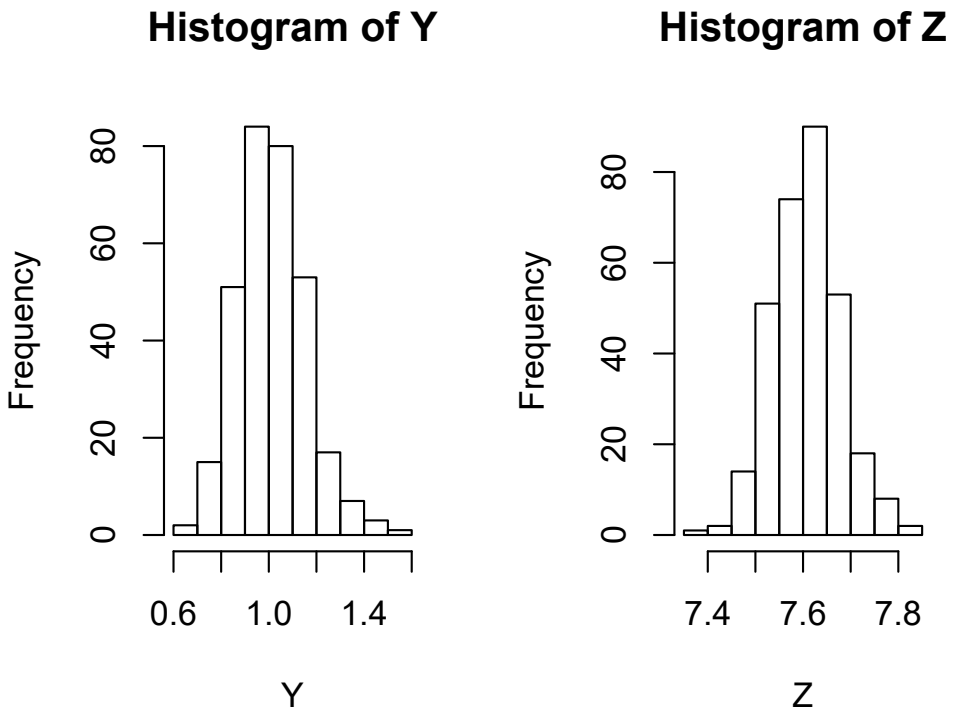
ratios no cumplen con la condición de normalidad. Utilizando un test Shapiro-Wilk se tiene un p-valor de 0,006 por lo que es significativo al 1 % y se rechaza la hipótesis nula de normalidad. Para que sea posible aplicar métodos gaussianos se aplica sobre ellos una transformación logarítmica como la que se propone en [1, p.395] o [6, p. 291].

$$Z = log(1000 \times (Y + 1))$$

Mediante un test Shapiro-Wilk se comprueba que la hipótesis gaussiana sí se puede aceptar en este caso pues se tiene un p-valor de 0,312 que no es significativo y por tanto, se acepta la hipótesis nula de normalidad.

```
  df.datos.concellos=read.csv2('df.datos.concellos.csv',check.names = F)
  Y=df.datos.concellos$RME.suavizada.T
  Z=log(1000*(Y+1))
  par(mfrow=c(1,2))
  hist(Y, breaks = 10)
  hist(Z, breaks=15)
  shapiro.test(Y)

>
>  Shapiro-Wilk normality test
>
```

```
> data:  Y
> W = 0.98694, p-value = 0.006291

  shapiro.test(Z)

>
>  Shapiro-Wilk normality test
>
> data:  Z
> W = 0.99443, p-value = 0.3121
```



Además de la posible dependencia espacial de la mortalidad por cáncer de pulmón, se consideraron otras tres variables independientes que se presentan en la figura 4.2

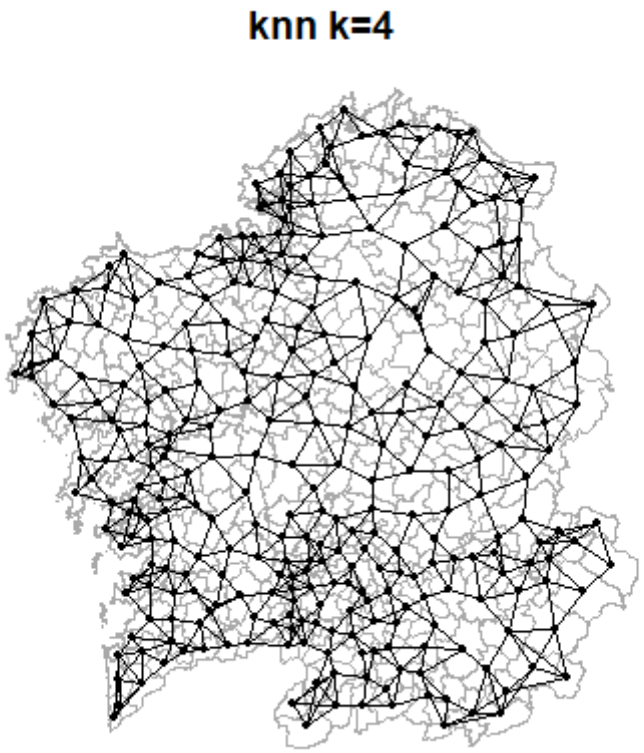
- **Medida promedio de radiación debida al radón en hogares:** Datos tomados del Laboratorio de Radón de Galicia medidos en bequerelios por metro cúbico. Existen numerosos estudios que vinculan la alta concentración de gas radón con la aparición de cáncer de pulmón[15].

- **Concentración promedio de PM10:** Datos tomados de la red Xunta de Galicia de sensores de la calidad del aire de Meteogalicia [11]. Las PM10 son partículas

contaminantes de menos de $10\mu m$ suspendidas en el aire que se consideran también como factor de riesgo para el cancer de pulmón [17]. Los datos son tomados de nueve sensores repartidos por Galicia y a cada municipio se le asigna el valor del sensor más cercano.

- **Población:** Datos del Instituto Galego de Estadística. La cantidad de población en un área o su densidad suelen ser un buenos indicadores de la industrialización o urbanización de un área.

Se estudiarán 3 posibles relaciones de vecindad: por adyacencia tipo torre, por esferas de influencia, y por KNN para $k = 4$; estas relaciones se muestran en 4.1. Además, se considerarán con pesos binarios y con pesos modificados por el inverso de la distancia. Un ejemplo de estas matrices se puede ver en 4.3.



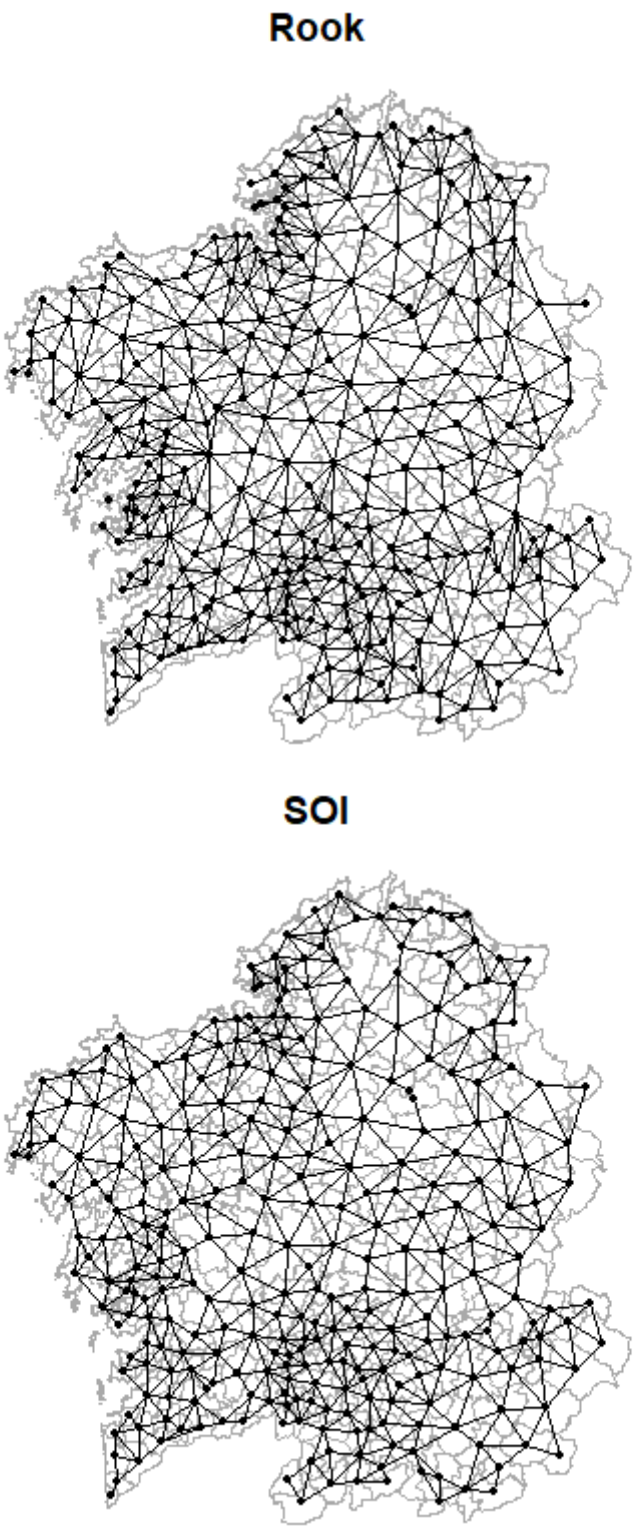


Figura 4.1: Distintas estructuras de vecindad entre municipios

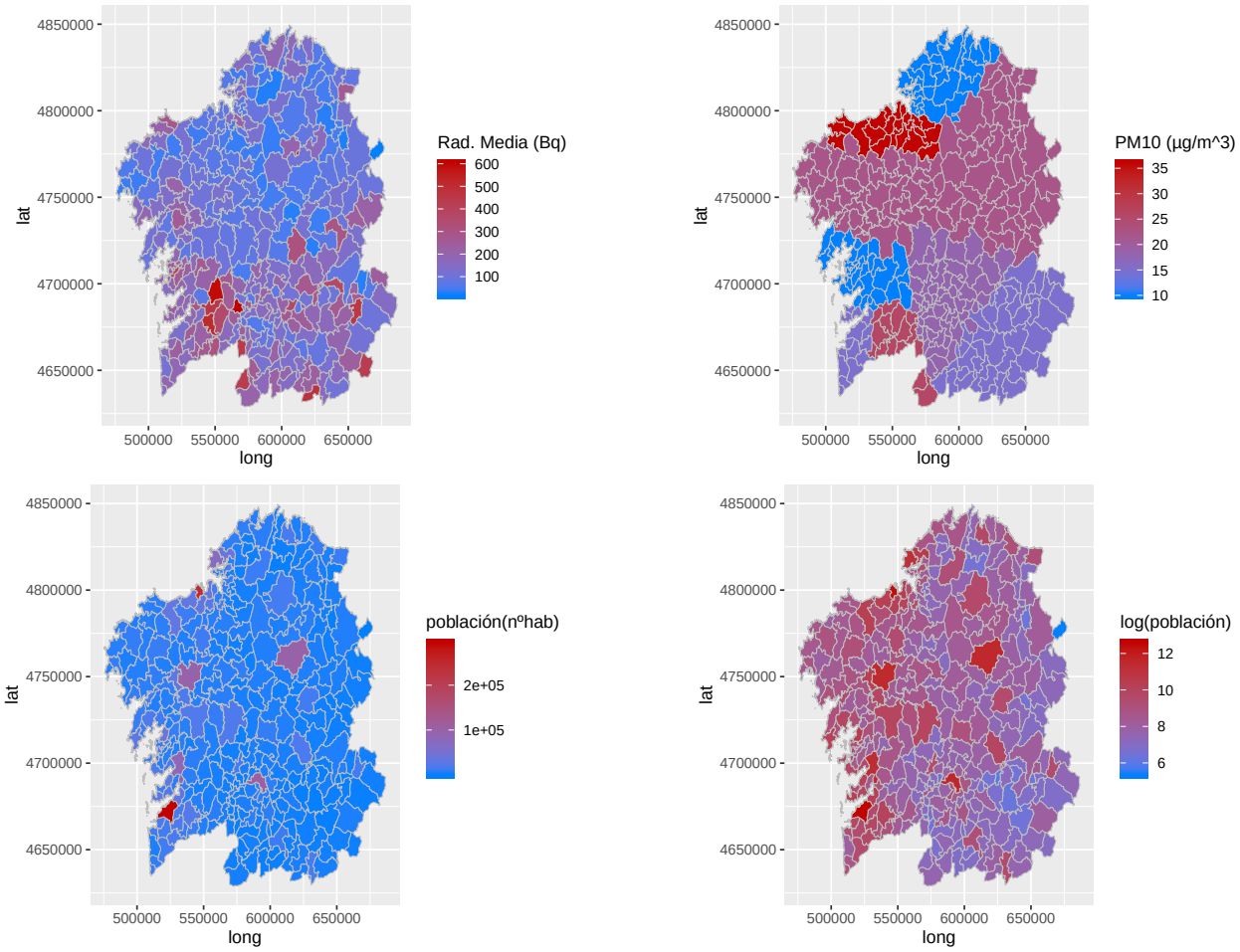


Figura 4.2: Mapas con los datos de las variables para cada municipio

```
> OGR data source with driver: ESRI Shapefile
> Source: "C:\Users\infoj\Documents\TFG\Concellos", layer: "Concellos_IGN"
> with 313 features
> It has 13 fields
```

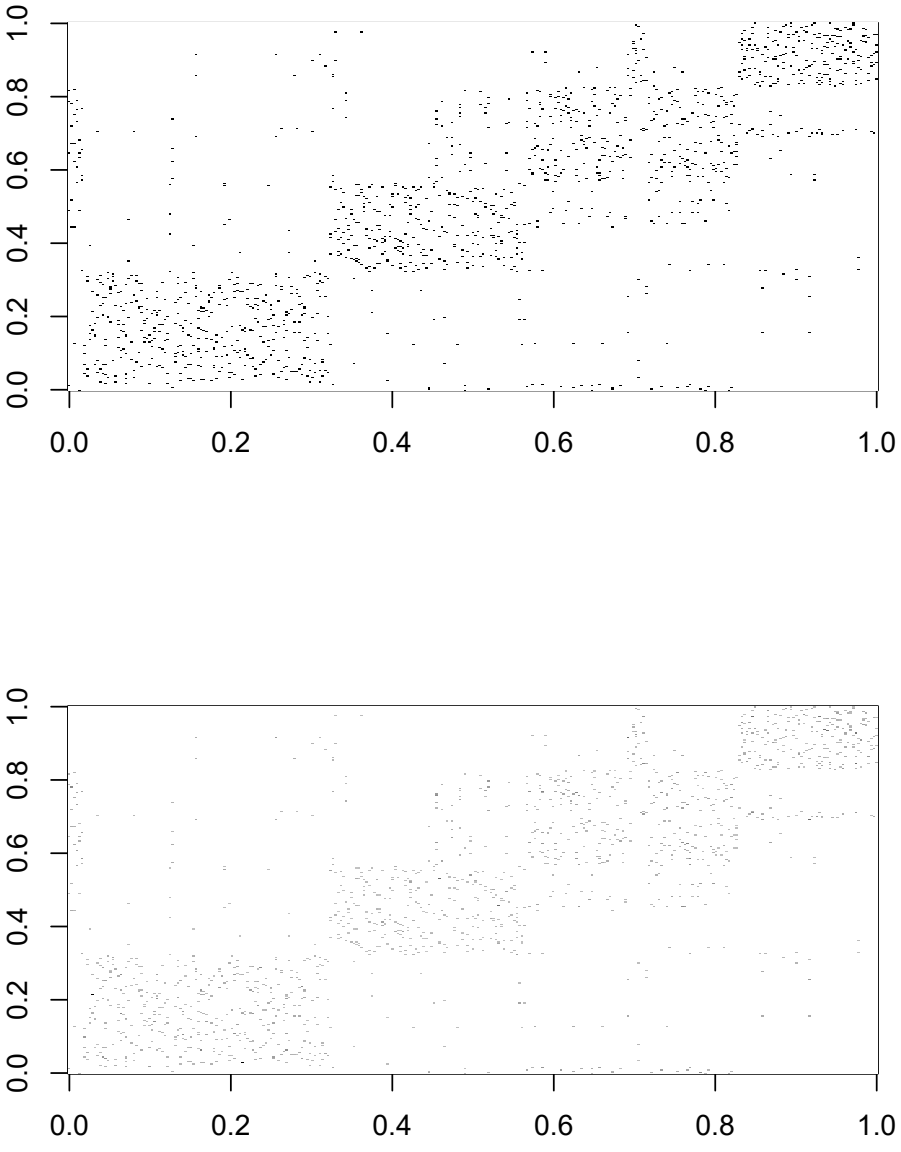
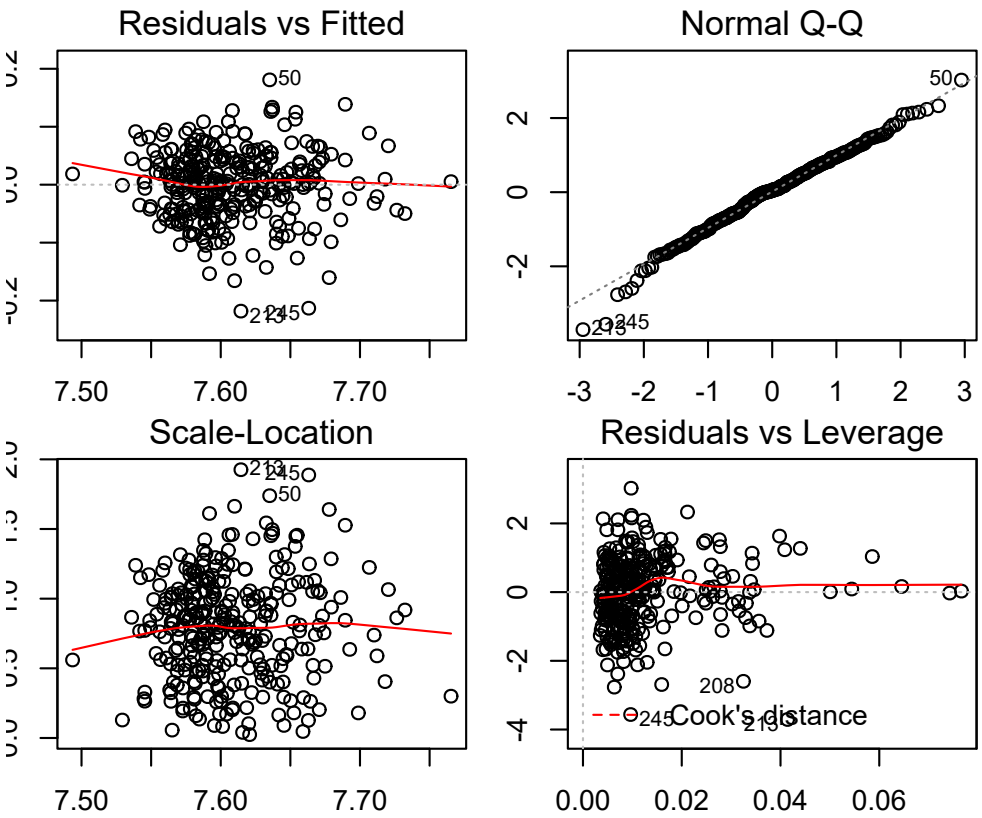


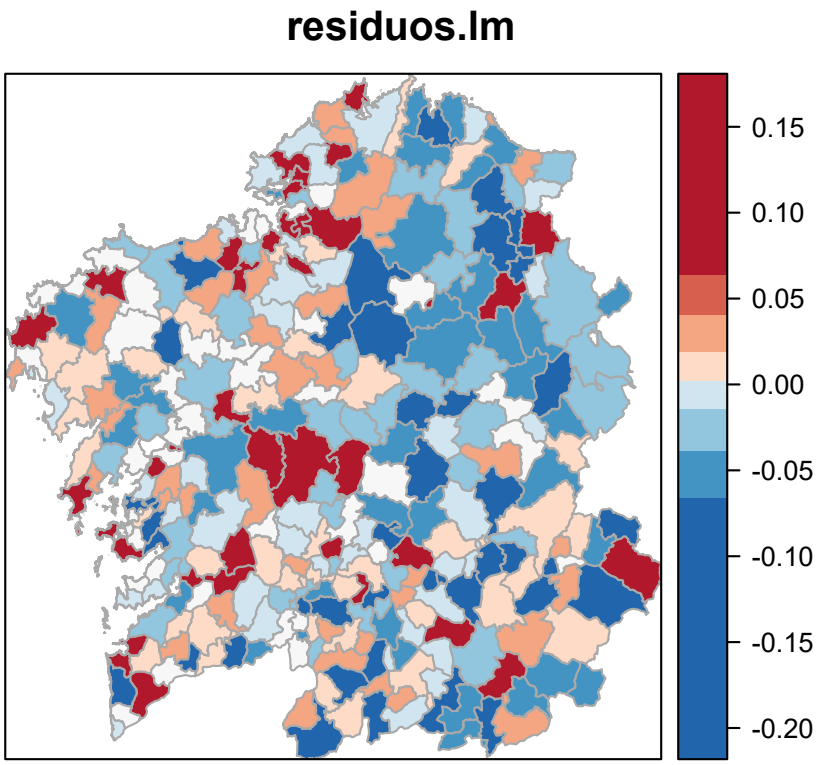
Figura 4.3: Matriz de pesos para KNN con $k = 4$ tomando vecindades binarias y con efecto de distancia

Para comenzar, se realiza un modelo de regresión lineal simple sobre los datos con respecto a las variables explicativas.

```
library(RColorBrewer)
library(rgdal, quietly = T)
library(rgeos, quietly = T)
mapa.concellos <- readOGR(dsn = "./Concellos", layer = "Concellos_IGN")

> OGR data source with driver: ESRI Shapefile
> Source: "C:\Users\infoj\Documents\TFG\Concellos", layer: "Concellos_IGN"
> with 313 features
> It has 13 fields

my.palette <- rev(brewer.pal(n = 9, name = "RdBu"))

df.datos.concellos = read.csv2("df.datos.concellos.csv")
lm = lm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T), data = df.datos.concellos)
summary(lm)

>
> Call:
> lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM +
>     log(T), data = df.datos.concellos)
>
> Residuals:
>       Min        1Q    Median        3Q       Max
> -0.218148 -0.038631  0.002583  0.040340  0.180873
>
> Coefficients:
>               Estimate Std. Error t value Pr(>|t|)
> (Intercept)  7.346e+00  2.792e-02 263.047  < 2e-16 ***
> PM10        -1.789e-03  5.136e-04  -3.483 0.000568 ***
> GM           6.467e-05  3.601e-05   1.796 0.073505 .
> log(T)       3.481e-02  3.059e-03  11.379  < 2e-16 ***
> ---
> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> Residual standard error: 0.06009 on 309 degrees of freedom
> Multiple R-squared:  0.3086,Adjusted R-squared:  0.3019
> F-statistic: 45.98 on 3 and 309 DF,  p-value: < 2.2e-16

par(mar = c(2, 1.5, 1.5, 1.5))
par(mfrow = c(2, 2))
plot(lm)
```

```
mapa.concellos$residuos.lm = residuals(lm)
grps = 9
brks <- quantile(mapa.concellos$residuos.lm, 0:(grps - 1)/(grps - 1), na.rm = TRUE)
spplot(mapa.concellos, "residuos.lm", at = brks, col = "darkgray", col.regions = my.palette,
    main = "residuos.lm")
```





Todas las covariables presentan un alto grado de significación, siendo la menos importante la radiación, con un p-valor de 0,07, y el logaritmo de la población, con un p-valor de 2×10^{-16}, la de mayor importancia. Lo único a considerar es que el efecto de los PM10 tiene un coeficiente negativo, lo cual es extraño pues implica que a mayores valores de contaminación, menos probable es la muerte por cancer de pulmón. Este dato se discute más en detalle en las conclusiones.

Como puede observarse en la taba de gráficos de control del modelo los datos respetan de forma razonable las hipótesis de normalidad (Gráfico Q-Q) y homocedasticidad (Localización-escala). En el gráfico de *Residuals vs Leverage* se aprecia como ningún dato muestra un apalancamiento excesivo en comparación con el resto, por lo que no parece que existan datos atípicos que se deban eliminar. Sin embargo, los datos pueden no respetar la hipótesis de independencia debido a efectos espaciales que se pueden apreciar en el gráfico *residuos.lm*. Para comprobar estos efectos se usa el test de Moran sobre los residuos del modelo. Se realiza el test una vez para cada tipo de vecindad de las escogidas anteriormente: tipo torre binaria, tipo torre con distancias, tipo *SOI* binaria, tipo *SOI* con distancias, tipo *KNN* con $k = 4$ binaria y tipo *KNN* con $k = 4$ con distancias.

```
> OGR data source with driver: ESRI Shapefile
> Source: "C:\Users\infoj\Documents\TFG\Concellos", layer: "Concellos_IGN"
> with 313 features
> It has 13 fields
>
>  Global Moran I for regression residuals
>
> data:
> model: lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM
> + log(T), data = df.datos.concellos)
> weights: concellos.ROOKB.lw
>
> Moran I statistic standard deviate = 1.8026, p-value = 0.03572
> alternative hypothesis: greater
> sample estimates:
> Observed Moran I      Expectation         Variance
>      0.058436723     -0.004259959      0.001209693
>
>  Global Moran I for regression residuals
>
> data:
> model: lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM
> + log(T), data = df.datos.concellos)
> weights: concellos.ROOKIDW.lw
>
> Moran I statistic standard deviate = 1.4543, p-value = 0.07292
> alternative hypothesis: greater
> sample estimates:
> Observed Moran I      Expectation         Variance
>      0.050705295     -0.004413988      0.001436382
>
>  Global Moran I for regression residuals
>
> data:
> model: lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM
> + log(T), data = df.datos.concellos)
> weights: concellos.SOIB.lw
>
> Moran I statistic standard deviate = 2.4407, p-value = 0.00733
> alternative hypothesis: greater
> sample estimates:
> Observed Moran I      Expectation         Variance
>      0.082031168     -0.004111448      0.001245704
>
```

```
>  Global Moran I for regression residuals
>
> data:
> model: lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM
> + log(T), data = df.datos.concellos)
> weights: concellos.SOIIDW.lw
>
> Moran I statistic standard deviate = 2.0799, p-value = 0.01877
> alternative hypothesis: greater
> sample estimates:
> Observed Moran I      Expectation         Variance
>      0.075090608     -0.004249604      0.001455157
>
>  Global Moran I for regression residuals
>
> data:
> model: lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM
> + log(T), data = df.datos.concellos)
> weights: concellos.4nnB.lw
>
> Moran I statistic standard deviate = 1.9566, p-value = 0.0252
> alternative hypothesis: greater
> sample estimates:
> Observed Moran I      Expectation         Variance
>      0.066797516     -0.004342536      0.001321929
>
>  Global Moran I for regression residuals
>
> data:
> model: lm(formula = log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM
> + log(T), data = df.datos.concellos)
> weights: concellos.4nnIDW.lw
>
> Moran I statistic standard deviate = 1.8515, p-value = 0.03205
> alternative hypothesis: greater
> sample estimates:
> Observed Moran I      Expectation         Variance
>      0.067980584     -0.004421702      0.001529249
```

En efecto, se puede observar que para todas las estructuras de vecindad planteadas, el test de Moran verifica que existe autocorrelación espacial. En lo siguiente, y por simplicidad, se trabajará solo con las dos estructuras espaciales distintas más significativas: la estructura de esferas de influencia con pesos binarios, con un valor de $I = 0{,}082031168$ y p-valor $p = 0{,}00733$, y la estructura generada por KNN para $k = 4$ con pesos binarios, con un

valor de $I = 0{,}06679751$ y p-valor $p = 0{,}0252$.

4.2. Modelos espaciales

Para las dos estructuras espaciales seleccionadas antes se construirán modelos SAR y CAR de la forma explicada en los apartados 1.6.2 y 1.6.1, donde las matrices B y C se consideran como el múltiplo de una dada W y se estima el factor de escala. La matriz W es, en este caso, la matriz de pesos de los pasos anteriores. Para comparar los modelos se utilizará el criterio de información de Akaike.

```
summary(slm.SAR.SOI)

>
> Call: spatialreg::spautolm(formula = formula, data = data, listw = listw,
>     na.action = na.action, family = family, method = method,
>     verbose = verbose, trs = trs, interval = interval, zero.policy = zero.policy,
>     tol.solve = tol.solve, llprof = llprof, control = control)
>
> Residuals:
>        Min         1Q     Median         3Q        Max
> -0.2163027 -0.0384020  0.0045669  0.0413688  0.1772887
>
> Coefficients:
>               Estimate  Std. Error  z value  Pr(>|z|)
> (Intercept)  7.3565e+00  2.7884e-02 263.8259 < 2.2e-16
> PM10        -1.7343e-03  5.0779e-04  -3.4154 0.0006369
> GM           6.8905e-05  3.5550e-05   1.9382 0.0525967
> log(T)       3.3263e-02  3.0747e-03  10.8184 < 2.2e-16
>
> Lambda: 0.035543 LR test value: 4.6707 p-value: 0.030682
> Numerical Hessian standard error of lambda: 0.016093
>
> Log likelihood: 440.3246
> ML residual variance (sigma squared): 0.0034892, (sigma: 0.059069)
> Number of observations: 313
> Number of parameters estimated: 6
> AIC: -868.65
```

Este es el modelo SAR con la matriz de pesos correspondiente a las vecindades de esferas de influencia. Los coeficientes vinculados a las variables explicativas son muy similares a los obtenidos en el modelo lineal general, con el mismo problema en los datos de PM10. Una vez más el elemento con menos significación es la radiación, que apenas es significativo al 5 %, y el más significativo el logaritmo de la población. Además, el factor de efecto espacial

λ, que cuantifica la importancia de la autocorrelación espacial por la estructura dada, toma el valor $\lambda = 0{,}035543$ con p-valor $p = 0{,}030682$, por lo que la estructura espacial es significativa al 5 %. Finalmente, destacar el valor de AIC $= -868{,}65$ para comparación poterior.

```
summary(slm.CAR.SOI)

>
> Call: spatialreg::spautolm(formula = formula, data = data, listw = listw,
>     na.action = na.action, family = family, method = method,
>     verbose = verbose, trs = trs, interval = interval, zero.policy = zero.policy,
>     tol.solve = tol.solve, llprof = llprof, control = control)
>
> Residuals:
>        Min         1Q     Median         3Q        Max
> -0.2183057 -0.0371597  0.0043895  0.0413414  0.1722427
>
> Coefficients:
>               Estimate  Std. Error  z value  Pr(>|z|)
> (Intercept)  7.3585e+00  2.7902e-02 263.7245 < 2.2e-16
> PM10        -1.7300e-03  5.0721e-04  -3.4107 0.0006479
> GM           6.9266e-05  3.5532e-05   1.9494 0.0512506
> log(T)       3.3009e-02  3.0780e-03  10.7242 < 2.2e-16
>
> Lambda: 0.074628 LR test value: 5.1944 p-value: 0.022659
> Numerical Hessian standard error of lambda: 0.029606
>
> Log likelihood: 440.5865
> ML residual variance (sigma squared): 0.0034518, (sigma: 0.058752)
> Number of observations: 313
> Number of parameters estimated: 6
> AIC: -869.17
```

Aquí se muestra el modelo CAR con la matriz de pesos correspondiente a las vecindades de esferas de influencia. Los coeficientes de las variables explicativas son muy similares al modelo anterior. El factor de efecto espacial es un poco más significativo, $\lambda = 0{,}074628$, p-valor $p = 0{,}022659$, por lo que se asegura un nivel de 2,5 %. El valor de AIC es $-869{,}17$.

```
summary(slm.SAR.KNN)

>
> Call: spatialreg::spautolm(formula = formula, data = data, listw = listw,
>     na.action = na.action, family = family, method = method,
>     verbose = verbose, trs = trs, interval = interval, zero.policy = zero.policy,
```

```
>     tol.solve = tol.solve, llprof = llprof, control = control)
>
> Residuals:
>        Min         1Q     Median         3Q        Max
> -0.2204392 -0.0380011  0.0034088  0.0414585  0.1766092
>
> Coefficients:
>               Estimate  Std. Error  z value  Pr(>|z|)
> (Intercept)  7.3548e+00  2.8101e-02 261.7229 < 2.2e-16
> PM10        -1.7628e-03  5.1206e-04  -3.4425 0.0005763
> GM           6.9912e-05  3.5666e-05   1.9602 0.0499708
> log(T)       3.3461e-02  3.0935e-03  10.8167 < 2.2e-16
>
> Lambda: 0.030997 LR test value: 3.1187 p-value: 0.077397
> Numerical Hessian standard error of lambda: 0.017241
>
> Log likelihood: 439.5487
> ML residual variance (sigma squared): 0.0035133, (sigma: 0.059273)
> Number of observations: 313
> Number of parameters estimated: 6
> AIC: -867.1
```

Este es el caso del modelo SAR usando la matriz de pesos obtenida por *KNN* con $k = 4$. En este caso, los coeficientes de las variables explicativas son algo más significativos, llegando GM a ser significativo al 5 %. El efecto espacial decrece, con $\lambda = 0{,}030997$ y p-valor $p = 0{,}077397$, por lo que no llega a ser significativo al 5 %. Se obtiene AIC $= -867{,}1$.

```
summary(slm.CAR.KNN)

>
> Call: spatialreg::spautolm(formula = formula, data = data, listw = listw,
>     na.action = na.action, family = family, method = method,
>     verbose = verbose, trs = trs, interval = interval, zero.policy = zero.policy,
>     tol.solve = tol.solve, llprof = llprof, control = control)
>
> Residuals:
>        Min         1Q     Median         3Q        Max
> -0.2209991 -0.0372576  0.0028621  0.0417726  0.1706017
>
> Coefficients:
>               Estimate  Std. Error  z value  Pr(>|z|)
> (Intercept)  7.3562e+00  2.8141e-02 261.4094 < 2.2e-16
> PM10        -1.7635e-03  5.1175e-04  -3.4460 0.0005689
> GM           7.0296e-05  3.5652e-05   1.9717 0.0486389
```

```
> log(T)       3.3260e-02  3.0996e-03  10.7306 < 2.2e-16
>
> Lambda: 0.064981 LR test value: 3.3986 p-value: 0.065251
> Numerical Hessian standard error of lambda: 0.032893
>
> Log likelihood: 439.6886
> ML residual variance (sigma squared): 0.0034884, (sigma: 0.059063)
> Number of observations: 313
> Number of parameters estimated: 6
> AIC: -867.38
```

Finalmente se presenta el modelo CAR con la matriz de pesos anterior. Es un modelo muy similar al anterior en lo que se refiere a significación. Al igual que en el otro CAR, el modelo considera que todas las variables (incluyendo la estructura espacial) son algo más significativas. A nivel espacial $\lambda = 0{,}064981$ con p-valor $p = 0{,}065251$. Se obtiene AIC $= -867{,}38$

4.3. Conclusiones

Los cuatro modelos propuestos son muy similares en rendimiento pero, teniendo en cuenta los resultados de test de Moran y los AIC, parece aceptable asumir que la relación espacial de esferas de influencia es una forma correcta de modelar la estructura espacial de los datos.

Observando 4.4 se puede ver como, aunque de forma modesta, los modelos CAR reducen más los los residuos y su autocorrelación. Comparando los valores de AIC se obtiene que el mejor modelo es el CAR con estructura de *SOI*. Sin embargo se debe destacar que la mejora con respecto del modelo lineal que asume independencia no es dramática, pues se puede calcular que para el modelo simple se tiene un AIC de -865, no muy superior al del mejor modelo. Esto es debido, probablemente, a la gran significación del logaritmo de la población que domina al resto de efectos. En conclusión, se pueden interpretar a partir de los estimadores que el efecto del gas radón y los núcleos de población (probablemente por ser áreas más urbanas e industrializadas), así como la proximidad a zonas de riesgo, tienen un efecto en la mortalidad por cancer de plumón. El efecto de los contaminantes PM10 también parece, según los modelos, considerable. Sin embargo, el hecho de que su estimador siempre tome valores negativos lleva a pensar que no son datos completamente válidos. La posible causa de esto es que solo existían 9 sensores en toda Galicia en el momento en el que se tomaron los datos, y a cada municipio se le asignó la medición del sensor más próximo. Debido a esto, es probable que los datos no representen la calidad real del aire y

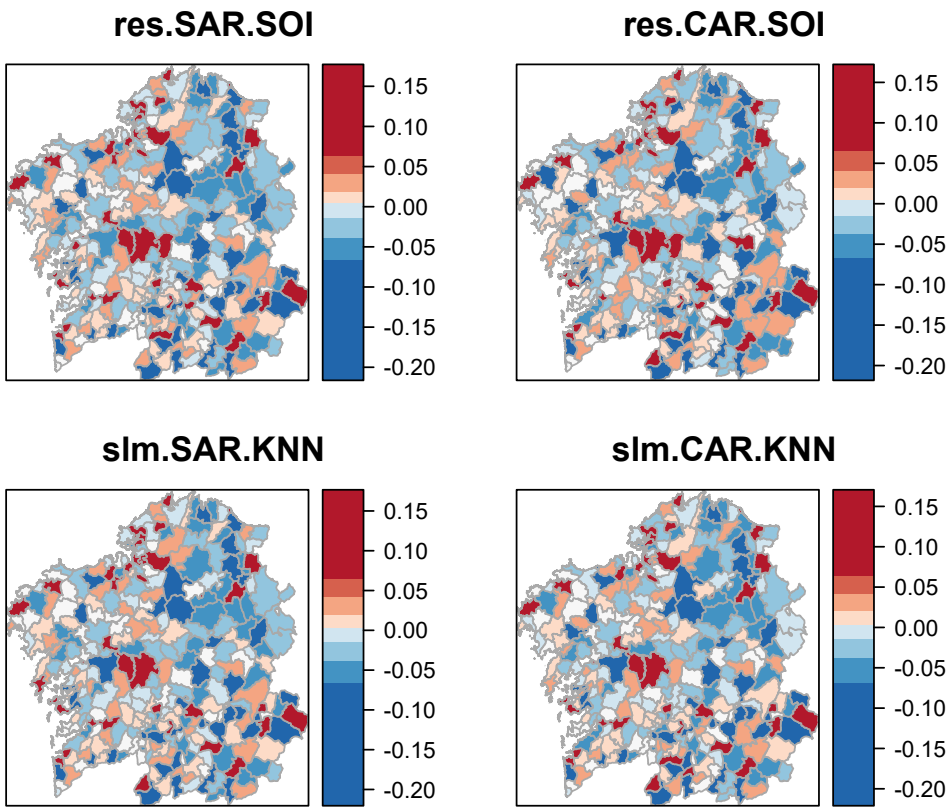


Figura 4.4: Residuos obtenidos por los modelos espaciales

se deberá tener en cuenta.

Tablas de datos

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
1	Abegondo	0.95	1.08	0.97	28.6	12.7	15.9	5585	50.29	36
2	Ames	1.13	1.02	1.12	10.5	4.4	6.1	29975	154.56	22
3	Aranga	1.01	0.99	1	38.3	17	21.3	2033	120.89	36
4	Ares	1.4	1.01	1.36	24.4	10.7	13.7	5741	128.7	10
5	Arteixo	1.1	1	1.09	14.1	6.3	7.8	30857	68.87	36
6	Arzúa	1.12	0.95	1.08	29.3	12.4	16.9	6261	81.02	22
7	Baña A	1.18	0.96	1.13	35.9	15.3	20.6	3754	209.79	22
8	Bergondo	0.99	1.03	1	25.9	11.6	14.3	6702	141.38	36
9	Betanzos	0.99	1.06	1.01	22.8	9.3	13.5	13352	49.83	36
10	Boimorto	1.03	0.95	0.99	36	15.1	20.9	2115	84.66	22
11	Boiro	1.37	0.94	1.28	19.8	8.4	11.4	19060	134.51	10
12	Boqueixón	1.08	1.1	1.1	23.3	10.2	13.1	4342	77.2	22
13	Brión	1.14	1.02	1.12	19.7	8.4	11.3	7519	201.52	22
14	Cabana de Bergantiños	1.17	1.01	1.14	30.5	12.8	17.7	4623	136.29	36
15	Cabanas	1.14	0.95	1.1	26.5	11	15.6	3294	165.82	10
16	Camariñas	1.2	0.79	1.09	26.8	11.1	15.7	5774	104.47	22
17	Cambre	1.05	0.97	1.04	15	6.6	8.4	24029	70.52	36
18	Capela A	1.14	1	1.12	30.2	13.6	16.6	1356	44.39	10
19	Carballo	1.01	0.89	0.98	21.4	9.3	12.1	31288	116.76	36
20	Cariño	1.37	0.95	1.29	35.1	14.6	20.5	4376	159.32	22
21	Carnota	0.97	0.89	0.94	23.1	10.1	13	6118	76.64	36
22	Carral	1.16	1.01	1.12	27.9	12.4	15.6	7147	101.98	10
23	Cedeira	1.04	0.8	0.97	23.7	10.2	13.5	7760	40.59	22

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
24	Cee	0.97	0.9	0.94	29.8	13.1	16.7	5156	88.68	36
25	Cerceda	1.04	0.93	1	37.9	15.8	22.1	1249	75.67	10
26	Cerdido	0.96	0.98	0.96	30.7	13.6	17.1	1747	68.19	36
27	Coirós	0.93	0.99	0.95	23.4	9.6	13.8	1672	22.91	22
28	Corcubión	1.09	0.85	1.03	30.9	12.8	18	6706	43.85	22
29	Coristanco	1.24	1.38	1.27	23	9.2	13.8	244810	76.94	36
30	Coruña A	1.01	1.2	1.05	15.6	6.9	8.7	29434	83.98	36
31	Culleredo	0.93	0.98	0.93	26.9	11.5	15.4	4048	41.25	36
32	Curtis	1.22	0.91	1.13	27.4	11.2	16.1	2913	163.02	22
33	Dodro	1.17	0.85	1.09	30.1	13.1	17	3137	40.87	22
34	Dumbría	1.15	1.15	1.16	26	11.1	14.9	13498	126.85	10
35	Fene	1.41	1.33	1.41	25.7	10.2	15.5	70389	65.02	10
36	Ferrol	0.98	0.81	0.9	23.5	10.8	12.8	4824	34.3	22
37	Fisterra	1.09	0.98	1.06	29.9	13	16.9	2485	59.29	22
38	Frades	1.07	0.94	1.04	37.3	16.7	20.6	1421	131.85	36
39	Irixoa	0.98	0.95	0.97	26.9	11.7	15.2	3207	173.45	22
40	Laracha A	0.96	0.92	0.94	24.5	10.6	13.9	11443	41.1	36
41	Laxe	1.2	0.88	1.12	30.3	12.8	17.5	3513	79.88	22
42	Lousame	1.17	0.87	1.1	31.2	13.2	18	5768	278.98	36
43	Malpica de Bergantiños	1.1	1.03	1.07	38.7	15.5	23.2	1455	80.77	10
44	Mañón	1.25	0.83	1.16	32.8	13.1	19.7	4289	74.91	22
45	Mazaricos	1.05	1.02	1.03	26	11.4	14.6	7578	78	22
46	Melide	1.02	0.97	1	32.6	14.6	18	2795	72.39	22

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
47	Mesía	1.08	0.96	1.05	23.8	10.9	12.9	5838	88.03	10
48	Miño	1.14	0.97	1.1	35.6	14.6	21	1303	37.83	10
49	Moeche	1.11	0.99	1.08	35	15.7	19.2	2089	90.19	10
50	Monfero	1.43	1.3	1.48	27.2	11.4	15.8	5417	127.52	10
51	Mugardos	1.14	0.82	1.05	29.6	12.9	16.7	5068	55.64	22
52	Muros	1.44	0.93	1.35	28.5	11.8	16.7	9117	191.88	22
53	Muxía	1.26	0.95	1.21	19.2	8.2	10.9	39574	41.96	10
54	Narón	1.18	1.07	1.17	29	12.3	16.8	5327	150.82	10
55	Neda	1.23	0.93	1.17	24.8	10.6	14.2	7009	136.94	22
56	Negreira	1.17	1.03	1.14	23.7	10.1	13.6	14571	118.38	22
57	Noia	1.06	1.27	1.11	18.1	8	10.1	34563	82.13	36
58	Oleiros	0.93	0.91	0.91	22	9.4	12.6	12844	68.83	22
59	Ordes	0.99	0.9	0.95	13.9	6.2	7.8	7400	71.38	22
60	Oroso	1.09	0.88	1.03	36.8	15.7	21.1	5997	183.3	10
61	Ortigueira	1.13	1	1.1	33.1	13.4	19.7	6804	143.21	22
62	Outes	0.99	0.98	0.98	33.6	14.9	18.6	2502	217.23	36
63	Oza-Cesuras	1.17	1.11	1.17	22.8	9.4	13.5	8693	167.98	22
64	Paderne	1.03	0.94	1.01	25.5	11.3	14.2	4680	123.93	22
65	Padrón	1.47	0.92	1.36	22	9.2	12.7	9672	99.27	10
66	Pino O	1.24	0.94	1.18	30.4	13.2	17.2	5893	206.68	36
67	Pobra do Caramiñal A	1.36	1.06	1.32	23.5	9.7	13.8	8117	75.99	10
68	Ponteceso	1.06	0.88	1.02	21.1	9.3	11.7	10634	24.75	10
69	Pontedeume	1.44	1.07	1.39	24.5	10.6	13.9	9571	111.95	10

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
70	Pontes de García Rodríguez As	1.28	0.81	1.17	24.1	10.4	13.7	11479	79.85	10
71	Porto do Son	1.51	1.3	1.51	19.1	8.1	11	27565	91.03	10
72	Rianxo	1.28	0.96	1.22	27	10.8	16.2	4767	257.64	22
73	Ribeira	0.91	0.97	0.91	22.1	10	12.1	15156	73.73	36
74	Rois	1.12	1.05	1.11	30.8	14.1	16.6	3019	107.8	10
75	Sada	1.17	0.83	1.09	28.2	11.9	16.3	9719	115.1	22
76	San Sadurniño	1.18	1.23	1.19	19.9	8	11.9	95800	104.9	22
77	Santa Comba	1	0.93	0.97	37.4	17.3	20.1	1742	53.53	22
78	Santiago de Compostela	0.97	0.91	0.94	38.6	17	21.6	1965	39.56	22
79	Santiso	1.09	0.94	1.04	28.5	12.1	16.4	1211	45.03	10
80	Sobrado	1.07	1.25	1.11	16.5	7.1	9.4	18254	181.7	22
81	Somozas As	0.98	0.94	0.95	32.4	15.5	16.8	1248	36.66	22
82	Teo	1.07	0.96	1.03	36.4	14.8	21.6	3706	77.74	22
83	Toques	1.08	1.11	1.1	31.6	14	17.7	3853	111.08	22
84	Tordoia	1.05	0.94	1.01	30.9	13.5	17.4	3301	106.43	22
85	Touro	1.14	1	1.11	28.6	13.1	15.5	6796	154.44	10
86	Trazo	1.17	0.84	1.09	32.3	13.9	18.4	4120	89.64	22
87	Val do Dubra	1.12	0.96	1.1	23.6	9.8	13.8	5073	118.97	22
88	Valdoviño	0.95	1.02	0.94	37.8	16.9	20.9	1316	110.91	22
89	Vedra	1.13	0.99	1.1	31.7	13.8	18	1236	51.63	10
90	Vilarmaior	1.02	0.97	0.99	30	12.7	17.3	7710	96.52	22
91	Vilasantar	1.08	0.85	1.02	31.1	12.9	18.3	4902	42.56	22
92	Vimianzo	1.12	0.86	1.04	31.7	14	17.8	4241	36.57	10

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
93	Zas	0.98	1.02	0.98	33.7	14.8	18.9	5275	35.02	36
94	Abadín	0.94	0.99	0.94	42.7	18.2	24.5	2646	97.53	22
95	Alfoz	0.98	0.91	0.95	36.8	16.1	20.7	1903	59.05	22
96	Antas de Ulla	0.8	0.95	0.8	40.7	18	22.7	2152	107.66	22
97	Baleira	0.79	1.04	0.83	39.3	18.9	20.4	1392	80.39	22
98	Baralla	1	0.95	0.99	33.2	13.8	19.4	2994	90.08	22
99	Barreiros	0.83	0.9	0.83	35.1	16.7	18.5	3012	95.62	22
100	Becerreá	0.94	0.95	0.93	35.7	15.6	20.2	3179	200	22
101	Begonte	0.92	1.02	0.94	38.9	17.4	21.5	1571	28.11	22
102	Bóveda	0.84	1.07	0.87	40.1	17.8	22.2	2401	129.38	18
103	Burela	0.92	1.28	0.99	31.6	13.1	18.5	5226	161.66	22
104	Carballedo	0.88	1.02	0.91	35.3	16.1	19.2	2823	78.21	22
105	Castro de Rei	0.81	1.02	0.83	40.7	22.4	18.3	1528	216.15	22
106	Castroverde	1.22	0.93	1.16	23.7	10.6	13.1	4369	62.16	10
107	Cervantes	0.92	0.98	0.93	34.7	15	19.6	3695	70.91	22
108	Cervo	0.91	1	0.93	34.6	14.5	20.1	4892	82.01	22
109	Chantada	0.79	0.91	0.79	30.8	13	17.8	8553	65.28	18
110	Corgo O	0.81	0.95	0.82	42.1	20.2	22	1106	137.15	22
111	Cospeito	0.75	1.2	0.81	39.9	18.9	21	3960	93.96	22
112	Folgoso do Courel	1.09	1	1.07	23.4	10.1	13.4	9899	57.79	22
113	Fonsagrada A	0.99	0.92	0.97	36.5	16.1	20.5	4004	114.8	22
114	Foz	1.01	1.02	1.01	34.1	14.4	19.6	2036	120	10
115	Friol	0.92	1.02	0.92	31	13.7	17.3	5564	35.46	22

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
116	Guitiriz	0.91	1	0.91	36.4	16.5	19.9	2927	52.03	22
117	Guntín	0.79	1.03	0.82	45.6	20	25.5	1797	71.6	22
118	Incio O	1.17	1.04	1.16	26.2	11.2	15	3456	136.51	10
119	Láncara	0.86	1.01	0.88	34.6	15.8	18.9	2789	99.93	22
120	Lourenzá	0.95	0.98	0.95	32	14	18	2338	70.13	22
121	Lugo	1.16	1.1	1.16	20.3	8.2	12.1	98560	57.36	22
122	Meira	0.88	1.08	0.92	29.4	12.9	16.5	1731	46.13	22
123	Mondoñedo	0.94	0.9	0.92	34.9	13.9	20.9	3991	46.05	22
124	Monforte de Lemos	0.97	1.04	0.98	28.5	12	16.6	19201	166.32	18
125	Monterroso	0.92	1.06	0.93	31.8	14.1	17.7	3779	91.57	22
126	Muras	1.05	0.95	1.03	41.2	20.3	21	696	72.1	10
127	Navia de Suarna	0.8	0.95	0.81	47.3	24.4	22.9	1254	93.71	22
128	Negueira de Muñiz	0.77	1.03	0.83	30.3	13.7	16.6	211	17.09	22
129	Nogais As	0.8	1.03	0.82	37.9	19.3	18.6	1228	240.24	22
130	Ourol	1.15	0.92	1.09	47.9	20.6	27.3	1090	155.11	10
131	Outeiro de Rei	0.95	0.94	0.95	25.9	11.4	14.4	5083	103.73	22
132	Palas de Rei	0.9	0.89	0.87	38.4	16.2	22.2	3601	46.36	22
133	Pantón	0.83	0.99	0.85	46.6	20.6	26	2708	157.64	18
134	Paradela	0.76	1.01	0.79	37.2	17.2	20	1967	115.65	22
135	Páramo O	0.89	1.03	0.91	36.5	16.8	19.7	1484	128.61	22
136	Pastoriza A	0.9	0.93	0.91	33.7	15.4	18.3	3309	27.87	22
137	Pedrafita do Cebreiro	0.86	0.95	0.86	40.2	20	20.2	1155	125.83	22
138	Pobra do Brollón A	0.86	1.02	0.89	36.1	15.2	20.9	1738	51.51	22

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
139	Pol	0.84	1.19	0.89	46.4	21	25.4	1830	162.44	18
140	Pontenova A	0.86	0.95	0.88	35.5	15.4	20.1	2534	117.89	22
141	Portomarín	0.79	1.02	0.81	38.2	17.6	20.6	1624	26.5	22
142	Quiroga	0.89	0.92	0.89	38.1	16.7	21.4	3501	81.57	15
143	Rábade	1.14	1.04	1.13	23.8	9.7	14	10010	80.22	22
144	Ribadeo	0.88	0.95	0.88	41.2	18.1	23.1	1011	367.81	15
145	Ribas de Sil	0.83	0.99	0.86	46.3	20.2	26.1	583	22.98	22
146	Ribeira de Piquín	0.92	1.15	0.96	41.6	19.2	22.4	1395	43.48	22
147	Riotorto	0.83	1.02	0.86	40.4	19.3	21.1	1458	292.03	22
148	Samos	0.98	1.04	0.99	27.1	11.1	16	1604	500.35	22
149	Sarria	0.86	1.01	0.88	26	11.6	14.4	13504	40.1	22
150	Saviñao O	0.79	1.01	0.81	41.5	18.6	22.9	4113	311.14	18
151	Sober	0.83	0.93	0.84	45.8	20.4	25.4	2453	271.03	18
152	Taboada	0.78	0.91	0.78	40.5	17.4	23.2	3101	51.43	22
153	Trabada	1.01	0.99	1.01	35.8	14.6	21.2	1203	257.39	22
154	Triacastela	0.8	0.97	0.82	33.1	15.4	17.8	721	100.8	22
155	Valadouro O	1.15	1.06	1.13	35.6	15.8	19.8	2069	44.94	22
156	Vicedo O	1.06	0.91	1	33.7	15.3	18.4	1882	173.44	10
157	Vilalba	0.88	1.08	0.91	28.4	11.9	16.4	14788	58.89	22
158	Viveiro	1.35	0.98	1.29	23.7	9.7	14	15932	82.04	10
159	Xermade	0.8	1.1	0.85	37.6	17.4	20.2	2774	116.15	22
160	Xove	1.22	0.9	1.14	15.4	6.5	8.9	9660	46	22
161	Allariz	0.97	1.07	0.99	25.5	11.1	14.4	6060	132.02	18

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
162	Amoeiro	0.93	0.96	0.94	33.7	14.7	19	2274	223.49	18
163	Arnoia A	1.09	1.11	1.11	39.6	17	22.6	1040	75.13	18
164	Avión	0.98	1.05	1	47.5	20.3	27.2	2053	254.19	10
165	Baltar	0.82	0.95	0.81	44	20.3	23.7	1019	150.24	15
166	Bande	1.02	0.93	0.99	45.4	20	25.4	1829	117.38	18
167	Baños de Molgas	1.14	1.02	1.13	44.1	20.3	23.8	1675	229.02	18
168	Barbadás	1.05	1.04	1.05	16.5	7.1	9.4	10371	184.25	18
169	Barco de Valdeorras O	0.84	1.05	0.86	19.4	8.9	10.5	13899	64.67	15
170	Beade	0.93	1.06	0.96	33.2	14.2	19	464	137.48	18
171	Beariz	1.02	1.04	1.02	43.8	15.4	28.4	1112	273.02	10
172	Blancos Os	0.85	0.96	0.84	42.6	20.9	21.7	885	204.11	15
173	Boborás	0.98	1.04	0.99	43.6	17.3	26.3	2617	198.03	18
174	Bola A	1.02	0.97	1	42.8	18.1	24.7	1350	89.31	18
175	Bolo O	0.85	0.96	0.85	49.5	22.5	26.9	1013	450.4	15
176	Calvos de Randín	0.88	0.99	0.87	51.2	22.4	28.7	985	219.2	15
177	Carballeda de Avia	0.83	0.9	0.8	30.6	13.8	16.8	1697	149.77	15
178	Carballeda de Valdeorras	0.97	1.11	1.01	41.1	18	23.1	1390	602.45	18
179	Carballiño O	0.93	1.09	0.96	26.4	11	15.4	14123	152.98	18
180	Cartelle	1.08	1.02	1.07	45	19.4	25.6	2944	58.97	18
181	Castrelo de Miño	0.74	0.93	0.75	41.4	20	21.4	1109	127.41	15
182	Castrelo do Val	0.99	1.07	1.01	42	18.4	23.6	1620	138.69	18
183	Castro Caldelas	0.86	0.97	0.87	39	17.6	21.4	1376	103.69	15
184	Celanova	1.12	1.01	1.1	34.4	14.3	20.1	5615	88.58	18

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
185	Cenlle	0.94	1.04	0.96	41.6	16.1	25.5	1235	262.85	18
186	Chandrexa de Queixa	0.93	1.14	0.98	31.5	13.6	18	3150	76.06	18
187	Coles	1.03	1.16	1.07	37	15.2	21.8	1191	241.65	18
188	Cortegada	0.82	1.03	0.83	42.6	18.6	24	1808	220.13	15
189	Cualedro	0.82	0.96	0.82	43.5	20.9	22.6	545	169.66	15
190	Entrimo	1.18	1	1.15	40.4	15	25.3	1323	432.42	26
191	Esgos	0.99	1.07	1.02	36.9	16.5	20.4	1216	364.69	18
192	Gomesende	0.91	0.99	0.92	25.3	11.2	14.1	10200	131.33	15
193	Gudiña A	1.09	1.06	1.09	49.8	19.1	30.7	833	169.17	18
194	Irixo O	0.79	0.96	0.79	32.4	15.5	16.9	1461	235.52	15
195	Larouco	0.95	1.12	0.98	49.5	20.6	28.8	1637	51.08	18
196	Laza	0.99	1.15	1.02	43.1	19.5	23.6	1546	89.46	15
197	Leiro	0.94	1.03	0.95	33.1	14.5	18.7	863	76.13	18
198	Lobeira	0.88	1.04	0.9	48.5	20.6	27.9	544	115.87	15
199	Lobios	0.79	0.92	0.8	45.1	21.4	23.7	1449	89.33	15
200	Maceda	0.98	1.08	1	35.4	14.9	20.5	1675	100.71	18
201	Manzaneda	1.05	1.06	1.05	50.5	19.7	30.8	837	265.12	18
202	Maside	1.02	1.1	1.04	37.5	16.8	20.7	2025	210.19	26
203	Melón	1.12	0.95	1.1	34.2	14.8	19.4	2978	245.64	15
204	Merca A	0.83	1.02	0.84	41.1	19.3	21.8	973	217.9	15
205	Mezquita A	0.98	1.09	1	35.3	14.7	20.7	2906	126.75	18
206	Montederramo	0.94	1.2	1.01	41.3	16	25.2	1340	245.89	26
207	Monterrei	1.01	1.14	1.02	40.7	17	23.7	2078	65.28	18

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
208	Muíños	0.72	0.91	0.7	38.4	17.9	20.4	1155	428.92	15
209	Nogueira de Ramuín	0.92	1	0.92	45	19	26	822	184.06	15
210	Oímbra	0.68	1.05	0.71	41.5	18.9	22.7	2809	229.44	15
211	Ourense	0.93	0.98	0.92	41.8	19.1	22.7	1622	171.04	15
212	Paderne de Allariz	0.85	1.17	0.9	36.7	15.8	20.9	2198	110.18	18
213	Padrenda	0.63	0.9	0.63	31.4	14.7	16.7	2020	477.38	15
214	Parada de Sil	1.11	1.43	1.17	23.7	9.7	14	106905	154.91	18
215	Pereiro de Aguiar	1.01	1.08	1.03	42	18.1	23.9	1515	269.49	18
216	Peroxa A	1.12	1.1	1.13	40.2	15.5	24.8	1951	442.07	26
217	Petín	0.88	1.02	0.9	50.5	22.1	28.4	634	107.66	18
218	Piñor	0.93	1.1	0.97	24.1	10.6	13.5	6350	212.33	18
219	Pobra de Trives A	0.88	1.06	0.92	39.6	17.1	22.5	1999	198.54	18
220	Pontedeva	0.93	0.94	0.92	29.2	11.9	17.4	961	255.53	15
221	Porqueira	0.82	1.11	0.86	39.3	18.3	21	1237	204.31	18
222	Punxín	0.94	1	0.93	47.8	20.6	27.2	931	130.44	15
223	Quintela de Leirado	0.83	0.97	0.83	33.8	15.6	18.2	2280	265.4	15
224	Rairiz de Veiga	1.04	1.1	1.06	36.3	16.2	20.2	625	169.1	18
225	Ramirás	0.96	1.02	0.98	40.5	15.4	25.1	749	191.84	18
226	Ribadavia	1.03	1.02	1.02	48.7	17.5	31.2	651	201.05	18
227	Riós	0.98	1.03	0.98	40.9	17.9	23.1	1456	219.42	18
228	Rúa A	1.1	0.99	1.07	48.5	19.9	28.6	1643	133.06	18
229	Rubiá	0.98	1.21	1.03	27.6	11.7	15.9	5187	116.79	18
230	San Amaro	0.83	0.99	0.83	52.7	21.7	31	632	98.35	15

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
231	San Cibrao das Viñas	0.85	1.16	0.91	42.8	19.7	23.2	1684	162.52	15
232	San Cristovo de Cea	0.99	1.07	1.01	28.5	12.6	15.9	4634	211.36	15
233	San Xoán de Río	0.77	0.9	0.75	39	17.1	21.9	1494	238.19	15
234	Sandiás	0.93	1.09	0.96	41.4	18.2	23.1	1180	66.51	18
235	Sarreaus	0.96	0.96	0.96	21.5	9.7	11.8	4972	126.19	18
236	Taboadela	0.84	1	0.86	41.5	18.8	22.7	2418	154	18
237	Teixeira A	1.03	0.95	1.02	41.2	16.8	24.4	1297	165.24	15
238	Toén	0.85	1.01	0.86	49.4	22.2	27.3	1313	76.84	15
239	Trasmiras	1.08	1.02	1.07	32.2	14.9	17.3	1561	147.67	18
240	Veiga A	0.87	1.03	0.88	53.7	24.8	28.9	391	294.24	15
241	Verea	0.94	1.05	0.96	32.8	14.6	18.2	2511	164.32	18
242	Verín	0.81	0.93	0.8	42.7	19.5	23.1	1456	142.77	15
243	Viana do Bolo	0.86	0.95	0.85	47.4	19.9	27.5	965	100.69	15
244	Vilamarín	1.03	1.14	1.05	49.7	20.7	29	1059	132.07	18
245	Vilamartín de Valdeorras	0.7	0.92	0.72	23	10.5	12.5	14652	163.78	15
246	Vilar de Barrio	0.79	0.85	0.77	39.5	17.3	22.2	3113	111.58	15
247	Vilar de Santos	0.86	1	0.88	34.7	16.1	18.6	2056	128.08	18
248	Vilardevós	0.99	0.93	0.97	37.2	15.7	21.5	1793	26.46	15
249	Vilariño de Conso	0.95	1.05	0.96	46.2	20.5	25.7	1500	138.71	15
250	Xinzo de Limia	0.99	1.07	1	37.3	15.3	22	922	117.46	15
251	Xunqueira de Ambía	0.83	0.92	0.83	45.9	21.7	24.2	2024	106.2	15
252	Xunqueira de Espadanedo	0.78	0.95	0.78	34.1	17.9	16.3	633	198.79	15
253	Agolada	0.95	1.26	1.04	33.2	14.2	19	2904	166.23	26

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
254	Arbo	1.01	1.13	1.03	22.5	8.9	13.5	3705	233.86	10
255	Baiona	1.21	1.15	1.2	17.5	7.8	9.7	12233	163.14	15
256	Barro	1.29	1.15	1.29	20.1	8.5	11.5	12352	232.56	15
257	Bueu	1.15	0.96	1.11	20.9	8.3	12.6	9895	71.77	10
258	Caldas de Reis	1.31	0.97	1.25	20.1	8.5	11.7	13399	226.08	10
259	Cambados	1.1	1.07	1.11	32.1	12.6	19.5	1962	264.2	10
260	Campo Lameiro	1.08	0.98	1.05	18.2	7.7	10.5	26567	199.73	15
261	Cangas	0.94	1.18	1.01	29	12	17	5342	168.3	26
262	Cañiza A	1.24	1.08	1.22	22.3	9.5	12.8	3402	101.54	10
263	Catoira	0.96	1.16	1.01	38.5	15.1	23.4	2719	362.51	26
264	CERDEDO-COTOBADE	1	1.07	1.04	36.8	14.8	22	2261	115.61	26
265	Covelo	1.12	0.98	1.08	29.3	11.2	18.1	4931	95.74	10
266	Crecente	0.84	0.98	0.85	37.6	17.1	20.4	1209	74.29	18
267	Cuntis	1.09	1.28	1.12	26.6	10.9	15.7	21197	94.67	22
268	Dozón	1.05	0.96	1.04	40.2	16.3	23.9	3696	100.31	10
269	Estrada A	1.09	1.17	1.13	31.6	12.6	18.9	1784	455.67	26
270	Forcarei	0.96	0.97	0.95	37.6	16.1	21.5	2651	101.73	22
271	Fornelos de Montes	1.16	1.14	1.17	16	7.1	8.9	14051	233.4	15
272	Gondomar	1.11	1.15	1.14	19.5	8.3	11.2	10851	206.51	10
273	Grove O	1.27	1.17	1.26	20.3	8.5	11.8	10314	163.01	15
274	Guarda A	0.92	1.09	0.94	24.4	10.6	13.8	20158	82.29	18
275	Illa de Arousa A	1.08	1	1.08	32.7	13.7	19	2763	591.14	10
276	Lalín	1.04	1.54	1.14	18.8	8.2	10.6	25329	186.1	10

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
277	Lama A	0.92	1.12	0.95	21.6	9.2	12.4	5428	201.86	10
278	Marín	1.11	0.99	1.1	24	9.4	14.7	5033	254.55	10
279	Meaño	1.1	1.01	1.08	18.6	7.9	10.7	19365	182.56	15
280	Meis	1.15	1.18	1.18	29.5	11.5	18	4742	510.93	26
281	Moaña	1.04	1.13	1.06	23.5	9.4	14.1	682	196.89	26
282	Mondariz	1.06	0.94	1.03	23.9	9.8	14.1	4362	171.85	10
283	Moraña	1.22	1.19	1.23	17.6	7.3	10.4	15324	256.22	15
284	Mos	1.09	1.44	1.21	28	11.2	16.7	4121	307.83	26
285	Neves As	1.13	1.19	1.15	19	8.4	10.6	17715	178.73	15
286	Nigrán	1.3	1.07	1.24	22.9	10.4	12.6	3042	117.61	15
287	Oia	1.17	1.2	1.2	23.8	9.3	14.4	3075	139.83	26
288	Pazos de Borbén	1.12	1.37	1.17	18.5	7.5	11	82946	160.92	10
289	Poio	1.24	1.34	1.29	15.4	6.5	8.9	18508	165.45	15
290	Ponte Caldelas	1.1	1	1.07	26	10.5	15.5	3025	210.97	10
291	Ponteareas	0.91	1.16	0.96	17.1	7.7	9.4	16794	200.41	10
292	Pontecesures	1.12	1.09	1.13	16.4	6.8	9.6	23115	218.76	26
293	Pontevedra	1.11	1.3	1.17	24.5	9.8	14.7	5625	71.59	10
294	Porriño O	1.09	1.05	1.09	16.4	7.1	9.3	3092	161.11	22
295	Portas	1.31	1.08	1.28	19.2	8	11.2	29909	185.11	15
296	Redondela	1.15	1.13	1.16	19.8	7.5	12.2	5103	400.28	10
297	Ribadumia	0.78	1.02	0.8	36.9	17.8	19.1	2752	110.19	18
298	Rodeiro	1.24	1.05	1.19	21.7	9.4	12.3	6455	188.79	15
299	Rosal O	1.11	1.38	1.2	13.6	5.9	7.8	8890	257.92	26

ID	Concello	RMES.H	RMES.M	RMES.T	Tpc64	Hpc64	Mpc64	T	GM	PM10
300	Salceda de Caselas	1.13	1.28	1.19	21	8.5	12.5	9626	270.58	26
301	Salvaterra de Miño	0.8	1.06	0.84	20.7	9.3	11.4	17543	155.66	10
302	Sanxenxo	1.05	1.15	1.09	25.9	11	14.9	8966	83.58	22
303	Silleda	1.16	1.2	1.19	17.1	7.1	10	7356	292.62	10
304	Soutomaior	1.19	1.1	1.18	19.5	8.8	10.7	13707	191.16	15
305	Tomiño	1.16	1.31	1.21	20	8.2	11.8	16884	160.95	15
306	Tui	1.22	0.95	1.17	20.8	8.5	12.3	6050	144.47	22
307	Valga	1.31	1.61	1.37	20	8.1	11.9	294997	128.13	15
308	Vigo	0.97	0.91	0.94	24	9.5	14.5	5972	183.95	10
309	Vila de Cruces	1.06	1.14	1.08	32.7	13.8	19	5676	68.34	22
310	Vilaboa	1.43	1.32	1.43	18.6	7.7	10.9	37712	189.2	10
311	Vilagarcía de Arousa	1.27	1.25	1.28	22.7	9.4	13.3	10459	250.16	10
312	Vilanova de Arousa	1.5	0.96	1.37	19.3	8.2	11.1	5006	197.77	10
313	Mondariz-Balneario	1.06	1.09	1.08	32.7	12.6	20.1	6187	212.42	10

Código completo en R

```
###### librerias utilizadas######
library(rgdal)
library(rgeos)
library(maptools)
library(gpclib)  # may be needed, may not be
library(ggplot2)
library(dplyr)
library(tidyr)
library(stringr)
library(spdep)
library(gridExtra)
library(RANN)



makefloat = function(x) as.numeric(as.character(x))
d = function(X1, X2, A1, A2) {
    d = sqrt((X1 - A1)^2 + (X2 - A2)^2)
    return(d)
}

###### Obtención de los datos gergráficos y la tabla de datos######


mapa.concellos <- readOGR(dsn = "./Concellos", layer = "Concellos_IGN")
df.mapa.concellos <- fortify(mapa.concellos, region = "Concello")

centros = data.frame(mapa.concellos$Concello, coordinates(mapa.concellos))
names(centros) = c("id", "X1", "X2")
centros$id = str_replace_all(toupper(centros$id), c(`AS ` = "", `, AS` = "",
    `OS ` = "", `, OS` = ""))
centros$id = str_replace_all(toupper(centros$id), c(`A ` = "", `, A` = "", `O ` = "",
    `, O` = ""))
sensores = read.csv2("snombres.csv")
```

```
colnames(sensores)[1] = "id"
sensores$id = str_replace_all(toupper(sensores$id), c(`AS ` = "", `, AS` = "",
    `OS ` = "", `, OS` = ""))
sensores$id = str_replace_all(toupper(sensores$id), c(`A ` = "", `, A` = "",
    `O ` = "", `, O` = ""))
sensores = cbind(sensores, centros[centros$id %in% sensores$id, ]$X1, centros[centros$id %in%
    sensores$id, ]$X2)
colnames(sensores)[c(3, 4)] = c("A1", "A2")
values = numeric()
for (i in 1:dim(centros)[1]) {
    dist = d(centros$X1[i], centros$X2[i], sensores$A1, sensores$A2)
    mindist = min(dist)
    minid = which(dist == mindist)
    values = rbind(values, c(minid, mindist))
}
atmconcellos = data.frame(centros$id, sensores$PM10[values[, 1]], sensores$id[values[,
    1]], values[, 2])
names(atmconcellos) = c("id", "PM10", "PROX", "dist")
atmconcellos$id = as.character(atmconcellos$id)


pconcellos = read.csv2("pulmon_concellos.csv")
colnames(pconcellos)[2] = "id"
pconcellos$id = as.character(pconcellos$id)
df.mapa.concellos$id = as.character(df.mapa.concellos$id)

pconcellos2 = pconcellos
pconcellos2$id = str_replace_all(toupper(pconcellos2$id), c(`AS ` = "", `, AS` = "",
    `OS ` = "", `, OS` = ""))
pconcellos2$id = str_replace_all(toupper(pconcellos2$id), c(`A ` = "", `, A` = "",
    `O ` = "", `, O` = ""))


pobconcellos = read.csv2("poblacion.csv")
colnames(pobconcellos)[1] = "id"
pobconcellos$id = str_replace_all(toupper(pobconcellos$id), c(`AS ` = "", `, AS` = "",
    `OS ` = "", `, OS` = ""))
pobconcellos$id = str_replace_all(toupper(pobconcellos$id), c(`A ` = "", `, A` = "",
    `O ` = "", `, O` = ""))

radconcellos = read.csv2("radiacionconcellos.csv")
colnames(radconcellos)[1] = "id"
radconcellos$id = str_replace_all(toupper(radconcellos$id), c(`AS ` = "", `, AS` = "",
    `OS ` = "", `, OS` = ""))
```

```
radconcellos$id = str_replace_all(toupper(radconcellos$id), c(`A ` = "", `, A` = "",
    `O ` = "", `, O` = ""))
radconcellos[, -1] = apply(radconcellos[, -1], 2, FUN = makefloat)


df.mapa.concellos2 = df.mapa.concellos
df.mapa.concellos2$id = str_replace_all(toupper(df.mapa.concellos2$id), c(`AS ` = "",
    `, AS` = "", `OS ` = "", `, OS` = ""))
df.mapa.concellos2$id = str_replace_all(toupper(df.mapa.concellos2$id), c(`A ` = "",
    `, A` = "", `O ` = "", `, O` = ""))

pconcellos3 = pconcellos2[!pconcellos2$id %in% c("COTOBADE", "CERDEDO"), ]
aux1 = pconcellos2[pconcellos2$id == "COTOBADE", -c(1, 2)]
aux2 = pconcellos2[pconcellos2$id == "CERDEDO", -c(1, 2)]
aux = rbind(aux1, aux2)
aux = t(as.data.frame(colMeans(aux)))
aux1 = data.frame(as.factor("??"), "CERDEDO-COTOBADE")
aux = cbind(aux1, aux)
names(aux) = names(pconcellos3)
pconcellos3 = rbind(pconcellos3, aux)


df.mapa.concellos2 <- left_join(df.mapa.concellos2, pconcellos3, by = "id")
df.mapa.concellos2 <- left_join(df.mapa.concellos2, pobconcellos, by = "id")
df.mapa.concellos2 <- left_join(df.mapa.concellos2, radconcellos, by = "id")
df.mapa.concellos2 <- left_join(df.mapa.concellos2, atmconcellos, by = "id")

df.datos.concellos <- pconcellos3
df.datos.concellos <- left_join(df.datos.concellos, pobconcellos, by = "id")
df.datos.concellos <- left_join(df.datos.concellos, radconcellos, by = "id")
df.datos.concellos <- left_join(df.datos.concellos, atmconcellos, by = "id")
write.csv2(df.datos.concellos, "df.datos.concellos.csv")

df.datos.latex = df.datos.concellos[, -c(1, 3, 5, 7, 9, 10, 11, 12, 13, 14,
    15, 16, 17, 18, 19, 20, 21, 22, 23, 29, 30, 33)]
write.csv2(df.datos.latex, "df.datos.latex.csv")


###### cloromapas de datos######
cloromapa2 <- ggplot(data = dfm.concellos, aes(x = long, y = lat, group = group))
p0 = cloromapa2 + geom_polygon(aes(fill = RME.suavizada.T), color = "gray",
    size = 0.1) + scale_fill_gradient(high = "#000067", low = "#0080ff", guide = "colorbar") +
    coord_fixed(1.3)
```

```
p1 = cloromapa2 + geom_polygon(aes(fill = RME.suavizada.H), color = "gray",
    size = 0.1) + scale_fill_gradient(high = "#000067", low = "#0080ff", guide = "colorbar") +
    coord_fixed(1.3)

p2 = cloromapa2 + geom_polygon(aes(fill = RME.suavizada.M), color = "gray",
    size = 0.1) + scale_fill_gradient(high = "#000067", low = "#0080ff", guide = "colorbar") +
    coord_fixed(1.3)

p3 = cloromapa2 + geom_polygon(aes(fill = RME.suavizada.T), color = "gray",
    size = 0.1) + scale_fill_gradient(high = "#c20101", low = "#0080ff", guide = "colorbar") +
    coord_fixed(1.3)

p33 = cloromapa2 + geom_polygon(aes(fill = log(1000 * (RME.suavizada.T + 1))),
    color = "gray", size = 0.1) + scale_fill_gradient(high = "#c20101", low = "#0080ff",
    guide = "colorbar", xlab("Z")) + coord_fixed(1.3)


p4 = cloromapa2 + geom_polygon(aes(fill = log(T)), color = "gray", size = 0.1) +
    scale_fill_gradient(high = "#c20101", low = "#0080ff", guide = "colorbar",
        xlab("log(población)")) + coord_fixed(1.3)
p44 = cloromapa2 + geom_polygon(aes(fill = T), color = "gray", size = 0.1) +
    scale_fill_gradient(high = "#c20101", low = "#0080ff", guide = "colorbar",
        xlab("población(nºhab)")) + coord_fixed(1.3)



p5 = cloromapa2 + geom_polygon(aes(fill = BQ300), color = "gray", size = 0.1) +
    scale_fill_gradient(high = "#c20101", low = "#0080ff", guide = "colorbar") +
    coord_fixed(1.3)

p6 = cloromapa2 + geom_polygon(aes(fill = PM10), color = "gray", size = 0.1) +
    scale_fill_gradient(high = "#c20101", low = "#0080ff", guide = "colorbar",
        xlab("PM10 (µg/m^3)")) + coord_fixed(1.3)

p7 = cloromapa2 + geom_polygon(aes(fill = GM), color = "gray", size = 0.1) +
    scale_fill_gradient(high = "#c20101", low = "#0080ff", guide = "colorbar",
        xlab("Rad. Media (Bq)")) + coord_fixed(1.3)

grid.arrange(p7, p6, p44, p4, layout_matrix = rbind(c(1, 1, 2, 2), c(1, 1, 2,
    2), c(3, 3, 4, 4), c(3, 3, 4, 4)))

grid.arrange(p7, p6, p44, layout_matrix = rbind(c(1, 1, 2, 2), c(1, 1, 2, 2),
    c(NA, 3, 3, NA), c(NA, 3, 3, NA)))
```

```
grid.arrange(p7, p6, p44, layout_matrix = rbind(c(1, 1), c(1, 1), c(2, 2), c(2,
    2), c(3, 3), c(3, 3)))


###### vecindades######
concellos.ROOK.nb = poly2nb(mapa.concellos, queen = F)

concellos.QUEEN.nb = poly2nb(mapa.concellos, queen = T)

coords <- coordinates(mapa.concellos)
IDs <- row.names(as(mapa.concellos, "data.frame"))

concellos.tri.nb <- tri2nb(coords, row.names = IDs)

concellos.SOI.nb <- graph2nb(soi.graph(concellos.tri.nb, coords), row.names = IDs)

concellos.gabriel.nb <- graph2nb(gabrielneigh(coords), row.names = IDs)
concellos.relative.nb <- graph2nb(relativeneigh(coords), row.names = IDs)

concellos.1nn.nb <- knn2nb(knearneigh(coords, k = 1), row.names = IDs)
concellos.2nn.nb <- knn2nb(knearneigh(coords, k = 2), row.names = IDs)
concellos.4nn.nb <- knn2nb(knearneigh(coords, k = 4), row.names = IDs)
concellos.2nnsym.nb <- make.sym.nb(concellos.2nn.nb)
concellos.4nnsym.nb <- make.sym.nb(concellos.4nn.nb)

# por inverso de d
dsts = nbdists(concellos.4nn.nb, coords = coords)
idw = lapply(dsts, function(x) 1/(x/1000))
concellos.4nnIDW.lw = nb2listw(concellos.4nn.nb, glist = idw, style = "B")
# exponencial negativa de d
edw = lapply(dsts, function(x) exp(-x/1000))
concellos.4nnEDW.lw = nb2listw(concellos.4nn.nb, glist = edw, style = "B")

###### genreador de mapas de vecindad######
concellos.nb.mat <- nb2listw(concellos.ROOK.nb, style = "W", zero.policy = TRUE)

par(mar = c(0, 0, 2, 2))
par(mfrow = c(2, 1))
q0 = plot(mapa.concellos, border = "darkgray", main = "Rook")
q1 = plot(concellos.nb.mat, coords = coordinates(mapa.concellos), pch = 19,
    cex = 0.5, col = "black", lwd = 1.5, add = T)


concellos.nb.mat <- nb2listw(concellos.SOI.nb, style = "W", zero.policy = TRUE)
```

```
q0 = plot(mapa.concellos, border = "darkgray", main = "SOI")
q1 = plot(concellos.nb.mat, coords = coordinates(mapa.concellos), pch = 19,
    cex = 0.5, col = "black", lwd = 1.5, add = T)




concellos.nb.mat <- nb2listw(concellos.2nnsym.nb, style = "W", zero.policy = TRUE)
par(mar = c(0, 0, 2, 2))
par(mfrow = c(2, 1))
q0 = plot(mapa.concellos, border = "darkgray", main = "knn k=2")
q1 = plot(concellos.nb.mat, coords = coordinates(mapa.concellos), pch = 19,
    cex = 0.5, col = "black", lwd = 1.5, add = T)


concellos.nb.mat <- nb2listw(make.sym.nb(concellos.4nn.nb), style = "W", zero.policy = TRUE)
q0 = plot(mapa.concellos, border = "darkgray", main = "knn k=4")
q1 = plot(concellos.nb.mat, coords = coordinates(mapa.concellos), pch = 19,
    cex = 0.5, col = "black", lwd = 1.5, add = T)


###### tests descriptiva//1er bloque del cap4######

df.datos.concellos = read.csv2("df.datos.concellos.csv")
Y = df.datos.concellos$RME.suavizada.T
Z = log(1000 * (Y + 1))
par(mfrow = c(1, 2))
hist(Y, breaks = 10)
hist(Z, breaks = 15)
shapiro.test(Y)
shapiro.test(Z)

###### modelo lineal general######
library(RColorBrewer)
library(rgdal, quietly = T)
library(rgeos, quietly = T)
mapa.concellos <- readOGR(dsn = "./Concellos", layer = "Concellos_IGN")
my.palette <- rev(brewer.pal(n = 9, name = "RdBu"))

df.datos.concellos = read.csv2("df.datos.concellos.csv")
lm = lm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T), data = df.datos.concellos)
summary(lm)
par(mfrow = c(2, 2))
```

```
plot(lm)

mapa.concellos$residuos.lm = residuals(lm)
grps = 9
brks <- quantile(mapa.concellos$residuos.lm, 0:(grps - 1)/(grps - 1), na.rm = TRUE)
spplot(mapa.concellos, "residuos.lm", at = brks, col = "darkgray", col.regions = my.palette,
    main = "residuos.lm")

###### moran######

library(rgdal, quietly = T)
library(rgeos, quietly = T)
library(spdep, quietly = T)
library(gridExtra, quietly = T)
library(RANN, quietly = T)
mapa.concellos <- readOGR(dsn = "./Concellos", layer = "Concellos_IGN")
coords <- coordinates(mapa.concellos)

df.datos.concellos = read.csv2("df.datos.concellos.csv")
lm = lm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T), data = df.datos.concellos)

# rook neighbors
concellos.ROOK.nb <- poly2nb(mapa.concellos, queen = F)
dsts = nbdists(concellos.ROOK.nb, coords = coords)
idw = lapply(dsts, function(x) 1/(x/1000))
concellos.ROOKB.lw = nb2listw(concellos.ROOK.nb, style = "B", zero.policy = T)
concellos.ROOKIDW.lw = nb2listw(concellos.ROOK.nb, glist = idw, style = "B",
    zero.policy = T)
lm.morantest(lm, listw = concellos.ROOKB.lw, zero.policy = T)
lm.morantest(lm, listw = concellos.ROOKIDW.lw, zero.policy = T)

# SOI neighbors
IDs <- row.names(as(mapa.concellos, "data.frame"))
concellos.tri.nb <- make.sym.nb(tri2nb(coords, row.names = IDs))
concellos.SOI.nb <- make.sym.nb(graph2nb(soi.graph(concellos.tri.nb, coords),
    row.names = IDs))
dsts = nbdists(concellos.SOI.nb, coords = coords)
idw = lapply(dsts, function(x) 1/(x/1000))
concellos.SOIB.lw = nb2listw(concellos.SOI.nb, style = "B")
concellos.SOIIDW.lw = nb2listw(concellos.SOI.nb, glist = idw, style = "B")
lm.morantest(lm, listw = concellos.SOIB.lw, zero.policy = T)
lm.morantest(lm, listw = concellos.SOIIDW.lw, zero.policy = T)

# k=4 knn nieghbors
```

```
concellos.4nn.nb <- make.sym.nb(knn2nb(knearneigh(coords, k = 4), row.names = IDs))
dsts = nbdists(concellos.4nn.nb, coords = coords)
idw = lapply(dsts, function(x) 1/(x/1000))
concellos.4nnB.lw = nb2listw(concellos.4nn.nb, style = "B")
concellos.4nnIDW.lw = nb2listw(concellos.4nn.nb, glist = idw, style = "B")
lm.morantest(lm, listw = concellos.4nnB.lw, zero.policy = T)
lm.morantest(lm, listw = concellos.4nnIDW.lw, zero.policy = T)


###### weight matplot######
library(rgdal, quietly = T)
library(rgeos, quietly = T)
library(spdep, quietly = T)
library(gridExtra, quietly = T)
library(RANN, quietly = T)
mapa.concellos <- readOGR(dsn = "./Concellos", layer = "Concellos_IGN")

coords <- coordinates(mapa.concellos)
IDs <- row.names(as(mapa.concellos, "data.frame"))
concellos.4nn.nb <- make.sym.nb(knn2nb(knearneigh(coords, k = 4), row.names = IDs))
dsts = nbdists(concellos.4nn.nb, coords = coords)
idw = lapply(dsts, function(x) 1/(x/1000))
concellos.4nnB.lw = nb2listw(concellos.4nn.nb, style = "B")
concellos.4nnIDW.lw = nb2listw(concellos.4nn.nb, glist = idw, style = "B")
image(listw2mat(concellos.4nnB.lw), col = gray.colors(50, end = 1, start = 0,
    rev = T))
image(listw2mat(concellos.4nnIDW.lw), col = c("#FFFFFF", gray.colors(49, end = 0.8,
    start = 0, rev = T)))

###### spatial models######
library(rgdal, quietly = T)
library(rgeos, quietly = T)
library(spdep, quietly = T)
library(gridExtra, quietly = T)
library(RANN, quietly = T)
mapa.concellos <- readOGR(dsn = "./Concellos", layer = "Concellos_IGN")
coords <- coordinates(mapa.concellos)

df.datos.concellos = read.csv2("df.datos.concellos.csv")


concellos.4nn.nb <- make.sym.nb(knn2nb(knearneigh(coords, k = 4), row.names = IDs))
concellos.4nnB.lw = nb2listw(concellos.4nn.nb, style = "B")
IDs <- row.names(as(mapa.concellos, "data.frame"))
```

```
concellos.tri.nb <- make.sym.nb(tri2nb(coords, row.names = IDs))
concellos.SOI.nb <- make.sym.nb(graph2nb(soi.graph(concellos.tri.nb, coords),
    row.names = IDs))
concellos.SOIB.lw = nb2listw(concellos.SOI.nb, style = "B")



slm.SAR.SOI = spautolm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T),
    data = df.datos.concellos, listw = concellos.SOIB.lw, family = "SAR")
mapa.concellos$slm.SAR.SOI = residuals(slm.SAR.SOI)
summary(slm.SAR.SOI)
slm.CAR.SOI = spautolm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T),
    data = df.datos.concellos, listw = concellos.SOIB.lw, family = "CAR")
mapa.concellos$slm.CAR.SOI = residuals(slm.CAR.SOI)
summary(slm.CAR.SOI)

slm.SAR.KNN = spautolm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T),
    data = df.datos.concellos, listw = concellos.4nnB.lw, family = "SAR")
mapa.concellos$slm.SAR.KNN = residuals(slm.SAR.KNN)
summary(slm.SAR.KNN)
slm.CAR.KNN = spautolm(log(1000 * (RME.suavizada.T + 1)) ~ PM10 + GM + log(T),
    data = df.datos.concellos, listw = concellos.4nnB.lw, family = "CAR")
mapa.concellos$slm.CAR.KNN = residuals(slm.CAR.KNN)
summary(slm.CAR.KNN)



library(RColorBrewer)
par(mar = c(0, 0, 0, 0))
par(mfrow = c(2, 2))
my.palette <- rev(brewer.pal(n = 9, name = "RdBu"))

grps = 9
brks <- quantile(mapa.concellos$slm.SAR.KNN, 0:(grps - 1)/(grps - 1), na.rm = TRUE)
p3 = spplot(mapa.concellos, "slm.SAR.KNN", at = brks, col = "darkgray", col.regions = my.palette,
    main = "slm.SAR.KNN")

brks <- quantile(mapa.concellos$slm.CAR.KNN, 0:(grps - 1)/(grps - 1), na.rm = TRUE)
p4 = spplot(mapa.concellos, "slm.CAR.KNN", at = brks, col = "darkgray", col.regions = my.palette,
    main = "slm.CAR.KNN")

brks <- quantile(mapa.concellos$slm.SAR.SOI, 0:(grps - 1)/(grps - 1), na.rm = TRUE)
p1 = spplot(mapa.concellos, "slm.SAR.SOI", at = brks, col = "darkgray", col.regions = my.palette,
    main = "res.SAR.SOI")
```

```
brks <- quantile(mapa.concellos$slm.CAR.SOI, 0:(grps - 1)/(grps - 1), na.rm = TRUE)
p2 = spplot(mapa.concellos, "slm.CAR.SOI", at = brks, col = "darkgray", col.regions = my.palette,
    main = "res.CAR.SOI")

grid.arrange(p1, p2, p3, p4, layout_matrix = rbind(c(1, 1, 2, 2), c(1, 1, 2,
    2), c(3, 3, 4, 4), c(3, 3, 4, 4)))
```

Bibliografía

[1] N. A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, New York, revised edition, 1993.

[2] W. B. Mercer, A. D. Hall. The experimental error of field trials. *Journal of Agricultural Science*, 4:107-132, 1911.

[3] M. M. Fischer, J. Wang. *Spatial Data Analysis: Models Methods and Techniques*. Springer, Heidelberg Dordrecht London New York, 2011.

[4] S. M. Cramb, E. W. Duncan, N. M. White, P. D. Baade, K. L. Mengersen. *Spatial Modelling Methods.*. Cancer Council Queensland and Queensland University of Technology, Brisbane, 6 2016.

[5] J. Dubé, D. Legros. *Spatial Econometrics Using Microdata*. John Wiley & Sons, Hoboken, ISTE, London, 2014.

[6] R. S. Bivand, E. Pebesma, V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer, Heidelberg Dordrecht London New York, 2 edition, 2013.

[7] B. D. Ripley. *Spatial Statistics*. John Wiley & Sons, Hoboken, 2004.

[8] R. S. Bivand. Creating Neighbours. The Comprehensive R Archive Network, 4 2019.

[9] R. Haining. *Spatial Data Analysis: Theory and Practice*. Cambridge University Press, Cambridge, 2004.

[10] L. A. Waller, C. A. Gotway. *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Hoboken, 2004.

[11] V. Pérez Muñuzuri et al. Informe anual da calidade do aire en Galicia. Informe Técnico, Consellería de Medio Ambiente, Territorio e Infraestruturas, Santiago de Compostela, 2014.

[12] N. A. Cressie, N. Chan. Spatial modeling of regional variables. *Journal Of The American Statistical Association*, 84(406):393-401, 1989.

[13] M. Tiefelsdorf, B. Boots. The Exact Distribution of Moran's I. *Environment and Planning A: Economy and Space*, 27(6):985-999, 6 1995.

[14] L. Anselin. Local Indicators of Spatial Association—LISA. *Geographical Analysis*, 27(2):93-115, 9 2010.

[15] D. Krewski et al. Residential Radon and Risk of Lung Cancer: A Combined Analysis of 7 North American Case-Control Studies. *Epidemiology*, 16(2):137-145, 3 2005

[16] S. H. Fairfield. An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28:1-23, 1938.

[17] S. H. Downs et al. Reduced Exposure to PM10 and Attenuated Age-Related Decline in Lung Function. *The New England Journal of Medicine*, 12 2017.

[18] L. Anselin. Moran's I and Geary's c [video]. 8 2016. Desde `https://youtu.be/GVTjYXX5BuA`

[19] L. Anselin. Global Spatial Autocorrelation (1) Moran Scatter Plot and Spatial Correlogram. 2018. Desde `https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn4.`