Modelo Lineal de Efectos Mixtos: Una aplicación a Datos Temporal y Espacialmente Correlacionados

Carolina Romero Coronado
Estadística

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Noviembre de 2019

Modelo Lineal de Efectos Mixtos: Una aplicación a Datos Temporal y Espacialmente Correlacionados

Carolina Romero Coronado
Estadística

Trabajo de Grado presentado para optar al título de
Magíster en Ciencias - Estadística

Director
Luis Guillermo Díaz Monroy
Profesor Asociado

Línea de investigación
Análisis de Medidas Repetidas

Universidad Nacional de Colombia
Facultad de Ciencias
Departamento de Estadística
Bogotá, D.C.
Noviembre de 2019

Título en español

Modelo Lineal de Efectos Mixtos: Una aplicación a Datos Temporal y Espacialmente Correlacionados

Title in English

Linear Mixed Effects Model: An application to Temporal and Spatially Correlated Data

Resumen: Modelar correlación espacial y temporal en simultáneo se ha convertido en un tema de interés para diferentes contextos, especialmente en el contexto geoestadístico, pues la realización de una predicción espacial óptima en sitios no muestreados para determinada variable regionalizada en estudio, se encuentra ligada a la correcta identificación de dependencias existentes entre dicha variable regionalizada y la componente longitudinal de la misma (esto es, el instante de tiempo en el que fue medida). Se contempla una ampliación de la metodología aplicada por (Militino et al., 2008), usando los *modelos lineales mixtos* (*LMM* por sus siglas en inglés), pero adicionando un análisis que involucre el uso de metodologías para datos espaciales y datos longitudinales, en aras de visualizar las implicaciones que tiene el modelar vía *LMM*, olvidando y contemplando la correlación espacial inherente a los datos del proceso a estudiar. La estimación del modelo propuesto se hará *vía máxima verosimilitud restringida* (*REML*, en inglés).

Abstract: Modeling spatial and temporal correlation simultaneously has become a topic of interest for different contexts, especially in the geostatistical context, since the realization of an optimal spatial prediction in sites not sampled for a given regionalized variable under study, is linked to the correct identification of existing dependencies between said regionalized variable and the longitudinal component thereof (that is, the moment of time in which it was measured). An extension of the methodology applied by (Militino et al., 2008) is contemplated, using the *mixed linear models* (*LMM* for its acronym in English), but adding an analysis that involves the use of methodologies for spatial data and longitudinal data, in order to visualize the implications of modeling via *LMM*, forgetting and contemplating the spatial correlation inherent in the data of the process to be studied. The estimation of the proposed model will be done by *restricted maximum likelihood* (*REML*).

Palabras clave: Análisis de medidas repetidas, correlación espacial y temporal, modelo lineal mixto, *REML*, modelos *kriging*.

Keywords: Analysis of repeated measures, spatial and temporal correlation, mixed linear model, *REML*, *Kriging* models.

Nota de aceptación

Trabajo de tesis
Aprobado

Jurado
Luz Marina Rondon Poveda

Jurado
Martha Patricia Bohorquez Castañeda

Director
Luis Guillermo Díaz Monroy

Bogotá, D.C., Noviembre de 2019

Dedicado a

A mis padres, Esther Elena Coronado Ríos y César Oswaldo Romero Gutiérrez, quienes se encargaron de convertirme en la mujer que soy. A mi abuelita, María Dila Ríos Varela, mi motor. A mis hermanos, Natalia y Sergio, por ser los mejores compañeros de camino.

Agradecimientos

El autor expresa sus agradecimientos a todos aquellos que de una u otra forma han colaborado, contribuido o aportado en el desarrollo de este trabajo.

Índice general

Índice de tablas

Índice de figuras

Introducción

El estudio de *medidas repetidas* o *datos correlacionados* involucra el tratamiento y manejo de la información asociada a una variable en estudio, para la cual se han tomado datos de manera reiterativa. Se tienen entonces n unidades en estudio, éstas pueden ser experimentales, muestrales o casos, sobre las cuales se mide la variable de interés repetidamente: cuando la reiteración en las mediciones sobre la misma unidad se hace considerando periodos de tiempo, se habla de *datos longitudinales* (Davis, 2002), mientras que si las unidades corresponden a n observaciones espaciales $Z(s_i), i = 1, \ldots, n$, con $Z(\cdot)$ una variable de interés, se habla de estudios de tipo *geoestadístico*, con interés particular en la variable regionalizada $Z(\cdot)$. (Diggle & Ribeiro Jr, 2007)

De esta forma, en el contexto de los datos longitudinales se evidencia correlación temporal entre las observaciones tomadas en tiempos diferentes sobre la misma unidad y en los datos geoestadísticos, hay presencia de correlación espacial entre la información recolectada en puntos diferentes sobre la misma zona geográfica; así pues, el fin último en ambos tipos de medidas repetidas recae en el modelamiento adecuado de dicha correlación, por medio de las matrices de varianzas y covarianzas involucradas en cada contexto. (Davis, 2002)

Sin embargo, existen escenarios en los que se hace necesario modelar la correlación temporal y espacial en simultáneo pues la naturaleza de los datos así lo exige. Militino et al. (2008) resaltan, por ejemplo, la importancia de considerar el impacto de agentes contaminantes como el Nitrato (NO_3) sobre las aguas subterráneas, siendo estás de gran utilidad en la industria agrícola y en el abastecimiento de agua potable; la disbutrición del NO_3 en el agua subterránea posee una estructura espacial, pero también depende de la estación del año en la que se realice la medición, debido a las características propias del suelo durante las mismas.

Otro ámbito en el que es conveniente modelar correlación temporal y espacial en simultáneo, es el ámbito epidemiológico, específicamente considere áreas de la medicina relacionadas con neurociencia. En dichas áreas, el registro electroencefalográfico (EEG) proporciona una poderosa medida de la dinámica neuronal subyacente a la cognición humana, la cual se encuentra asociada con entender cómo trabaja la memoria. Sin embargo, el análisis de datos EEG multidimensionales es un desafío porque requiere el modelado de correlaciones temporales y espaciales para determinar las características del EEG que brindan información respecto al rendimiento de la memoria. (Bi et al., 2015)

Así, la necesidad de modelar en conjunto ambos tipos de variabilidad ha sido tema de estudio y varios son los enfoques que han sido propuestos: modelos espacio-temporales geoestadísticos (Kyriakidis & Journel, 1999), uso de variogramas espacio-temporales que

tratan el tiempo como una tercera dimensión (Bogaert & Christakos, 1997), diversos enfoques semiparamétricos (Angulo et al., 1998), modelos espacio-temporales no separables (Stein, 2005), filtro espacio-temporal de Kalman (Mardia et al., 1998) y modelos de variograma espacio-temporales flexibles (Fernández-Casal et al., 2003); sin embargo, siguiendo a Militino et al. (2008), en la mayoría de trabajos se considera que se cuenta con una gran cantidad de datos temporales y en la práctica se requieren métodos alternativos, en particular cuando el objetivo es proporcionar predicciones globales en un área determinada a partir de los datos recopilados en el tiempo dentro del contexto de datos longitudinales, esto es, cuando se dispone de pocos tiempos de medición para cada individuo de estudio.

Para subsanar los inconvenientes antes mencionados, Militino et al. (2008) proponen modelar la correlación espacial y temporal, en el caso de datos longitudiales, vía un modelo lineal mixto que tenga en cuenta, de manera adecuada, la dependencia espacial y temporal presente en los datos. El resultado final involucra una extensión del *Kriging*, incorporando la dimensión longitudinal. Cepeda (2011) también propone una metodología para modelar correlación espacial y temporal en datos de tipo longitudinal, usando herramientas de la teoría bayesiana.

Así, en este proyecto se contempla una ampliación de la metodología aplicada por Militino et al. (2008), usando también los *modelos lineales mixtos* (*LMM* por sus siglas en inglés), pero adicionando un análisis que involucre el uso de metodologías para datos espaciales y datos longitudinales, en aras de visualizar las implicaciones que tiene el modelar vía *LMM*, olvidando y contemplando la correlación espacial inherente a los datos del proceso a estudiar. La estimación del modelo propuesto se hará *vía máxima verosimilitud restringida* (*REML*, en inglés), desarrollando una aplicación a datos tomados sobre el territorio de España.

Los dos primeros capítulos se encargan de presentar la teoría entorno al *Modelo Lineal para Datos Espaciales* y el *Modelo Lineal de Efectos Mixtos*, necesaria para el desarrollo de la aplicación, implementando además la modelación vía *LMM*, sin contemplar la dependencia espacial presente en los datos que serán estudiados. En el tercer capítulo se enuncia el modelo teórico de Militino et al. (2008) que será aplicado, realizando el ajuste del *LMM* que contempla ambos tipos de correlación, para en el cuarto capítulo realizar la comparación entre el modelo ajustado sin tener en cuenta la dependencia espacial y el *LMM* que involucra la correlación espacial y temporal.

CAPÍTULO 1

Modelo lineal para datos espaciales

Siguiendo a Giraldo (2011) y a Schabenberger & Gotway (2005), la *estadística espacial* es un conjunto de metódos apropiados para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. Formalmente, la estadística espacial trata con el análisis de realizaciones de un proceso estocástico $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$, en el que s es la ubicación en el espacio euclidiano d-dimensional y $\mathbf{Z}(s)$ es una variable aleatoria en la ubicación s.

Los datos espaciales pueden ser de tres tipos, según el conjunto de índices, D, en el que el proceso se lleve a cabo: *Geoestadística*, *Datos de Área* y *Patrones Puntuales* (Giraldo, 2011). Para efectos del presente proyecto, se trabajará con datos geoestadísticos.

1.1. Geoestadística

Cuando el conjunto de índices D en el que se está trabajando es continuo y fijo, la *Geoestadística* es la rama de la estadística espacial encargada de tratarlos (Giraldo, 2011). Que D sea continuo hace referencia a que $\mathbf{Z}(s)$ puede ser observada en cualquier punto dentro de D, esto es, entre cualquier par de ubicaciones muestreadas s_i y s_j pueden encontrarse infinitas ubicaciones en las que también se puede obtener información de la variable en estudio (Schabenberger & Gotway, 2005); que D sea fijo hace referencia a que las ubicaciones s son elegidas a juicio del investigador.

De acuerdo con Cressie (1993), áreas como la geología, ciencias del suelo, agronomía, ingeniería forestal, astronomía, o cualquier otra que trabaje con datos colectados en diferentes locaciones espaciales, necesita desarrollar modelos que indiquen cuándo hay dependencia entre las medidas de los diferentes sitios. Usualmente dicha modelación se relaciona con la predicción espacial, pero hay otras áreas importantes como la simulación y el diseño muestral.

Así, la geoestadística se encarga de estudiar fenómenos espaciales como los que las anteriores áreas involucran, con el objetivo principal de estimar, predecir y simular dichos fenómenos (Warrick & Myers, 1987). Permite además describir la continuidad espacial, rasgo distintivo en muchos fenómenos naturales y proporciona adaptaciones de las técnicas clásicas de regresión para tomar ventajas de esta continuidad (Isaaks & Srivastava, 1989).

Si lo que interesa es la predicción espacial, la goestadística opera en dos etapas, tal como lo menciona Giraldo (2011): la primera corresponde al análisis estructural, describiéndose la correlación entre puntos en el espacio; en la segunda fase se obtiene la predicción en localizaciones no muestreadas por medio de la técnica *kriging*, la cual consiste en calcular un promedio ponderado de las observaciones muestrales. Los pesos son escogidos según la estructura espacial de correlación establecida en la primera etapa y por la configuración de muestreo (Petitgas, 1996).

En los siguientes apartados se estudiará cada una de éstas dos etapas, mencionando las características teóricas de mayor relevancia.

1.1.1. Concepto de variable regionalizada

Se define *variable regionalizada* como aquella variable medida en el espacio de forma que presente una estructura de correlación. Formalmente se puede definir como un proceso estocástico con dominio contenido en un espacio euclidiano d-dimensional $\mathbb{R}^d$, $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$ (Giraldo, 2011). Si $d = 2$, $\mathbf{Z}(s)$ puede asociarse a una variable medida en un punto s del plano (Schabenberger & Gotway, 2005). Así, $\mathbf{Z}(s)$ puede verse como una medición de una variable aleatoria (por ejemplo, pH del suelo) en un punto s de una región de estudio. (Giraldo, 2011)

Debido a que tal proceso estotástico se define como una colección de variables aleatorias indexadas, para cada s en el conjunto de índices D, $\mathbf{Z}(s)$ es una variable aleatoria (s representa las coordenadas del punto en estudio y $\mathbf{Z}$ la variable en cada una de éstas). (Giraldo, 2011)

Sea $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$ el proceso estocástico que define la variable regionalizada. Para cualquier n puntos $s_1, s_2, \ldots, s_n$, el vector aleatorio $\mathbf{Z}(s) = [Z(s_1), Z(s_2), \ldots, Z(s_n)]'$ está definido por su función de distribución conjunta:

$$F(z_1, z_2, \ldots, z_n) = P(Z(s_1) < z_1, Z(s_2) < z_2, \ldots, Z(s_n) < z_n).$$

De este modo, explicitadas las densidades marginales univariadas y bivariadas, es posible establecer los momentos asociados a las mismas, a través de la expresiones relacionadas con toda pareja $Z(s), Z(s+h)$:

- $E(Z(s)) = \mu(s)$
- $Var(Z(s)) = E[Z(s) - \mu(s)]^2 = \sigma^2$
- $C(s,h) = Cov(Z(s), Z(s+h)) = E[Z(s) - \mu(s)][Z(s+h) - \mu(s+h)]$; conocida como *función de covarianza* de un proceso espacial.
- $$\gamma(s,h) = \frac{1}{2} Var[Z(s) - Z(s+h)]; \tag{1.1}$$
  conocida como *función de semivarianza* de un proceso espacial.

1.1.2. Supuesto de estacionariedad y tipos de estacionariedad

Una variable regionalizada es *estrictamente estacionaria* o cumple el supuesto de *estacionariedad fuerte*, si su función de distribución conjunta es invariante bajo una translación

de las coordenadas, esto es, la función de distribución del vector aleatorio $\mathbf{Z}(s)$, definida como $P(Z(s_1) < z_1, Z(s_2) < z_2, \ldots, Z(s_n) < z_n)$, es igual a la función de distribución del vector $\mathbf{Z}(s+h)$, la cual es $P(Z(s_1+h) < z_1, Z(s_2+h) < z_2, \ldots, Z(s_n+h) < z_n)$, para todo n y h. (Schabenberger & Gotway, 2005)

Debido a que la teoría geoestadística se fundamenta en los momentos de la variable regionalizada en estudio (Giraldo, 2011), los supuestos de estacionariedad pueden ser definidos en términos de dichos momentos, dando lugar a un segundo y tercer tipo de estacionaridad conocidos como *estacionariedad de segundo orden* y *estacionariedad débil* o *intrínseca*.

La variable regionalizada $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$ cumple el supuesto de *estacionariedad de segundo orden* si $E[Z(s)] = \mu$ y si $Cov[Z(s), Z(s+h)] = C(h)$; así, la media de una variable regionalizada estacionaria de segundo orden es constante y finita para todo punto en el dominio, D, y la covarianza entre toda pareja $\{Z(s), Z(s+h)\}$, existe, es finita y depende únicamente del vector de separación h. (Schabenberger & Gotway, 2005)

Bajo el supuesto de estacionariedad de segundo orden, que la covarianza exista, implica que la varianza existe, es finita y no depende de h, pues $Cov[Z(s), Z(s+0)] = Var[Z(s)] = C(0) = \sigma^2$. De este modo, la variabilidad en un proceso estacionario de segundo orden es la misma en todo lugar sobre el dominio en estudio. Además, de (1.1) se tiene la siguiente relación entre la función de covarianza y la de semivarianza:

$$\begin{aligned} \gamma(s,h) &= \frac{1}{2}\{Var[Z(s)] + Var[Z(s+h)] - 2Cov[Z(s), Z(s+h)]\} \\ &= \frac{1}{2}\left\{2\sigma^2 - 2C(h)\right\} \\ &= C(0) - C(h) \end{aligned} \tag{1.2}$$

Por otro lado, existen procesos espaciales, definidos por una variable regionalizada particular, en los que no se cuenta con la propiedad de varianza finita. En estos casos, según menciona Clark (1979) (citado en Giraldo (2011)), se debe considerar sólo el supuesto en el que los incrementos $\mathbf{Z}(s) - \mathbf{Z}(s+h)$ sean estacionarios, pues aún cuando $\mathbf{Z}(s)$ viole el supuesto de estacionariedad de segundo orden, tales incrementos podrían cumplirlo.

Procesos que presenten esa última característica se dice cumplen el supuesto de *estacionariedad intrínseca*, definida formalmente como sigue: el proceso $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$ es intrínsecamente estacionario si $E[Z(s)] = \mu$ y $\gamma(h) = \frac{1}{2}Var[Z(s) - Z(s+h)]$. Dicho de otro modo, si $Z(s)$ tiene media finita y constante para todo punto en D y si para cualquier vector h, la varianza está definida y depende únicamente de la distancia entre puntos en D, el proceso espacial es estacionario intrínseco. Bajo estacionariedad intrínseca la esperanza de los incrementos será nula, pues $E[Z(s) - Z(s+h)] = E[Z(s)] - E[Z(s)] = 0$. (Schabenberger & Gotway, 2005)

En síntesis, existen tres tipos de estacionariedad para procesos espaciales: *estacionariedad estricta o fuerte*, de *segundo orden* e *intrínseca*; la estacionariedad fuerte implica estacionariedad de segundo orden, pero no se tiene el recíproco; la estacionariedad de segundo orden implica estacionaridad intrínseca, pero el recíproco no se tiene tampoco en este caso; así, por transitividad, estacionaridad fuerte implica estacionariedad intrínseca. (Schabenberger & Gotway, 2005)

De este modo, las relaciones de implicación entre los tipos de estacionariedad pueden presentarse como sigue:

Estacionariedad Fuerte $\implies$ *Estacionariedad de* 2^{do} *Orden* $\implies$ *Estacionariedad Intrínseca.*

Detectar que un proceso espacial cumple el supuesto de estacionariedad puede ser complejo, pues debe garantizarse que para todo el dominio en estudio la media es constante y que se mantiene la misma forma funcional de la covarianza en todas la direcciones en las que el proceso pueda desarrollarse. Además, el proceso también debe ser *isotrópico*, esto es, que la correlación entre los datos no dependa de la dirección en la que misma sea medida, en caso contrario, se hablará de presencia de *anisotropía*. (Giraldo, 2011)

En la práctica, suelen utilizarse diagramas de dispersión de la variable regionalizada en función de las coordenadas en las que tiene lugar, para identificar tendencias de dicha variable en la región de estudio. El supuesto de isotropía se verifica calculando funciones de covarianza o de semivarianza en diversas direcciones, si éstas son significativamente diferentes para cualquier dirección considerada, podría ser un proceso anisotrópico. Así, de acuerdo con Schabenberger & Gotway (2005), ante presencia de anisotropía, se debe hacer una corrección de la misma, pues de otro modo las conclusiones a las que se llegue en la etapa de predicción serían erradas o alejadas del comportamiento real del proceso espacial.

1.2. Funciones de correlación espacial muestral y estimación de modelos

Retomando lo dicho anteriormente, existen dos etapas en el desarrollo de un análisis espacial de tipo geoestadístico: la primera corresponde a la definición de la estructura de dependencia espacial presente en los datos de estudio, conocido también como *análisis estructural* y la segunda se refiere a la *predicción espacial* vía *kriging*. Se abordará en esta sección lo relacionado con la primera etapa del análisis geoestadístico.

Para desarrollar el análisis estructural del proceso espacial, se emplea la información muestral recolectada y tres funciones: el *semivariograma*, el *covariograma* y el *correlograma*. La estacionariedad del proceso espacial es fundamental en esta etapa y en adelante se considerará implícita, pues de no cumplirsen se puede caer en conclusiones erradas al aplicar estimadores de semivariogramas y modelos de semivariogramas a datos de procesos espaciales no estacionarios. (Schabenberger & Gotway, 2005)

1.2.1. Semivariograma, covariograma y correlograma

Sea el proceso espacial $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$ y considérese

$$\begin{aligned} \gamma(s_i - s_j) &= \frac{1}{2} Var[Z(s_i) - Z(s_j)] \\ &= \frac{1}{2} Var[Z(s_i)] + Var[Z(s_j)] - 2Cov[Z(s_i), Z(s_j)]; \end{aligned} \tag{1.3}$$

$\gamma(s_i - s_j)$ es función de la diferencia de coordenadas entre puntos muestreados únicamente y se conoce como el *semivariograma* de un proceso espacial. El semivariograma corresponde a la mitad del *variograma* del proceso espacial, de este modo, $2\gamma(s_i - s_j)$ define al variograma del proceso.

Bajo estacionariedad de segundo orden, el semivariograma puede ser expresado en términos de la función de covarianza del proceso $C(s_i - s_j) = Cov[Z(s_i), Z(s_j)]$ como sigue

$$\gamma(s_i - s_j) = C(0) - C(s_i - s_j), \tag{1.4}$$

tal como en (1.2). Así, el nombre semivariograma es utilizado indistintamente tanto para hacer referencia a la función $\gamma(s_i - s_j)$, como para referirse a la gráfica de $\gamma(h)$ frente a h. Cuando se trabaja con covarianzas, $C(s_i - s_j)$ es la función de covarianza y la gráfica de $C(h)$ frente a h se conoce como *covariograma*.

Por otro lado, el *correlagrama* de un proceso espacial se define de forma tal que

$$R(s,h) = \frac{C(h)}{\sqrt{Var[Z(s)]Var[Z(s+h)]}};\ R(h) = \underbrace{\frac{C(h)}{C(0)} = \frac{C(h)}{\sigma^2}}_{\text{Bajo estacionariedad de } 2^{do} \text{ orden}};$$

sin embargo, igual que para las anteriores funciones, el término correlograma puede referirse también a la gráfica de $R(h)$ contra h.

Siempre que existan, es posible emplear las tres funciones antes descritas para definir la estructura de dependencia espacial, debe mencionarse la forma en la que éstas se aplican a datos muestreados para una variable regionalizada particular, esto es, se deben presentar los estimadores que permiten el cálculo directo sobre la muestra de las funciones de semivarianza y de covarianza.

De este modo, siguiendo a Matheron (1963) y a Schabenberger & Gotway (2005), como $\gamma(s_i - s_j) = \frac{1}{2}E[(Z(s_i) - Z(s_j))^2]$, la función de semivarianza puede estimarse usando el método de los momentos, a través de lo que se conoce como *semivariograma experimental* y que se calcula de acuerdo a (1.5):

$$\widehat{\gamma}(s_i - s_j) = \frac{1}{2|N(s_i - s_j)|} \sum_{N(s_i - s_j)} \{Z(s_i) - Z(s_j)\}^2, \tag{1.5}$$

donde $N(s_i - s_j)$ corresponde al conjunto de pares de localización con diferencia de coordenadas $s_i - s_j$ (o lo que es lo mismo, distancia entre pares de localización) y $|N(s_i - s_j)|$ es el número de pares distintos en ese conjunto. Así, el correspondiente estimador para la función de covarianza $C(h)$ sería

$$\widehat{C}(s_i - s_j) = \frac{1}{|N(s_i - s_j)|} \sum_{N(s_i - s_j)} (Z(s_i) - \overline{Z})(Z(s_j) - \overline{Z}), \tag{1.6}$$

donde $\overline{Z} = n^{-1}\sum_{i=1}^{n} Z(s_i)$. Puede tomarse $h = s_i - s_j$ y de este modo hacer referencia a la distancia de separación en h unidades del par de localización s_i al par de localización s_j, para todo par muestreado en el dominio de estudio.

En cuanto a las propiedades que cumplen los estimadores con anterioridad presentados, se tiene que $\widehat{\gamma}(h)$ es insesgado para $\gamma(h)$ si $\mathbf{Z}(s)$ es intrínsecamente estacionario. Si la media es estimada desde la muestra, $\widehat{C}(h)$ es un estimador sesgado de la función de covarianza en el rezago h. Adicionalmente,

$$\begin{aligned} \widehat{\gamma}(h) &= \frac{1}{2|N(h)|} \sum_{N(h)} \left\{Z(s_i) - \overline{Z} - Z(s_j) + \overline{Z}\right\}^2 \\ &= \frac{1}{|N(h)|} \sum_{N(h)} \left\{Z(s_i) - \overline{Z}\right\}^2 - \widehat{C}(h), \end{aligned}$$

pero $\widehat{C}(0) = n^{-1}\sum_{i=1}^{n} \left\{Z(s_i) - \overline{Z}\right\}^2$. En consecuencia, $\widehat{C}(0) - \widehat{C}(h) \neq \widehat{\gamma}(h)$ lo cual implica que el estimador del semivariograma construído desde (1.6) también será sesgado, sin embargo, como $|N(h)|/n \to 1$, tal estimador es asintóticamente insesgado. (Schabenberger & Gotway, 2005)

Cualquiera de las tres funciones antes descritas puede emplearse para definir la estructura de dependencia espacial del proceso en estudio, siempre y cuando éste sea estacionario: bajo estacionariedad de segundo orden, es posible estimar tanto el covariograma como el semivariograma; si no se tiene esta propiedad pero existe estacionariedad intrínseca, aún se puede estimar la función de semivarianza. Sin embargo, en la práctica suele utilizarse con mayor frecuencia el semivariograma,

pues este no involucra la estimación de parámetros adicionales, como si sucede en los otros dos casos. (Giraldo, 2011)

La interpretación del semivariograma experimental, así como cualquiera de las otras dos funciones, se basa en la distancia entre pares de localización sobre el dominio de interés, pues de acuerdo con *la primera ley de la geografía* de Tobler: "todo está relacionado con todo lo demás, pero las cosas cercanas están más relacionadas que las cosas distantes" (Tobler, 1970). De este modo, se espera que en presencia de autocorrelación, para valores pequeños de h, el semivariograma experimental tenga magnitudes menores a las que este tomaría cuando las distancias h se incrementan, en otras palabras, las correlaciones disminuyen con el aumento de la separación espacial.

1.2.2. Modelos teóricos de semivariogramas

Para poder pasar a la segunda etapa del análisis espacial geoestadístico, la predicción espacial, es necesario conocer la estructura de autocorrelación asociada con la variable regionalizada en estudio, para cualquier posible distancia entre sitios sobre el dominio. Una primera idea de dicha autocorrelación, se obtiene al graficar para diferentes rezagos el semivariograma experimental, sin embargo, éste sólo brinda una aproximación para algunas distancias, pues debido a irregularidades en el muestreo, en la práctica se toman intervalos de distancia $\{[0, h], (h, 2h], (2h, 3h], \ldots\}$ y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo, más no a una distancia h específica (Giraldo, 2011). Luego, se hace necesario generalizar lo que inicialmente se obtiene con el semivariograma experimental, para cualquier distancia vía modelos teóricos de semivariogramas.

El desarrollo conceptual en torno a modelos teóricos de semivariograma para generalizar lo evidenciado en el semivariograma experimental es bastante amplio. Por ejemplo, en Schabenberger & Gotway (2005) se presentan familias y clases de funciones de covarianza, así como las condiciones que éstas deben cumplir para ser modelos válidos. En Samper & Carrera (1990) también se aborda una discusión en torno a las características y condiciones asociadas con los modelos teóricos de covarianza y semivarianza.

En términos generales, si se tiene una función de covarianza $C(h)$ y el semivariograma $\gamma(h)$ asociados con un proceso estacionario de segundo orden e isotrópico, las siguientes proposiciones son válidas (Schabenberger & Gotway, 2005):

- Si $C(h)$ es válida en $\mathbb{R}^d$, también es válida en $\mathbb{R}^s$ para $s < d$.
- Si $C_1(h)$ y $C_2(h)$ son funciones de covarianza válidas, entonces $aC_1(h) + bC_2(h)$, $a, b \geq 0$, es una función de covarianza válida.
- Si $\gamma_1(h)$ y $\gamma_2(h)$ son semivariogramas válidos, entonces $a\gamma_1(h) + b\gamma_2(h)$, $a, b \geq 0$, es un semivariograma válido.
- Una función de covarianza válida $C(h)$ es una función definida positiva, esto es,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(s_i - s_j) \geq 0,$$

  para cualquier conjunto de números reales $a_1, \ldots, a_n$ y sitios de estudio.
- Un semivariograma válido $\gamma(h)$ es condicionalmente definido negativo, esto es,

$$2\sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \gamma(s_i - s_j) \leq 0,$$

  para cualquier conjunto de números reales $a_1, \ldots, a_k$ tal que $\sum_{i=1}^{k} a_i = 0$ y para cualquier número finito de sitios de estudio.

- Una condición necesaria para que $\gamma(h)$ sea un semivariograma válido es que $2\gamma(h)$ crezca más lentamente que $||h||^2$. Esto se conoce usualmente como la *hipótesis intrínseca* y su cumplimiento garantiza que la varianza de las predicciones sean positivas.

En presencia de modelos válidos de semivariograma, éstos pueden dividerse en *acotados* y *no acotados* (Giraldo, 2011). En los modelos acotados, se garantiza el cumplimiento del supuesto de estacionariedad de segundo orden, pues la covarianza de los incrementos es finita. Además, éstos modelos tienen en común, entre otros, tres parámetros conocidos como *efecto pepita* (*nugget* en inglés), *silla* o *meseta* (*sill* en inglés) y *rango* (*Figura 1.1*).

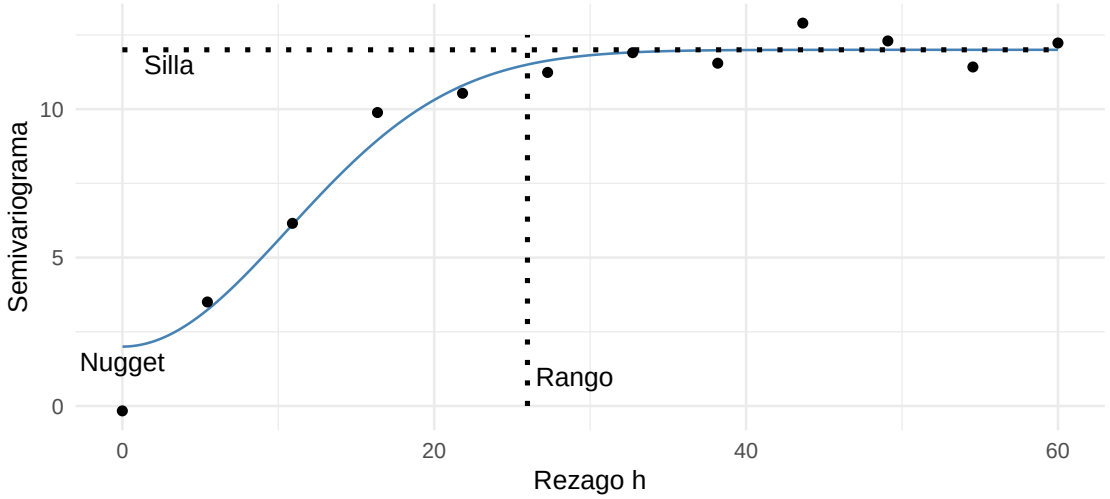


FIGURA 1.1. Representación gráfica de los parámetros usuales un semivariograma experimental acotado y su respectivo modelo teórico de ajuste.

El *nugget*, denotado por c_0, corresponde a una discontinuidad en el semivariograma, la cual hace que éste se desplace del origen al valor c_0, cuando la distancia entre sitios es nula, esto es, $\gamma(h) \to 0$ cuando $h \to 0$. Tal discontinuidad aparece por errores de medición de la variable regionalizada en estudio o por la escala de la misma. Otra causa de la aparición del *nugget*, según comenta Giraldo (2011), puede deberse a que la estructura espacial se concentra a distancias inferiores a las observadas.

Por otro lado, la *silla* del semivariograma representa la cota superior del mismo, corresponde al valor de $C(0) = \sigma^2$ y está presente únicamente en los modelos que cumplen el supuesto de estacionariedad de segundo orden (modelos acotados). Cuando el proceso espacial tiene varianza infinita, la silla también es infinita y en este caso dicho proceso cumpliría sólo la hipótesis intrínseca (modelos no acotados). Usualmente, la silla se denota con c_1 y si el *nugget* no es nulo, la silla quedaría representada por $c_0 + c_1$.

El *rango* de los modelos teóricos de semivariogramas se define como la distancia a la cual las observaciones pierden o dejan de estar correlacionadas. De este modo, los puntos separados por una distancia inferior al rango se consideran espacialmente correlacionados y aquellos separados por una distancia superior al mismo se consideran independientes. Pueden presentarse procesos espaciales en los que su modelo de semivariograma asociado no describa una distancia finita para la cual se hable de independencia entre observaciones, es por esto que suele utilizarse lo que se conoce como *rango práctico* o *rango efectivo* y se define como la distancia a la que el semivariograma alcanza el 95 % de la silla (Schabenberger & Gotway, 2005).

Así, una vez definidos los parámetros usuales de un modelo de semivariograma teórico, en la *Tabla 1.1* se presentan algunos de éstos modelos, especificando $\gamma(h)$ y características básicas asociadas a los mismos. La representación gráfica para tales modelos puede visualizarse en la *Figura 1.2*.

TABLA 1.1. Modelos teóricos de semivariogramas asociados a un proceso espacial isotrópico $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$, bajo los supuestos usuales de estacionariedad.

Modelo	$\gamma(h)$
1. Efecto Nugget. Se ajusta en presencia de un problema de microestructura (la variación espacial ocurre a menores distancias, luego fue incorrecta la distancia de muestreo elegida); o cuando no existe autocorrelación espacial. (*Fig. 1.2(a)*)	$\gamma(h) = \begin{cases} 0 & \text{si } \|\|h\|\| = 0 \\ c_0 & \text{si } \|\|h\|\| \neq 0 \end{cases}$
2. Lineal. Es un caso particular de los modelos monómicos ($\lambda = 1$). Cumple sólo el supuesto de estacionariedad intrínseca, luego el semivariograma no se estabiliza. (*Fig. 1.2(c)*)	$\gamma(h) = \begin{cases} 0 & \text{si } \|\|h\|\| = 0 \\ c_0 + \theta\|\|h\|\| & \text{si } \|\|h\|\| \neq 0 \end{cases}$
3. Esférico. Es acotado, lo cual implica estacionariedad de segundo orden. Presenta un crecimiento acelerado cerca al origen, similar a un modelo lineal, pero a valores superiores al rango tal crecimiento es inexistente. (*Fig. 1.2(b)*)	$\gamma(h) = \begin{cases} 0 & \text{si } h = \alpha \\ c_0 + c_1 \left[\frac{3}{2} \frac{\|\|h\|\|}{\alpha} - \frac{1}{2} \left(\frac{\|\|h\|\|}{\alpha} \right)^3 \right] & \text{si } 0 < \|\|h\|\| \leq \alpha \\ c_0 + c_1 & \text{si } h > \alpha \end{cases}$
4. Exponencial. Aplicado a procesos en los que la dependencia espacial tiene un crecimiento de tipo exponencial en función de la distancia entre observaciones. Es ampliamente utilizado pues la mayoría de procesos espaciales presentan este tipo de crecimiento, además al ser acotado, garantiza covarianza finita entre sitios de estudio. (*Fig. 1.2(b)*)	$\gamma(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_0 + c_1 \left[1 - \exp\left\{ \frac{-3\|\|h\|\|}{\alpha} \right\} \right] & \text{si } \|\|h\|\| > 0 \end{cases}$
5. Gaussiano. Su comportamiento cerca al origen tiene forma parabólica y al igual que el exponencial es ampliamente utilizado. Tanto el modelo gaussiano como el exponencial pertenecen a la *clase Matérn de funciones de covarianza*, luego presentan características similares. (*Fig. 1.2(b)*) (Schabenberger & Gotway, 2005)	$\gamma(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_0 + c_1 \left[1 - \exp\left\{ - \left(\frac{\|\|h\|\|}{\alpha} \right)^2 \right\} \right] & \text{si } \|\|h\|\| > 0 \end{cases}$

TABLA 1.1. Modelos teóricos de semivariogramas asociados a un proceso espacial isotrópico $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$, bajo los supuestos usuales de estacionariedad.

Modelo	$\gamma(h)$
6. Potencia Exponencial. Reportado en la literatura como *powered exponential model*, por su nombre en inglés (Minasny & McBratney, 2005; Gneiting, 2002; Chen, 2013); es una extención del modelo exponencial para potencias ω tales que $0 < \omega \leq 2$. (*Fig. 1.2(b)*)	$\gamma(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_0 + c_1 \left[1 - \exp\left\{-\left(\frac{\|\|h\|\|}{\alpha}\right)^{\omega}\right\}\right] & \text{si } \|\|h\|\| > 0 \\ \text{donde } 0 < \omega \leq 2,\ \alpha > 0. & \end{cases}$;
7. Monómicos. Corresponde al grupo de modelos no acotados, luego éstos no poseen silla. Al cumplir únicamente el supuesto de estacionariedad intrínseca, usarlos es delicado pues tal supuesto podría en realidad no cumplirse para alguna dirección particular. (*Fig. 1.2(c)*) (Giraldo, 2011)	$\gamma(h) = \begin{cases} 0 & \text{si } h, \theta = 0 \\ \theta \|\|h\|\|^{\lambda} & \text{si } \|\|h\|\| \neq 0 \\ \text{donde } 0 \leq \lambda < 2,\ \theta \geq 0. & \end{cases}$;

1.2.3. Tratamiento de procesos anisotrópicos

Tal como mencionan Schabenberger & Gotway (2005), si la función de covarianza de un proceso estacionario de segundo orden es anisotrópica, la estructura espacial depende de la dirección. De este modo, mientras que en el caso isotrópico, los contornos de igual correlación sobre todo el dominio son esféricos, un caso particular de anisotropía da lugar a contornos elípticos. Este caso se conoce como *anisotropía geométrica* y se puede corregir mediante una transformación lineal del sistema de coordenadas (*Figura 1.3*).

Una anisotropía geométrica se manifiesta en semivariogramas que tienen la *misma forma* y *silla* en la dirección de los ejes mayor y menor de la elipse que describen, pero con diferentes *rangos*. Para corregirla, puede considerarse una transformación lineal $\mathbf{s}^* = \mathbf{A}\mathbf{s}$ del espacio Euclidiano que proporcione el espacio apropiado para expresar la covarianza. De este modo, una rotación del sistema de coordenadas para alinear los ejes mayor y menor de los contornos elípticos y una compresión del eje mayor para hacer que los contornos sean esféricos, producirían el proceso isotrópico deseado. (Schabenberger & Gotway, 2005)

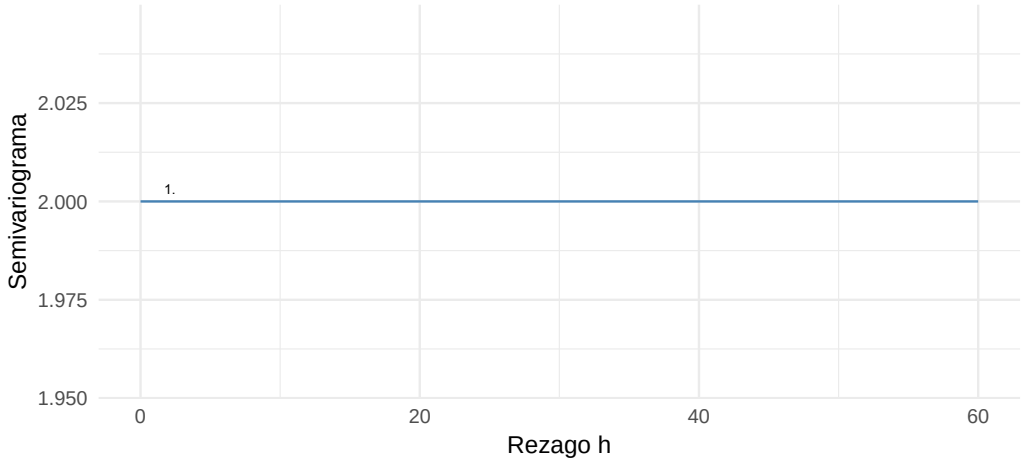
Para tal fin, se toma

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} \cos(\phi) & -\sin(\phi) \\ \sin(\phi) & \cos(\phi) \end{bmatrix},$$

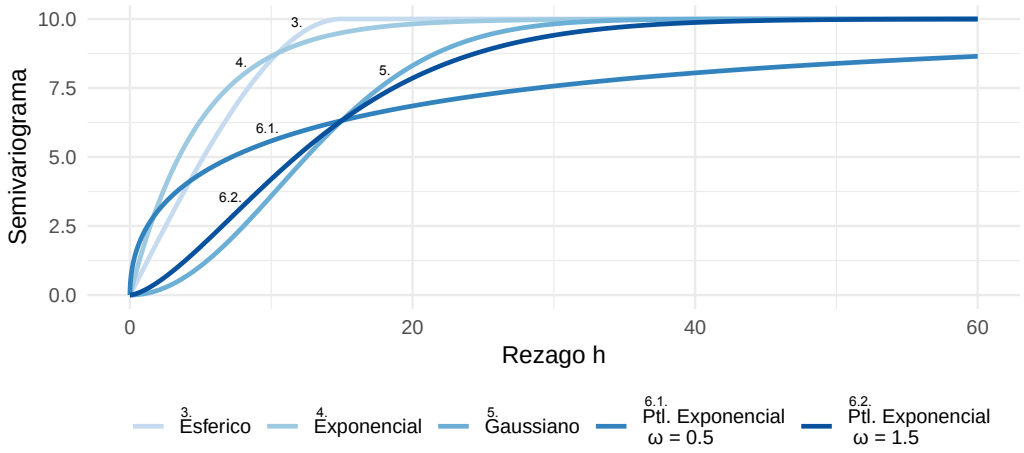
donde λ es la razón de anisotropía, definida como el cociente de los rangos asociados a las direcciones de los ejes mayor y menor de la elipse y ϕ es la dirección del eje mayor.

1.2.4. Estimación de modelos teóricos de semivariogramas

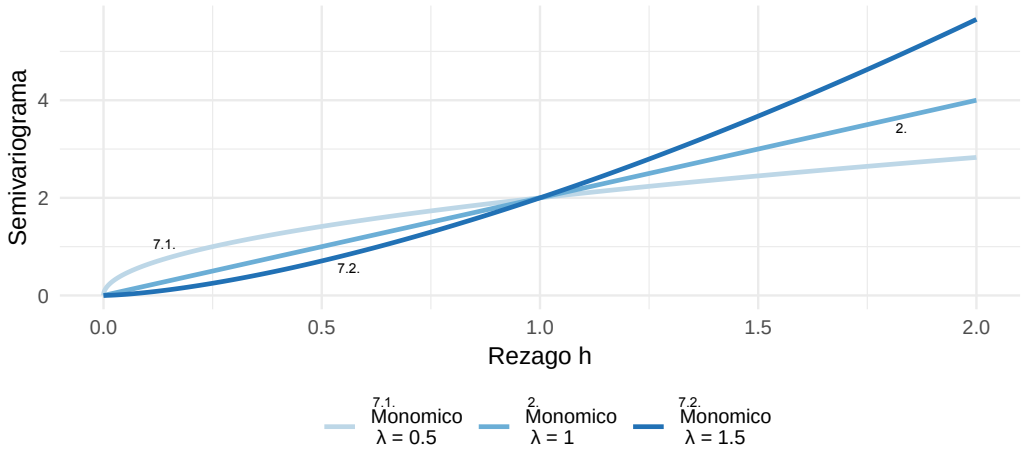
Tal como se vió anteriorme, los modelos teóricos de semivariogramas involucran parámetros en su definición que deben ser estimados para completar la caracterización de los mismos. Los métodos que comúnmente son empleados para realizar tal estimación son, entre otros, *mínimos cuadrados ordinarios* (*OLS*, en inglés), *mínimos cuadrados ponderados* (*WLS*, en inglés), *máxima verosimilitud* (*ML*, en inglés) y *máxima verosimilitud restringida* (*REML*, en inglés). En Schaben-



(a) *Modelo Nugget*.



(b) Modelos de la clase *Matérn* de funciones de semivarianza.



(c) *Modelos Monómicos* de funciones de semivarianza.

FIGURA 1.2. Representación gráfica del comportamiento usual asociado a algunos modelos de semivarianza teóricos válidos.

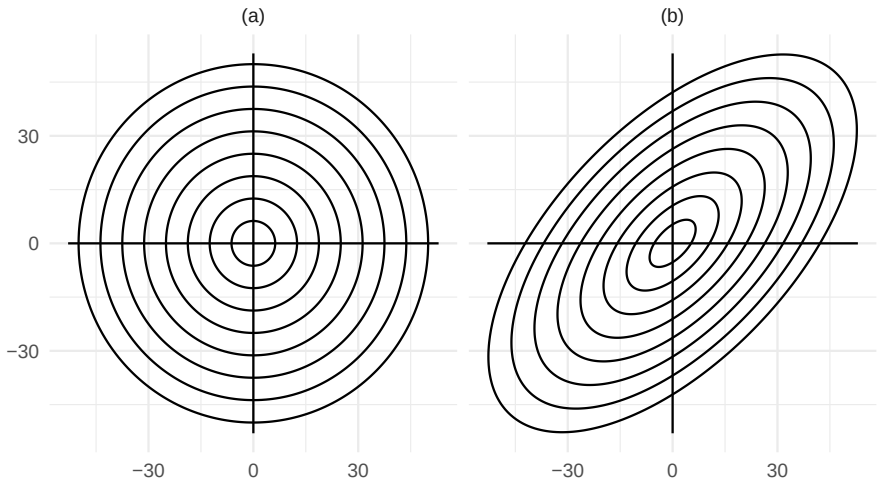


FIGURA 1.3. Gráficos de contornos asociados a dos procesos espaciales. El modelo isotrópico en (a) describe contornos circulares, mientras que el modelo en (b) corresponde a una rotación de 45° de las coordenadas del proceso, obteniendo contornos elípticos característicos de una anisotropía geométrica.

berger & Gotway (2005) se describen cada uno de éstos métodos, mencionando detalles relevantes para su implementación, al igual que en Cressie (1993).

Para aplicar métodos de máxima verosimilitud se requiere que la distribución del vector aleatorio $\mathbf{Z}(s) = [Z(s_1), Z(s_2), \ldots, Z(s_n)]'$ sea conocida; en este caso se asume generalmente como normal multivariada, pues según comentan Schabenberger & Gotway (2005), tales métodos han sido desarrollados sólo para campos aleatorios Gaussianos. De este modo, $\mathbf{Z}(s) \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\theta))$, donde $\boldsymbol{\Sigma}(\theta) = Cov(\mathbf{Z}(s)) \in \mathcal{M}_{(n\times n)}$ (θ es el vector que contiene los parámetros del modelo a estimar) y $\mathbf{X} \in \mathcal{M}_{(n\times q)}$ con $q < n$, la cual contiene variables explicativas que usualmente corresponden a las coordenadas de ubicación de los sitios en estudio. La componente ij de $\boldsymbol{\Sigma}(\theta)$ corresponde a la covariaza espacial entre $Z(s_i)$ y $Z(s_j)$, es decir, $C(s_i - s_j; \theta) = C(h; \theta)$. De este modo, se minimiza (1.7), el negativo de la función de log-verosimilitud, con respecto a $\boldsymbol{\beta}$ y a θ y se obtiene el estimador de (1.8), cuando θ es conocido. $\widehat{\boldsymbol{\beta}}_{GLS}$ es nombrado en la literatura como el estimador de mínimos cuadrados generalizados (*GLS*, por su siglas en inglés).

$$2L(\boldsymbol{\beta}; \theta; \mathbf{Z}(s)) = \ln(|\boldsymbol{\Sigma}(\theta)|) + n\ln(2\pi) + (\mathbf{Z}(s) - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\theta)(\mathbf{Z}(s) - \mathbf{X}\boldsymbol{\beta}) \quad (1.7)$$

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\theta)\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\theta)\mathbf{Z}(s) \quad (1.8)$$

Sin embargo, como el interés recae en explicitar $\widehat{\theta}$, pues contiene los parámetros asociados al modelo de covariograma que ahora nos ocupa, el procedimiento a seguir consiste en reemplazar (1.8) en (1.7), obteniendo una función que depende únicamente de θ. Así, $\widehat{\boldsymbol{\beta}}_{GLS}$ se obtiene desde una función objetivo (el negativo de la función log-verosímil) y la función resultante al remplazar (1.8) en (1.7) corresponde a un perfil del negativo de la función log-verosímil. Esta función se minimiza, por lo general, mediante métodos numéricos iterativos, esto debido a que la log-verosimilitud generalmente es una función no lineal de los parámetros de covariación (θ). Una vez que se han obtenido las estimaciones de máxima verosimilitud $\widehat{\theta}_{ML}$, su valor se sustituye en la función perfil antes mencionada para obtener sus estimaciones de verosimilitud. (Schabenberger & Gotway, 2005)

El estimador $\widehat{\theta}_{ML}$ es sesgado pero asintóticamente eficiente (Schabenberger & Gotway, 2005). Sin embargo, tener una muestra grande implica realizar una gran cantidad de operaciones para computar $\widehat{\theta}_{ML}$, pues deben obtenerse de forma explícita tanto el determinante como la inversa de la matriz de covarianza en forma iterativa. Una variación del estimador máximo verosímil que

reduce el sesgo de las estimaciones y en ciertas situaciones balanceadas lo elimina (Patterson & Thompson, 1971), es el estimador obtenido vía máxima verosimilitud restringida (*REML*); éste sustituye la maximización de la verosimilitud del vector $\mathbf{Z}(s)$ por la del vector $\mathbf{KZ}(s)$, donde la matriz $\mathbf{K} \in \mathcal{M}_{[n\times(n-p)]}$ es elegida tal que $E[\mathbf{KZ}(s)] = \mathbf{0}$.

De lo anterior, $\mathbf{KZ}(s) \sim \mathbf{N}_n(\mathbf{0}, \mathbf{K\Sigma}(\theta)\mathbf{K}')$ y los estimadores por máxima verosimilitud restringida son todos los valores de θ que minimizan (1.9). Tal como se observa, (1.9) no depende de $\boldsymbol{\beta}$, luego este método de estimación no modela la tendencia, sino que opera directamente sobre el vector de incrementos $\mathbf{KZ}(s)$, el cual posee media $\mathbf{0}$. Ahora, aún cuando (1.9) no dependa de $\boldsymbol{\beta}$, es posible definir $\widehat{\boldsymbol{\beta}}_{REML}$, reemplazando el $\widehat{\theta}_{REML}$ obtenido de la minimización de (1.9) en (1.8), llegando a lo presentado en (1.10).

$$2L(\theta; \mathbf{KZ}(s)) = \ln(|\mathbf{K\Sigma}(\theta)\mathbf{K}'|) + (n-p)\ln(2\pi) + \mathbf{Z}(s)'\mathbf{K}'(\mathbf{K\Sigma}(\theta)\mathbf{K}')^{-1}\mathbf{KZ}(s) \quad (1.9)$$

$$\widehat{\boldsymbol{\beta}}_{REML} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\widehat{\theta}_{REML})\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}(\widehat{\theta}_{REML})\mathbf{Z}(s) \quad (1.10)$$

Si bien el método de estimación *REML* minimiza el sesgo de $\widehat{\theta}$, sigue siendo significatimente alto el número de operaciones que deben computarse para obtenerlo. Sin embargo, pueden emplearse s*oftwares estadísticos* que cuenten con un módulo de modelamiento geoestadístico, de forma que puedan obtenerse tales estimaciones, por ejemplo, *R* cuenta con varios paquetes para modelamiento de variables regionalizadas, o *GS+* (Design, 1995) que trae incorporados procedimientos iterativos como el de *Gauss-Newton* para llevar a cabo la estimación. (Giraldo, 2011)

Otro método de estimación comúnmente utilizado para el modelamiento geoestadístico es el de *mínimos cuadrados ponderados* (*WLS*, en inglés), que se obtiene de una generalización del método *OLS*. El procedimiento de estimación *OLS* consiste en minimizar la suma de cuadrados en (1.11), la cual se encuentra asociada al modelo espacial de la forma $\widehat{\gamma}(h) = \gamma(h,\theta) + \mathbf{e}(h)$, donde $\gamma(h,\theta) = [\gamma(h_1,\theta), \ldots, \gamma(h_k,\theta)]'$ y k, es el número de rezagos considerados; además, se tiene que $Var[\mathbf{e}(h)] = \mathbf{E}(\theta)$.

$$(\widehat{\gamma}(h) - \gamma(h,\theta))'\mathbf{E}^{-1}(\theta)(\widehat{\gamma}(h) - \gamma(h,\theta)) \quad (1.11)$$

Uno de los enfoques de *WLS* para el ajuste del semivariograma reemplaza a $\mathbf{E}(\theta)$ por una matriz diagonal de ponderaciones $\mathbf{W}(\theta)$, cuyas entradas, al rezago m, están dadas por (1.12) (Cressie, 1993). De este modo, en lugar de la suma generalizada de cuadrados en (1.11), este enfoque minimiza la suma de cuadrados ponderados, al rezago m, presentada en (1.13).

$$Var[\widehat{\gamma}(h_m)] \approx 2\frac{\gamma(h_m,\theta)^2}{|N(h_m)|} \quad (1.12)$$

$$(\widehat{\gamma}(h) - \gamma(h,\theta))'\mathbf{W}^{-1}(\theta)(\widehat{\gamma}(h) - \gamma(h,\theta)) = \sum_{m=1}^{k} \frac{|N(h_m)|}{2\gamma(h_m,\theta)^2}\left\{\widehat{\gamma}(h_m) - \gamma(h_m,\theta)\right\}^2 \quad (1.13)$$

En la práctica, suelen utilizarse rezagos espaciales hasta alcanzar la mitad de la máxima distancia entre cualquier par de ubicaciones, pues para localizaciones muy separadas, la cantidad de puntos incluídos en la estimación del semivariograma disminuye significativamente. (Bohorquez, 2009)

En términos generales, el método *WLS* estima mejor que los métodos basados en la verosimilitud y además, dado que (1.13) puede escribirse como una suma de cuadrados ponderados sobre las clases de rezago de orden k, es muy sencillo ajustar un modelo de semivariograma con un paquete (*software*) de estadísticas no lineales, siempre que se pueda acomodar en el mismo las ponderaciones pertinentes. (Schabenberger & Gotway, 2005)

1.3. Predicción espacial y kriging

Cuando la etapa de la definición de la estructura de dependencia espacial ha sido finalizada, esto es, cuando ya se han verificado los supuestos de estacionariedad e isotropía del proceso y

además se tiene un modelo de semivariograma teórico ajustado, que generaliza el comportamiento de la estructura de correlación espacial para cualquier distancia, puede proseguirse con la segunda etapa conocida como *predicción espacial*. La predicción espacial en procesos geoestadísticos busca predecir el valor que toma cierta variable regionalizada, $\mathbf{Z}(s)$, en una localización $s_0 \in D$ que no ha sido muestreada, a partir de lo observado en las n localizaciones muestreadas; lo anterior se realiza bajo un procedimiento conocido como *kriging* y que será abordado en los siguientes subsecciones. (Bohorquez, 2009; Schabenberger & Gotway, 2005)

En estadística suelen utilizarse los términos de *predicción* y *estimación*, en ocasiones, indistintamente. Sin embargo, es importante aclarar su diferencia, pues *estimar* involucra el proceso de inferir sobre parámetros fijos pero desconocidos y *predecir* hace alusión a la inferencia sobre cantidades aleatorias (Giraldo, 2002). De este modo, en el contexto geoestadístico y dado el proceso aleatorio $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^d\}$, se puede pensar tanto en estimación como en predicción, pues si dicho proceso sigue, entre otros, el modelo

$$\mathbf{Z}(s) = \mathbf{X}(s)\boldsymbol{\beta} + \mathbf{e}(s),\ \mathbf{e}(s) \sim (\mathbf{0}, \boldsymbol{\Sigma});$$

el interés puede recaer en estimar $E[\mathbf{Z}(s)] = \mathbf{X}(s)\boldsymbol{\beta}$ o en predecir $\mathbf{Z}(s)$. Sin embargo, tal como menciona Schabenberger & Gotway (2005), en las aplicaciones geoestadísticas, la predicción es a menudo más importante que la estimación de la media. Lo anterior se debe a que $E(\mathbf{Z}(s))$, al calcularse con respecto a la distribución de las posibles realizaciones de la variable regionalizada en las ubicaciones s, podría no ser de mayor utilidad. Además, es frecuente estar interesado en la cantidad real $\mathbf{Z}(s)$ que está allí y no en una cantidad promedio conceptual.

De este modo, se han desarrollado metodologías que usan propiedades estadísticas de los datos para predecir $Z(s_0)$, usando la información de lugares muestreados, $\mathbf{Z}(s)$, de forma que se logre minimizar el error cuadrático medio de predicción (*MSPE*, por sus siglas en inglés). A dichas metodologías se les conoce como *predicción óptima* y son comúnmente referenciadas como *kriging*. El kriging debe su nombre al Ingeniero de Minas sudafricano D.G. Krige, cuyo trabajo en las minas de oro de Witwatersrand sentó las bases para el campo de la geoestadística. (Krige, 1951; Matheron, 1963)

Tal como se dijo, el objetivo del kriging consiste en lograr una predicción óptima del proceso geoestadístico para una localización $s_0 \in D$ no observada, teniendo en cuenta las que si se observaron y de forma que se minimice el *MSPE*. Así, la derivación del predictor, notado como $p(\mathbf{Z}; s_0)$[1], involucra la minimización del

$$E\left[(Z(s_0) - p(\mathbf{Z}; s_0))^2\right] = MSE[p(\mathbf{Z}; s_0); Z(s_0)].$$

Los predictores a derivar deben cumplir las condiciones de *linealidad*, *insesgamiento* y *varianza mínima*. La condición de linealidad establece que el predictor debe ser una combinación lineal de los valores de la variable en sitios muestreados, de acuerdo a un vector de ponderaciones $\boldsymbol{\lambda}$, el cual se obtiene de forma que se cumplan las dos condiciones finales, esto es, insesgamiento y varianza mínima. El insesgamiento asegura que el valor esperado de predictor debe coincidir con el valor esperado de la variable en el punto de predicción, es decir, $E[p(\mathbf{Z}; s_0)] = E[Z(s_0)]$. Finalmente, la condición de varianza mínima pide al predictor minimizar $Var[p(\mathbf{Z}; s_0) - Z(s_0)] = MSE[p(\mathbf{Z}; s_0); Z(s_0)]$. (Bohorquez, 2009; Giraldo, 2011)

De este modo, debe encontrarse $\boldsymbol{\lambda}$ minimizando el *MSE*; al minimizar el *MSE* se obtendrán ecuaciones, las cuales cambiarán según sean los supuestos sobre la media y la covarianza del proceso espacial en estudio, generando así predictores diferentes y con éstos diferentes tipos de kriging. (Bohorquez, 2009)

[1] $p(\mathbf{Z}; s_0)$ es el predictor de $Z(s_0)$ en la localización s_0, basado en el vector de localizaciones observadas $\mathbf{Z}(\mathbf{s})$. (Schabenberger & Gotway, 2005)

CAPÍTULO 2

Modelo lineal de efectos mixtos

Los modelos lineales de efectos mixtos (*LMM*, en inglés) son populares en el análisis de datos longitudinales pues permiten el modelamiento y análisis explícito tanto de la variación entre unidades como la variación dentro de las unidades. De este modo, los efectos aleatorios varían entre los sujetos e inducen la dependencia dentro del sujeto para las medidas repetidas tomadas sobre el mismo, después del condicionamiento en las covariables observadas. (Fitzmaurice et al., 2008)

Así, los modelos lineales mixtos para medidas repetidas corresponden a modelos de regresión de respuesta univariada con errores correlacionados. Una ventaja importante de esta metodología es que se acomoda a la estructura y a las complejidades de los datos longitudinales, tal como antes se mencionaba. Mediante el enfoque de los modelo mixtos se puede dar más importancia a consideraciones específicas sobre la unidad estudiada, que a la metodología estadística misma. (Davis, 2002)

Por ejemplo, Diggle et al. (2002) sugieren pensar en una regresión lineal simple para el crecimiento infantil donde el intercepto representa el peso al nacer y la pendiente corresponde a la tasa de crecimiento. Obviamente, los niños nacen con diferentes pesos y tienen diferentes tasas de crecimiento debido a factores genéticos y ambientales que son difíciles o imposibles de cuantificar, lo que hace pensar en el intercepto y en la pendiente como variables aleatorias. Un modelo de efectos aleatorios es una descripción razonable si el conjunto de coeficientes de una población (intercepto, pendiente) de niños puede considerarse como una muestra de una distribución. Dados los coeficientes reales para un niño, el modelo lineal de efectos aleatorios supone además que las observaciones repetidas para esa persona son independientes. La correlación entre las observaciones repetidas surge porque no es posible observar la curva de crecimiento subyacente, es decir, los verdaderos coeficientes de regresión, pues sólo se tienen mediciones inexactas del peso en cada niño.

Se presenta entonces la formulación teórica del modelo en la sección 2.1, para continuar con su estimación en la sección 2.2. Al final del capítulo se realiza una aplicación del *LMM* a datos relacionados con la calidad del aire en España.

2.1. Formulación del modelo

La formulación teórica de los *LMM*, se hará siguiendo a Davis (2002) y a Fitzmaurice et al. (2008). Supóngase que n sujetos se miden repetidamente a lo largo del tiempo. Sea Y_{ij} la variable respuesta para el *i-ésimo* sujeto en la *j-ésima* ocasión de medición ($i = 1, \ldots, n; j = 1, \ldots, n_i$). Así, el *LMM* puede escribirse como

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i, \tag{2.1}$$

donde $\mathbf{Y}_i$ es un vector n_i-dimensional de respuestas Y_{ij} por unidad i, $\mathbf{X}_i$ es una matriz $n_i \times p$ de covariables con efectos fijos $\boldsymbol{\beta}_{(p\times 1)}$, $\mathbf{Z}_i$ es una matriz $n_i \times q$ de covariables con efectos aleatorios $\mathbf{b}_{i(q\times 1)}$ y $\mathbf{e}_i$ es un vector de errores, n_i-dimensional.

De acuerdo con lo anterior, en general no se asume que los sujetos tengan el mismo número de medidas repetidas o que se midan en un conjunto común de ocasiones. Para acomodar dichos datos longitudinales *desbalanceados* (es decir, mediciones repetidas que no se obtienen en un conjunto común de ocasiones), se asume que hay n_i mediciones repetidas de la respuesta en el i-ésimo sujeto y que cada Y_{ij} se observa en el momento j.

Es posible agrupar las respuestas en un vector de dimensión $n_i \times 1$ denotado por $\mathbf{Y}_i$. Asociado a cada respuesta Y_{ij}, hay un vector $p \times 1$ de covariables, $\mathbf{x}_{ij}$; éste puede incluir covariables cuyos valores no cambian a lo largo de la duración del estudio y covariables cuyos valores cambian con el tiempo. Las primeras se denominan covariables estacionarias en el tiempo o entre sujetos (p.e., sexo, tratamientos o intervenciones experimentales fijas), mientras que las segundas se conocen como covariables que varían en el tiempo o dentro de los sujetos (p.e., el tiempo desde la línea de base, el hábito de fumar actual y exposiciones ambientales que pueden variar con el tiempo). Los vectores de covariables pueden disponerse en una matriz $n_i \times p$ de covariables denotada por $\mathbf{X}_i$.

De esta forma, para una ocasión j dada, el modelo lineal mixto se puede expresar como

$$Y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i + e_{ij},$$

donde $\mathbf{x}'_{ij}$ y $\mathbf{z}'_{ij}$ corresponden a la *j-ésima* fila de las matrices $\mathbf{X}_i$ y $\mathbf{Z}_i$, respectivamente.

Se permite que los vectores de covariables $\mathbf{x}_{ij}$ y $\mathbf{z}_{ij}$ sean aleatorios, ya que generalmente no es razonable tratarlos como fijos en entornos no experimentales. Así, se supone que las covariables son estrictamente exógenas (Chamberlain, 1984) en el sentido de que

$$E(e_{ij}|\mathbf{b}_i, \mathbf{X}_i, \mathbf{Z}_i) = E(e_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = 0 \text{ y } E(\mathbf{b}_i|\mathbf{X}_i, \mathbf{Z}_i) = E(\mathbf{b}_i) = \mathbf{0}.$$

De este modo, la esperanza condicional de la respuesta, dados los efectos aleatorios y las covariables, es

$$\mu_{ij} \equiv E(Y_{ij}|\mathbf{b}_i, \mathbf{X}_i, \mathbf{Z}_i) = E(Y_{ij}|\mathbf{b}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i;$$

por lo tanto, una vez se ha condicionado respecto a $\mathbf{b}_i$, $\mathbf{x}_{ij}$ y $\mathbf{z}_{ij}$, no hay efectos de $\mathbf{x}_{ij'}$ y $\mathbf{z}_{ij'}$ en Y_{ij} para $j' \neq j$.

Se supone entonces que los efectos aleatorios siguen una distribución normal multivariada $\mathbf{b}_i \sim \mathbf{N_q}(\mathbf{0}, \mathbf{G})$, al igual que los errores $\mathbf{e}_i \sim \mathbf{N_{n_i}}(\mathbf{0}, \mathbf{R}_i)$; además los $\mathbf{b}_i$ y los $\mathbf{e}_i$ son independientes a través de los sujetos dadas las covariables e independientes entre sí, esto es, $\mathbf{e}_i$ es independiente de $\mathbf{b}_i$. Así, $\mathbf{R}_i$ es una matriz de covarianza, $\mathbf{R}_i \in \mathcal{M}_{(n_i\times n_i)}$, que caracteriza la varianza y la correlación debida a fuentes dentro de la unidad y $\mathbf{G}$ es una matriz de covarianza, $\mathbf{G} \in \mathcal{M}_{(q\times q)}$, que caracteriza la variación debida a las fuentes entre unidades. (Davidian, 2019; Fitzmaurice et al., 2008)

Lo anterior implica que los $\mathbf{Y}_i$ son vectores aleatorios de dimensión n_i con una estructura particular de matriz de covarianzas, $\mathbf{V}_i$ y además siguen el modelo marginal normal multivariado $\mathbf{N_{n_i}}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i)$, donde $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}\mathbf{Z}'_i + \mathbf{R}_i$. Nótese que las matrices $\mathbf{X}_i$, $\mathbf{Z}_i$ y $\mathbf{R}_i$ son específicas para cada sujeto, lo que hace de (2.1) un modelo bastante general, pues los sujetos pueden tener un número variable de observaciones y además los tiempos de observación pueden diferir entre los mismos. La matriz de covarianza dentro del sujeto, $\mathbf{R}_i$, depende de i sólo a través de su dimensión n_i, es decir, cualquier parámetro desconocido en $\mathbf{R}_i$ no depende de i. Así, se puede considerar una amplia variedad de estructuras de covarianza para $\mathbf{b}_i$ y para $\mathbf{e}_i$.

La opción más común para $\mathbf{R}_i$, en el contexto de datos longitudinales, es el modelo que contempla igualdad de varianza en todos los instantes de medición, para todas las unidades y que asume independencia entre los $\mathbf{Y}_i$, esto es, $\mathbf{R}_i = \sigma^2\mathbf{I_{n_i}}$; sin embargo, también es posible considerar otras estructuras de covarianza para $\mathbf{e}_i$. En el caso de $\mathbf{G}$, es posible permitir que ésta tenga una forma particular o que se comporte de acuerdo a una estructura de covarianza *no estructurada*. (Davidian, 2019)

Algunas de éstas estructuras de covarianza fueron estudiadas por Jennrich & Schluchter (1986), para quienes el *LMM* de (2.1) resulta ser un caso especial, al considerar $\mathbf{\Sigma}_i = \mathbf{V}_i$, del modelo $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i$, donde $e_1, \ldots, e_n$ son independientes y siguen una distribución $\mathbf{N_{n_i}}(\mathbf{0}, \mathbf{\Sigma}_i)$; se asume que la matriz de covarianza $\mathbf{\Sigma}_i$, submatriz de $\mathbf{\Sigma} \in \mathcal{M}_{(n\times n)}$ (se obtiene un número fijo n de mediciones de cada sujeto, pero no se observan todas las respuestas), es una función de un vector $\boldsymbol{\theta}$ de m parámetros de covarianza desconocidos. Siguiendo ésta idea, si el interés particular recae en conocer la estructura de correlación de $\mathbf{Y}_i$ en función de $\boldsymbol{\theta}$, en la *Tabla 2.1* se presentan algunas de las opciones más comunes para dicha estructura, considerando un modelo de datos incompletos con un máximo de n ocasiones de medición.

TABLA 2.1. Algunas estructuras de matrices de covarianza en el modelo de datos incompletos presentado por Jennrich & Schluchter (1986).

Estructura	m	Descripción
No estructurada	$\frac{n(n+1)}{2}$	$\sigma_{ij} = \sigma_{ji}$
Toeplitz	n	$\sigma_{ij} = \theta_k$; $k = \lvert i-j\rvert + 1$
Autoregresiva de orden uno - *AR(1)*	2	$\sigma_{ij} = \sigma^2\rho^{\lvert i-j\rvert}$
Simétrica compuesta	2	$\mathbf{\Sigma} = \sigma^2\mathbf{I_n} + \sigma_b^2\mathbf{1_n}\mathbf{1_n}'$
Observaciones independientes	1	$\mathbf{\Sigma} = \sigma^2\mathbf{I_n}$
Estructura de Markov	2	$\mathbf{\Sigma} = \sigma^2\mathbf{I_n} + \sigma^2\rho^{\lvert t_j - t_{j'}\rvert}$; $j, j' = 1, \ldots, n$

Los *softwares* estadísticos como *R* o *SAS* (SAS-Institute, 1999) son de gran utilidad al estimar $\mathbf{V}_i$, pues contemplan una gran variedad de estructuras de correlación, dentro de las cuales puede escogerse la que mejor explique la estructura de dependencia, de acuerdo a algún criterio de selección de modelos (Davis, 2002). Tales *softwares* utilizan la teoría que será presentada en 2.2 y métodos iterativos para llegar a las estimaciones de $\boldsymbol{\theta}$, el vector cuyas entradas corresponden a los componentes de varianza de $\mathbf{G}$ y $\mathbf{R}_i$, tal cómo se verá en la siguiente subsección.

2.2. Estimación del modelo

En aras de hacer claridad en la notación empleada, considérese inicialmente que los elementos de $\mathbf{G}$ y $\mathbf{R}_i$ son funciones conocidas de un vector de parámetros $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$. El espacio de parámetros para el modelo (2.1) es considerado como el conjunto $\{(\boldsymbol{\beta}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbf{\Omega}\}$, donde $\mathbf{\Omega}$ es el conjunto de valores $\boldsymbol{\theta}$ para los cuales $Var(\mathbf{Y}_i) = \mathbf{V}_i$ es definida positiva, es decir, una auténtica matriz de covarianzas. Cuando $\mathbf{R}_i = \sigma^2\mathbf{I_{n_i}}$ y $\mathbf{Z}_i = 0$, el modelo mixto en (2.1) se reduce al modelo lineal usual $\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{e}_i$. (Davis, 2002)

De este modo, una vez que se ha postulado el modelo, el problema es estimar los parámetros que lo definen. El propósito de toda estimación de parámetros consiste en encontrar, desde los datos observados, el valor que *mejor explique* a los mismos. Los métodos de *máxima verosimilitud* y *máxima verosimilitud restringuida* (*REML*, en inglés) pueden ser usados para estimar los parámetros que caracterizan la *media* o la parte sistemática del modelo, $\boldsymbol{\beta}$, y aquellos que caracterizan la *variación* o parte aleatoria del modelo, es decir, los diferentes parámetros que conforman las matrices $\mathbf{G}$ y $\mathbf{R}_i$.

Lo anterior involucra la "estimación" del vector de efectos aleatorios $\mathbf{b}_i$; las comillas sobre la acción de *estimar* se usan en el sentido de que técnicamente $\mathbf{b}_i$ no es una constante fija como $\boldsymbol{\beta}$, más bien, es un efecto aleatorio que varía entre las unidades. En este sentido, cuando se busca "estimar" $\mathbf{b}_i$, se quiere caracterizar una cantidad aleatoria, no una fija. En situaciones en las que el interés se centra en caracterizar una cantidad aleatoria, es habitual utilizar una terminología diferente para

preservar la idea de que se está interesado en algo que varía. Así, la "estimación"de una cantidad aleatoria a menudo se denomina *predicción* para enfatizar el hecho de que se está tratando de obtener algo que no es fijo e inmutable, pero si algo cuyo valor surge de manera aleatoria (a través de, por ejemplo, el hecho que las unidades se seleccionan al azar de la población). Por lo tanto, para caracterizar el comportamiento de una unidad individual, se debe desarrollar un método para la predicción de $\mathbf{b}_i$. (Davidian, 2019)

Siguiendo a Henderson (1953) y a Davidian (2019), quienes desarrollan métodos para la obtención del estimador y el predictor para $\boldsymbol{\beta}$ y $\mathbf{b}_i$, respectivamente, los vectores de efectos fijos y efectos aleatorios son estimados a partir del *mejor estimador lineal insesgado* (*BLUE*, en inglés) $\widehat{\boldsymbol{\beta}}$ y el mejor predictor lineal insesgado (*BLUP*) $\widehat{\mathbf{b}}_i$, dados por:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}_i'\widehat{\mathbf{V}}_i^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\widehat{\mathbf{V}}_i^{-1}\mathbf{Y}_i = \mathbf{A}_i\mathbf{X}_i'\widehat{\mathbf{V}}_i^{-1}\mathbf{Y}_i, \tag{2.2}$$

$$\widehat{\mathbf{b}}_i = \widehat{\mathbf{G}}\mathbf{Z}_i'\widehat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}), \tag{2.3}$$

donde $\mathbf{A}_i = (\mathbf{X}_i'\widehat{\mathbf{V}}_i^{-1}\mathbf{X}_i)^{-1}$. Davidian (2019) señala además que si se está interesado en caracterizar trayectorias individuales para cada sujeto en estudio, esto es, describir el comportamiento medio del *i-ésimo* sujeto a través de los tiempos de medición considerados, es usual emplear el predictor *BLUP* de (2.4) para este propósito.

$$\begin{aligned} \widehat{E(\mathbf{Y}_i|\mathbf{b}_i)} &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{b}}_i \\ &= \mathbf{X}_i\widehat{\boldsymbol{\beta}} + \mathbf{Z}_i\widehat{\mathbf{G}}\mathbf{Z}_i'\widehat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) \end{aligned} \tag{2.4}$$

El predictor (2.4) se interpreta como un promedio ponderado que combina informacion del sujeto i únicamente e información de la población. En otras palabras, tal predictor considera la información (desde los datos) específica para el sujeto y la información (desde los datos) de la población de donde procede tal sujeto.

Así, de (2.2) y (2.3) se sigue que las estimaciones de $\boldsymbol{\beta}$ y $\mathbf{b}_i$ están sujetas a la estimación de $\boldsymbol{\theta}$, para lo cual se usará el método de máxima verosimilitud restringida abordado en la siguiente subsección.

Estimación vía máxima verosimilitud restringida (*REML*)

La estimación vía máxima verosimilitud (*ML*) de (2.1) conduce a estimadores sesgados de los componentes de varianza alojados en $\boldsymbol{\theta}$, pues éstos dependen de una estimación previa de los efectos fijos del modelo (Davis, 2002). Para evitar estimadores sesgados de los componentes de varianza, Patterson & Thompson (1971) proponen el estimador de máxima verosimilitud restringida (*REML*, por sus siglas en inglés).

En esta aproximación, el vector de efectos fijos es eliminado de la funcion de verosimilitud, y por lo tanto se le denomina *verosimilitud restringida*, la cual sirve para estimar los parámetros de covarianza, o lo que es lo mismo, estimar $\boldsymbol{\theta}$. Cuando los datos son balanceados, este método produce estimadores insesgados. Así, la idea general del método será expuesta a continuación, considerando un modelo mixto balanceado como el de (2.5), por ser de interés para la presente aplicación, donde $N = nt$ y t corresponde al número común de tiempos de medición.

$$\mathbf{Y}_{(N\times 1)} = \mathbf{X}_{(N\times p)}\boldsymbol{\beta}_{(p\times 1)} + \mathbf{Z}_{(N\times q)}\mathbf{b}_{(q\times 1)} + \mathbf{e}_{(N\times 1)}. \tag{2.5}$$

Tal como menciona Davis (2002), la estimación *REML* aplica técnicas de estimación *ML* a la función de verosimilitud asociada con un conjunto de *contrastes de error*, en lugar de la asociada con las observaciones originales. Un contraste de error es una combinación lineal $\mathbf{w}'\mathbf{Y}$ de los elementos de $\mathbf{Y}$, tal que $E(\mathbf{w}'\mathbf{Y}) = 0$ para cualquier $\boldsymbol{\beta}$ (es decir, si $\mathbf{w}'\mathbf{X} = \mathbf{0}_p'$).

Considérese, por ejemplo, $\mathbf{S} = \mathbf{I}_N - \mathbf{P}_\mathbf{X}$, donde $\mathbf{P}_\mathbf{X} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es la matriz de proyección ortogonal en el espacio columna de $\mathbf{X}$. El valor esperado de $\mathbf{SY}$ es

$$E(\mathbf{SY}) = (\mathbf{I}_N - \mathbf{P}_\mathbf{X})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0}_N.$$

De este modo, cada elemento de $\mathbf{SY}$ es un error de contraste. Sin embargo, de la teoría de modelos lineales se sabe que aunque $\mathbf{S} \in \mathcal{M}_{(N\times N)}$, su rango es $N - p$; luego, existen algunas redundancias entre los elementos de $\mathbf{SY}$. Surge entonces la inquietud acerca de cuántos contrastes de error esencialmente diferentes pueden incluirse en un solo conjunto de los mismos. Como solución a tal duda, Davis (2002) comenta que puede demostrarse que cualquier conjunto de contrastes de error contiene a lo sumo $N - p$ contrastes de error linealmente independientes, así los contrastes de error $\mathbf{w}_1'\mathbf{Y}, \ldots, \mathbf{w}_k'\mathbf{Y}$ son linealmente independientes, si los vectores $\mathbf{w}_1, \ldots, \mathbf{w}_k$ también lo son.

Sea $\mathbf{W}$ una matriz, $\mathbf{W} \in \mathcal{M}_{\{N\times(N-p)\}}$, tal que $\mathbf{W}'\mathbf{W} = \mathbf{I}_{(N-p)}$ y $\mathbf{W}\mathbf{W}' = \mathbf{I}_N - \mathbf{P}_\mathbf{X}$. Se puede mostrar que $\boldsymbol{\tau} = \mathbf{W}'\mathbf{Y}$ es un vector de $N - p$ contrastes de error linealmente independientes, sin embargo, no es el único vector de este tipo. De este modo, tal como se había mencionado, el enfoque *REML* aplica técnicas de *ML* a $\boldsymbol{\tau} = \mathbf{W}'\mathbf{Y}$ en lugar de a $\mathbf{Y}$. Bajo el modelo asumido en (2.5), $\mathbf{Y} \sim \mathbf{N}_N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$, con $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$; lo anterior lleva a que $\boldsymbol{\tau} \sim \mathbf{N}_{(N-p)}(\mathbf{0}_{(N-p)}, \mathbf{W}'\mathbf{V}\mathbf{W})$.

Es natural pensar si el estimador $\widehat{\boldsymbol{\theta}}$, obtenido al maximizar $\ell_{\boldsymbol{\tau}}(\boldsymbol{\tau}; \boldsymbol{\theta})$, la función de verosimilitud asociada con el vector $\boldsymbol{\tau}$ de contrastes de error, es el mismo que el obtenido al maximizar la función de verosimilitud asociada con algún otro vector de $N - p$ contrastes de error linealmente independientes. Puede demostrarse que si $\mathbf{u} = \mathbf{C}'\mathbf{Y}$ es cualquier vector de $N - p$ contrastes de error linealmente independientes, la función de verosimilitud asociada con $\mathbf{u}$ es un múltiplo escalar de $\ell_{\boldsymbol{\tau}}(\boldsymbol{\tau}; \boldsymbol{\theta})$ que no depende de $\boldsymbol{\theta}$. (Davis, 2002)

Así, ignorando una constante aditiva que no depende de $\boldsymbol{\theta}$, la función de log-verosimilitud $L_R(\boldsymbol{\theta}; \mathbf{Y})$ asociada con cualquier vector de $N-p$ contrastes de error linealmente independientes es

$$L_R(\boldsymbol{\theta}; \mathbf{Y}) = -\frac{1}{2}\left[\ln|\mathbf{V}| + \ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\right], \tag{2.6}$$

donde $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{Y}$. En aras de realizar una comparación, se presenta en (2.7) la función de log-verosimilitud para $\mathbf{Y}$. Tal como se observa, la única diferencia es que $L_R(\boldsymbol{\theta}; \mathbf{Y})$ tiene el término adicional $-\frac{1}{2}\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|$.

$$L(\boldsymbol{\theta}; \mathbf{Y}) = -\frac{1}{2}\left[\ln|\mathbf{V}| + (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\right]. \tag{2.7}$$

Así, el estimador $\widehat{\boldsymbol{\theta}}$ es un estimador *REML* de $\boldsymbol{\theta}$ si $L_R(\boldsymbol{\theta}; \mathbf{Y})$ alcanza su valor máximo en $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. En el modelo lineal usual, $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, las ecuaciones *REML* de (2.6) tienen solución única, $\widehat{\boldsymbol{\theta}}$, la cual coincide con el estimador insesgado de mínima varianza

$$\widehat{\sigma}^2_{\text{REML}} = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{N - p}. \tag{2.8}$$

Ahora, en los *LMM* balanceados tipo ANOVA, las ecuaciones del método *REML* en (2.6) también tienen solución única que coincide con la estimación del modelo ANOVA. En general, sin embargo, obtener una estimación *REML* de $\boldsymbol{\theta}$ requiere de métodos iterativos para maximizar la función no lineal $L_R(\boldsymbol{\theta}; \mathbf{Y})$ sujeta a la restricción $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. Se pueden utilizar algoritmos como *Newton-Raphson* y el método de puntuación de Fisher, conocido en inglés como el método *Fisher scoring*.

Respecto a lo anterior, Jennrich & Schluchter (1986) desarrollaron un algoritmo *EM* generalizado (*GEM*, en inglés) para calcular estimaciones *REML* de $\boldsymbol{\theta}$. En este algoritmo, la verosimilitud se incrementa (en lugar de maximizarse) en cada paso *M*. El algoritmo *GEM* está restringido al modelo de datos incompletos, pero tiene la ventaja de poder ajustar matrices de covarianza con un gran número de parámetros.

Segun comenta Davis (2002), Jennrich & Schluchter (1986) describen cómo calcular estimaciones de error estándar a partir de la inversa de la matriz de información de *Fisher* (cuando se usa *Fisher scoring*) y de la matriz de información empírica (cuando se usa *Newton-Raphson*). Concluyen que las estimaciones de error estándar de la matriz de información empírica son preferibles cuando los datos están incompletos. Aunque su algoritmo *GEM* no produce errores estándar para los elementos de $\boldsymbol{\theta}$, éstos pueden obtenerse dando un solo paso adicional, ya sea por *Newton-Raphson* o por *Fisher scoring*, después de la convergencia.

Jennrich & Schluchter (1986) también comparan los algoritmos computacionales antes mencionados. El algoritmo de *Newton-Raphson* tiene una tasa de convergencia cuadrática y generalmente converge en un pequeño número de iteraciones (pero con un mayor costo computacional por iteración). El algoritmo *GEM* tiene el menor costo por iteración, pero puede requerir un gran número de las mismas. El algoritmo *Fisher scoring* se comporta de manera intermedia en términos de costo por iteración y número de iteraciones; el costo por iteración a menudo no es mucho menor que el de *Newton-Raphson*, pero puede requerir un número mucho mayor de iteraciones.

Finalmente, llegan a la conclusión de que cuando m, el número de parámetros de covarianza, es pequeño, se prefiere el algoritmo de *Newton-Raphson* porque éste no está restringido al modelo de datos incompletos y porque la convergencia es generalmente rápida. Sin embargo, con m grande, como cuando se ajusta una matriz de covarianza no estructurada a más de diez puntos de tiempo, sólo es posible emplear el algoritmo *GEM*. (Davis, 2002)

2.3. Criterios de selección del modelo

Tal como se vió anteriormente, el ajuste del modelo en (2.5) depende de la estimación de $\boldsymbol{\theta}$, el vector que contiene las componentes de varianza. En este sentido, la selección del mejor modelo debe involucrar criterios que permitan determinar si la estructura de covarianza elegida es la adecuada. Jiang (2007) y Davis (2002) coinciden en que el *criterio de información de* Akaike (1973) (AIC, por sus siglas en inglés) y el *criterio de información bayesiano* de Schwarz (1978) (BIC, en inglés), se constituyen como los criterios pioneros de selección de modelos y pueden ser empleados en aras de seleccionar la mejor estructura de covarianza.

Tanto el AIC como el BIC penalizan la log-verosimilitud por el número de parámetros y/o el número de observaciones. Así, la mayoría de las referencias estadísticas definen el AIC como

$$\text{AIC} = -2L(\boldsymbol{\theta}; \mathbf{Y}) + 2p, \tag{2.9}$$

donde p es el número de parámetros del modelo. De manera similar, el BIC se suele definir como

$$\text{BIC} = -2L(\boldsymbol{\theta}; \mathbf{Y}) + p\log(n), \tag{2.10}$$

donde n es el número de observaciones. El modelo con el AIC (BIC) más pequeño se considera mejor; el BIC tiene una mayor penalización por sobreajuste en comparación con el AIC y los dos criterios pueden no estar de acuerdo en cuanto a qué modelo es el mejor. Jones (1993, p. 46-47) recomienda usar el AIC como criterio de selección al mencionar que aquellos modelos que se encuentran dentro de los dos valores de AIC más bajos se consideran competitivos; de los modelos competitivos, generalmente se selecciona el que tiene menos parámetros. (Duong, 1984)

Tal como suele implementarse en la mayoría de *softwares* y paquetes estadísticos, el AIC y el BIC se pueden usar, en el contexto de los *LMM*, para comparar modelos con los mismos efectos fijos, pero diferentes estructuras de covarianza. Nótese que en (2.9) y en (2.10), $L(\boldsymbol{\theta}; \mathbf{Y})$ puede ser reemplazada por $L_R(\boldsymbol{\theta}; \mathbf{Y})$ y el criterio seguiría proporcionando resultados similares, en cuanto a la selección del mejor modelo, o lo que es lo mismo, la mejor estructura de covarianza.

2.4. Predicción de nuevas observaciones

En ocasiones lo que se desea es poder predecir el comportamiento medio, a través de los tiempos de medición considerados, de una variable particular en sujetos o unidades que no formaban parte del conjunto inicial a partir del cual se ajustó un *LMM*. Para este propósito, puede hacerse uso del *LMM* ajustado y de los vectores de covariables asociados a los nuevos sujetos, de modo que pueda obtenerse su predicción.

Jiang (2007) presenta un enfoque para predecir nuevas observaciones a partir de un modelo *LMM* ajustado, asumiendo generalidad distribucional, tal como se verá a continuación. Sea Y_{ij*} una observación futura que se desea predecir. Supóngase que Y_{ij*} satisface un modelo lineal mixto usual. Entonces, Y_{ij*} puede expresarse como

$$Y_{ij*} = \mathbf{x}'_{ij*}\boldsymbol{\beta} + b_{*1} + \cdots + b_{*q} + e_{ij*},$$

donde $\mathbf{x}_{ij*}$ es un vector conocido de covariables (no necesariamente presente en los datos), b_{*r} son efectos aleatorios y e_{ij*} es un error, tal que $b_{*i} \sim F_{ir}, \leq i \leq q$, $e_{ij*} \sim F_0$, donde las F's son distribuciones desconocidas (no necesariamente normales) y $b_{*1}, \ldots, b_{*q}$, e_{ij*} son independientes.

Se asume que Y_{ij*} es independiente de $Y_{ij} = (Y_{ij})_{1\leq i\leq n}$. De este modo, el mejor predictor puntual de Y_{ij*}, cuando se conoce $\boldsymbol{\beta}$, es $E(Y_{ij*}|Y_{ij}) = E(Y_{ij*}) = \mathbf{x}'_{ij*}\boldsymbol{\beta}$. Debido a que $\boldsymbol{\beta}$ es desconocido, se reemplaza por un estimador consistente, $\widehat{\boldsymbol{\beta}}$, que puede ser el estimador *OLS* (estimador mínimos cuadrados ordinarios, por sus siglas en inglés) o el estimador *BLUE* de (2.2).

Lo anterior da como resultado el mejor predictor empírico (2.11) en el contexto de los *LMM*, el cual permite predecir para nuevas o futuras observaciones, teniendo en cuenta el ajuste realizado para datos previamente recolectados.

$$\widehat{Y}_{ij*} = \mathbf{x}'_{ij*}\widehat{\boldsymbol{\beta}}. \tag{2.11}$$

Error cuadrático medio de predicción

El error cuadrático medio de predicción asociado al predictor de (2.11) se obtiene a través de

$$\begin{aligned} E[(\widehat{Y}_{ij*} - Y_{ij*})^2] &= Var(\widehat{Y}_{ij*} - Y_{ij*}) + E^2(\widehat{Y}_{ij*} - Y_{ij*}) \\ &= \widehat{\sigma}^2_* + \mathbf{x}'_{ij*}\mathbf{A}'\mathbf{x}_{ij*} \\ &= \widehat{\sigma}^2_* + \mathbf{x}'_{ij*}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}_{ij*}; \end{aligned} \tag{2.12}$$

donde $\widehat{\sigma}^2_*$ es la varianza del proceso la cual se estima como en (2.13).

$$\widehat{\sigma}^2_* = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\widehat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{N - (p+1)}. \tag{2.13}$$

2.5. Aplicación: modelo LMM del ozono como contaminante del aire en España

de Castro González (2012) menciona que la contaminación atmosférica en España se debe a la cantidad de partículas contaminantes presentes en el aire (causantes del 1.6 % de las muertes mundiales), los óxidos de nitrógeno y el ozono; sin embargo, la contaminación por ozono tiene mayor relevancia, por presentarse en mayor proporción.

El ozono se origina a partir de *productos precursores* (óxidos de nitrógeno, NO_x, y los *compuestos orgánicos volátiles*, *COV*, especialmente hidrocarburos) bajo condiciones de intensa insolación y elevadas temperaturas, factores que abundan en España. Luego, no es de extrañar que los um-

brales definidos por la Unión Europea (*UE*) sean sobrepasados incluso en periodos de tiempos consecutivos, debiéndose también a las corrientes de aire provenientes desde la costa mediterránea, las cuales transportan y elevan los valores de ozono a zonas residenciales o rurales en las que no debería presentarse. (de Castro González, 2012)

De este modo, el viento y el comportamiento aerodinámico general sobre el territorio español juegan un papel determinante a la hora de estudiar el perfil del ozono como contaminante del aire. Tal comportamiento puede consultarse, por ejemplo, en `https://www.eltiempo.es`, cuyas mediciones en tiempo real permiten identificar la velocidad y dirección del mismo. Así, imágenes como la de la *Figura 2.1* pueden ser visualizadas, evidenciando una tendencia circular en los vientos contrario a las manecillas del reloj, cuyo origen se da en sentido norte-sur por el occidente de España, para luego moversen hacia el oriente del país, retomando al final su dirección predominante, esto es, hacia el norte de la península; se observa además vientos con velocidades de entre 18 y 35 *km/h*, mayoritariamente. Las corrientes de aire provenientes del norte, se encuentran con las que se generan en la costa del mediterráneo y ascienden con dirección norte, de forma circular tal como antes fue mencionado. (Molina, 2015)

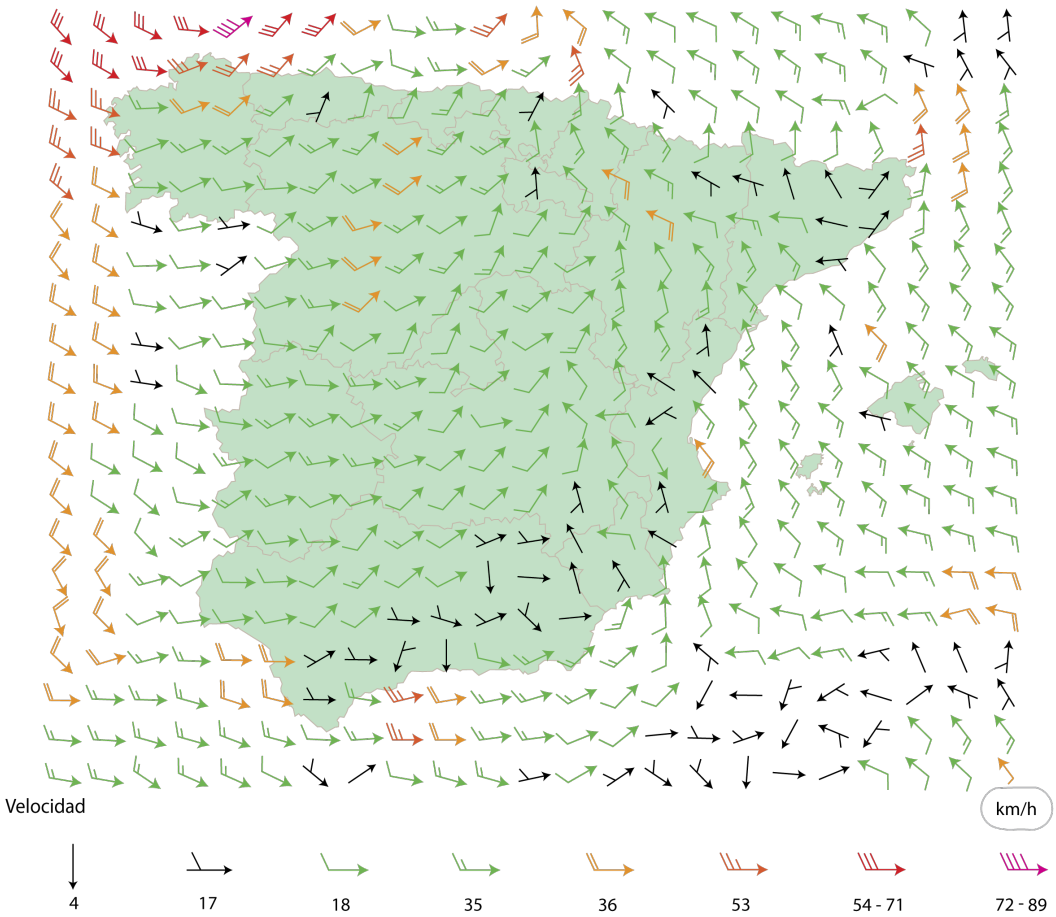


FIGURA 2.1. Comportamiento del viento sobre el territorio español peninsular. Fuente: `https://www.eltiempo.es/viento`.

Ahora, Shirk (2000) menciona los efectos que la contaminación del aire por ozono trae sobre el ser humano (ver *Tabla 2.2*), los cuales han llevado a la especificación de umbrales, límites o valores de referencia regulados por la *UE*; en Europa el valor de referencia para la concentración de ozono en exteriores es de 180 $\mu g/m^3$. Las concentraciones de ozono pueden variar significativamente en función del lugar y del tiempo. Al final de la tarde, por ejemplo, las concentraciones de este gas pueden ser mucho mayores en las afueras de una ciudad o en las áreas rurales, que las que se detectan en el centro de la ciudad. Este efecto se debe a que el ozono se degrada más rápidamente

en presencia de otros contaminantes; es decir, la concentración de ozono disminuye con anterioridad en las zonas urbanas que en áreas con bajas concentraciones de otros contaminantes aéreos (NO_x, partículas de polvo).

Así, conocer la distribución del ozono sobre el territorio nacional español, con el objetivo de identificar posibles zonas de intervención o focos de localización del mismo, es de interés para las agencias ambientales, pues atacarían de manera focalizada la problemática de contaminación atmosférica debida a este gas. Además, se tendrían herramientas para concientizar a la comunidad civil en general acerca de la problemáticas ambientales y la forma en que pueden mitigarse. (de Castro González, 2012)

Europa cuenta con estaciones de monitoreo de la calidad del aire, que además de registrar datos para el ozono, miden los niveles de diversos contaminantes atmosféricos. La *Agencia Europea de Medio Ambiente* (*European Environment Agency*, *EEA*, en inglés) es la encargada de proporcionar información sólida e independiente sobre el medio ambiente para quienes participan en el desarrollo, adopción, implementación y evaluación de la política ambiental, y también del público en general. En estrecha colaboración con la *Red Europea de Información y Observación Ambiental* (*Environmental Information and Observation Network*, *Eionet*, en inglés) y sus 33 países miembros, la *EEA* recopila datos y produce evaluaciones sobre una amplia gama de temas relacionados con el medio ambiente, incluída la calidad del aire. (EEA, 2019)

AirBase es el sistema de información sobre la calidad del aire que mantiene la *EEA* a través del centro europeo de temas sobre la contaminación del aire y la mitigación del cambio climático. Contiene datos de monitoreo de la calidad del aire emitidos anualmente y que han sido reportados por los países participantes en toda Europa, quienes en el marco de la decisión *97/101/EC* y al pertenecer a la *UE*, están obligados a participar en un intercambio recíproco de información (*Exchange of Information*, *EoI*, en inglés) sobre la calidad del aire en sus naciones. La EEA (2019), en conjunto con sus países miembros y colaboradores, recopilan la información prevista por la decisión de *EoI*, ya que la contaminación del aire es un problema continental; la agencia por su parte se encarga de realizar evaluaciones de la calidad del aire y de cubrir toda el área geográfica de Europa.

Tabla 2.2. Efectos de la contaminación del aire por ozono. (Shirk, 2000)

$\mu g/m^3$	Efectos
30	Perceptible al olfato; sin embargo, la habituación es muy rápida.
70	Irritaciones iniciales de la conjuntiva ocular.
100	Probabilidad de jaquecas.
160	En animales se reduce la capacidad de resistencia a infecciones pulmonares bacterianas.
160-200	Disfunción pulmonar, especialmente al hacer ejercicio.
200	Aumento de la cantidad de leucocitos, inactivación del sistema de inmunidad.
240-300	Mayor frecuencia de ataques de asma.
240-700	Reducción de la fuerza física.
400	Tos, dolor toráxico. Después de 4 horas de exposición a 400 $\mu g/m^3$ de ozono tienen lugar cambios hormonales y enzimáticos.
800	Reacción inflamatoria de los tejidos.
1000	Después de 6-10 horas de exposición: daños iniciales en los cromosomas humanos.

Agregando a lo anterior, la base *AirBase* consta de una serie de tiempo plurianual de datos y estadísticas de medición de la calidad del aire para una serie de contaminantes del mismo. También contiene metainformación sobre las redes de monitoreo involucradas, sus estaciones y sus mediciones. La base de datos cubre geográficamente todos los Estados miembros de la *UE*, los

países miembros de la *EEA* y algunos países colaboradores de la misma, dentro de los cuales se encuentra España. (EEA, 2019)

Así, debido a que se tienen mediciones anuales de contaminantes atmosféricos, arrojadas por un número contable de estaciones de monitoreo sobre el territorio español, puede pensarse en un enfoque a través de modelos *LMM*, de modo que se explique el comportamiento de tales contaminantes en función de la variabilidad debida a cada estación en particular y la debida a la interacción entre estaciones de medición.

Datos de aplicación y ajuste del LMM

Tal como se mencionó, *AirBase* almacena información para una gran variedad de contaminantes atmosféricos dentro de los cuales se encuentra el ozono (O_3), medido en microgramos por metro cúbico ($\mu g/m^3$). Además, cuenta con información para una ventana de tiempo anual desde 1969 hasta 2011; las mediciones son reportadas a diario por cada estación, por lo que el dato anual corresponde al promedio de los reportes hechos para cada uno de los 365 días que lo componen. Cada estación de monitoreo se encuentra identificada con un código único y georreferenciada en coordenadas geográficas (*longitud, latitud*).

Como el objetivo es modelar el comportamiento del ozono en España, a través de modelos *LMM*, la información de *AirBase* fue filtrada de forma que se obtuvieran sólo los datos asociados a las estaciones españolas. Esto, arrojó una ventana de tiempo anual desde 2007 hasta 2010 y un total de $n = 296$ estaciones de monitoreo, conformando un conjunto de datos balanceados. En la *Figura 2.2* se presenta la configuración de las 296 estaciones en estudio sobre el territorio español.

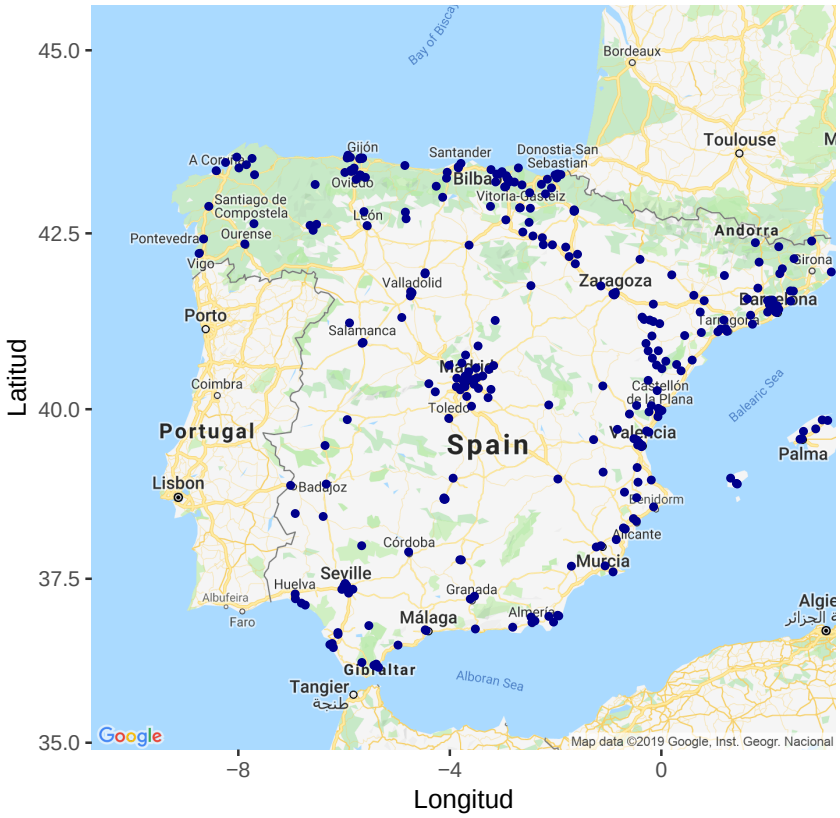
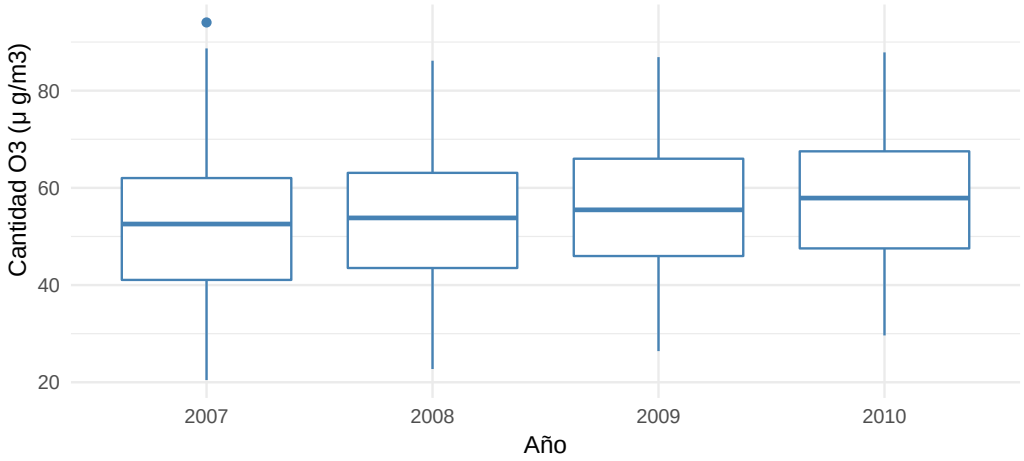
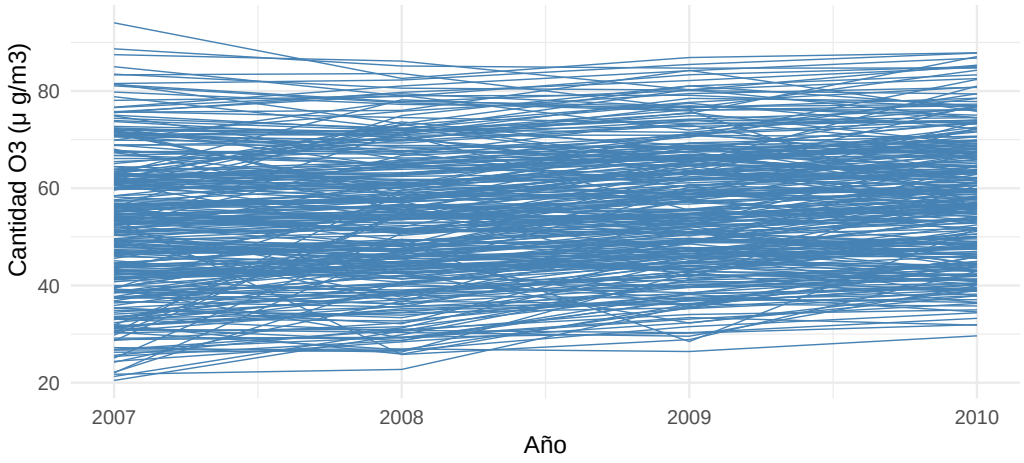


FIGURA 2.2. Configuración de la estaciones de monitoreo de la calidad del aire en estudio sobre el territorio español. Fuente: Google Maps 2019, `maps.google.com`.

Un análisis descriptivo inicial será abordado a continuación. Los *boxplots* de la *Figura 2.3(a)* señalan un comportamiento simétrico de los niveles de ozono para cada año de estudio; no se presenta una cantidad elevada de valores atípicos lo cual podría ser evidencia de tasas bajas en errores de medición (baja variabilidad), además ni los valores medios ni la variabilidad de ozono presentada para cada año parece diferir en gran medida. Se presentan también los perfiles individuales de ozono promedio para cada una de la 296 estaciones de monitoreo; éstos permiten analizar la evolución temporal de la variable en estudio para cada estación (*Figura 2.3(b)*). Tal como se observa, la mayoría de las estaciones no presentan grandes diferencias en los niveles promedio de ozono por año, los cuales fluctúan entre 20 $\mu g/m^3$ y 90 $\mu g/m^3$.



(a) *Boxplots* de la cantidad promedio de ozono por año.



(b) Gráfico de perfiles por estación de monitero (2007-2010).

FIGURA 2.3. Boxplots y gráficos de perfiles individuales para los niveles promedio de ozono por año (2007-2010).

La presencia de baja variabilidad en los niveles de ozono reportados por las estaciones, indicando que las mediciones obtenidas son bastante homogéneas, puede corroborarse con los valores de los coeficientes de variación de la *Tabla 2.3*. Éstos son menores del 30 % y por consiguiente indicadores de poca heterogeneidad en la información. Puede observarse que la concentración de ozono, para la ventana de tiempo estudiada, se mantiene dentro de los límites establecidos por la *UE* (máximo 180 $\mu g/m^3$) y los riesgos a los que se estaría expuesto serían irritaciones iniciales de la conjuntiva ocular y la probable aparición de jaquecas (ver *Tabla 2.2*).

TABLA 2.3. Medidas de tendencia central, localización y variabilidad para los niveles de ozono promedio en España (2007-2010).

	2007	2008	2009	2010
Media	52.07	53.42	56.12	57.91
Mediana	52.56	53.82	55.48	57.90
Mínimo	20.46	22.73	26.41	29.64
Máximo	94.04	86.16	86.90	87.87
Cuartíl Inf.	41.06	43.51	45.97	47.54
Cuartíl Sup.	62.02	63.08	66.00	67.52
Desv. Estándar	15.05	13.95	13.30	13.07
Coef. de Variación	0.29	0.26	0.24	0.23

Con el objetivo de estudiar la variabilidad de la concentración de ozono dentro y entre estaciones de monitoreo, se usará un modelo lineal mixto de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \mathbf{e}$, el cual puede ser particularizado para la presente aplicación como en (2.15). Así, Y_{ij} indica la concentración de O_3 registrada por la estación de monitoreo i ($i = 1, \ldots, 296$) en el *j-ésimo* año, con j desde el 2007 hasta el 2010. Las dos covariables x_{i1} y x_{i2} consideradas corresponden a las coordenadas planas de cada estación, en longitud y latitud, respectivamente.

Se tiene además que $b_i \sim N(0, \sigma_b^2)$ es el efecto aleatorio común para los registros tomados en la *i-ésima* estación. Finalmente, los e_{ij} son términos de error para la *i-ésima* estación en el *j-ésimo* año, de modo que $\mathbf{e} \sim \mathbf{N_4}(\mathbf{0}, \mathbf{R})$, donde $\mathbf{R} \in \mathcal{M}_{(4\times4)}$ es una matriz que refleja la estructura de covariación dentro de cada estación de monitoreo, esto es, entre años de medición.

De este modo, tal como se vió en la subsección 2.2 y en la sección 2.3, el ajuste del *LMM* adecuado depende de la estimación de $\boldsymbol{\theta}$, el vector que aloja las componentes de varianza de $\mathbf{G}$ y $\mathbf{R}$; entonces, si $\boldsymbol{\theta}_1$ es el vector que aloja las componentes de varianza de $\mathbf{R}$, $\boldsymbol{\theta} = (\sigma_b^2, \boldsymbol{\theta}_1')'$, asumiendo independencia entre estaciones de monitoreo e igualdad de varianza en las mismas.

Haciendo uso de `lme()` de la librería `nlme` en *R*, se ajustaron 3 modelos usando *REML*. El primero de éstos contempló una estructura de correlación simétrica compuesta para $\mathbf{R}$, el segundo una *AR(1)* y el tercero una matriz $\mathbf{R}$ no estructurada (ver *Tabla 2.1*). El AIC y BIC obtenido para cada modelo se presenta en la *Tabla 2.4*; luego, siguiendo a Jones (1993), el mejor modelo es el que minimiza el AIC y por consiguiente se elige el tercero de éstos.

En ese orden de ideas, el *LMM* seleccionado cuenta con una $\mathbf{R}$ no estructurada, ajustada en términos de correlación como en (2.14); además, $\sigma_b^2 = 26.907$. Entonces, $\widehat{\boldsymbol{\theta}} = (26.907, \widehat{\boldsymbol{\theta}}_1')'$, donde cada componente de varianza en $\widehat{\boldsymbol{\theta}}_1$ es explicitado en (2.14). La varianza del proceso estimada por `lme()` corresponde a $\widehat{\sigma}_*^2 = 27.5053$.

$$\begin{aligned} \widehat{\boldsymbol{\rho}} &= \text{Diag}(\widehat{\mathbf{R}})^{-1/2}\widehat{\mathbf{R}}\text{Diag}(\widehat{\mathbf{R}})^{-1/2} \\ &= \begin{pmatrix} 1 & 0.221 & -0.319 & -0.667 \\ 0.221 & 1 & 0.413 & -0.020 \\ -0.319 & 0.413 & 1 & 0.479 \\ -0.667 & -0.020 & 0.479 & 1 \end{pmatrix}. \end{aligned} \tag{2.14}$$

$$\begin{pmatrix} Y_{17} \\ Y_{18} \\ Y_{19} \\ Y_{110} \\ Y_{27} \\ \vdots \\ Y_{2967} \\ Y_{2968} \\ Y_{2969} \\ Y_{29610} \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{11} & x_{12} \\ 1 & x_{11} & x_{12} \\ 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{2961} & x_{2962} \\ 1 & x_{2961} & x_{2962} \\ 1 & x_{2961} & x_{2962} \\ 1 & x_{2961} & x_{2962} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{295} \\ b_{296} \end{pmatrix} + \begin{pmatrix} e_{17} \\ e_{18} \\ e_{19} \\ e_{110} \\ e_{27} \\ \vdots \\ e_{2967} \\ e_{2968} \\ e_{2969} \\ e_{29610} \end{pmatrix} \tag{2.15}$$

Tabla 2.4. Selección del modelo usando los valores del AIC y el BIC.

Modelo	Estructura de Covarianza	AIC	BIC
1	Simétrica compuesta	7990.373	8020.818
2	*AR(1)*	8195.611	8226.056
3	No estructurada	7986.93	8042.745

De este modo, los efectos fijos ajustados usando `lme()` se encuentra en la *Tabla 2.5*. Tal como se observa, el efecto asociado a la longitud es el menos significativo, lo cual se debe a que la dirección predominante del viento es en sentido norte; esto hace que la concentración de O_3 cambie mayoritariamente sobre los paralelos que atraviesan España y en una menor proporción sobre los meridianos asociados al mismo país.

Tabla 2.5. Efectos fijos ajustados.

Parámetro	Estimación	Std. Error	*df*	Valor *t*	*p-valor*
β_0	164.98608	13.587659	888	12.14235	0.0000
β_1	0.57460	0.239117	293	2.40302	0.0169
β_2	-2.66622	0.332815	293	-8.01112	0.0000

Se observa que a medida que aumenta un grado de longitud, la concentración de ozono aumenta aproximadamente 0.57 $\mu g/m^3$, es decir, se evidencia un aumento de tal concentración hacia el oriente del país. De forma similar, si se aumenta un grado de latitud, la concentación de ozono disminuye en aproximadamente 2.66 $\mu g/m^3$, esto es, hacia el norte de España.

Una vez explicitada la estimación de los efectos fijos, es posible hallar $\widehat{\mathbf{b}}$ a través de (2.3), con lo que el *LMM* queda definido. Finalmente, se observa en la *Tabla 2.6* que los residuales *dentro* de estaciones estandarizados tienen un comportamiento simétrico marginal similar al de una distribución normal.

Tabla 2.6. Medidas de localización asociadas a los residuales *dentro* de estaciones de monitoreo estandarizados.

Mínimo	Q_1	Mediana	Q_3	Máximo
-3.890929501	-0.497828007	-0.009367372	0.504742595	3.423728997

Como la normalidad marginal no implica normalidad multivariante, se efectúa además el *test* de Shapiro - Wilk multivariado sobre los errores del modelo para verificar el supuesto de multinormalidad asumido en el mismo; se obtiene una estadística de prueba de $W = 0.992$ junto con un *p-valor* asociado de 0.0879. De este modo, trabajando a un nivel de significancia del 5 %, la hipótesis de multinormalidad no es rechazada y existe evidencia estadística para garantizar el cumplimiento del supuesto antes mencionado.

Con el modelo estimado es posible caracterizar el comportamiento medio de una estación de monitereo particular a través de los cuatro años considerados, haciendo uso del *BLUP* presentado en (2.4). Sin embargo, si lo que se quiere es predecir la concentración de ozono para lugares en España diferentes a las estaciones de monitoreo, debe usarse el predictor descrito en la sección 2.4, tal como se verá en el *Capítulo 4*.

Los resultados hasta el momento obtenidos únicamente contemplan la dependencia temporal asociada a la variabilidad *dentro* de cada estación de monitoreo para los años de estudio; sin embargo, existen situaciones en las que debe estudiarse tanto la correlación temporal como la dependencia espacial por ser inherentes a las variables involucradas en el estudio. De este modo, resulta de interés tener en cuenta la distribución geográfica particular de las observaciones y la componente longitudinal de las mismas.

Militino et al. (2008) proponen entonces un *LMM* espacio-longitudinal para modelar ambos tipos de variabilidad en simultáneo, acomodando adecuadamente la covariación *entre* y *dentro* de ubicaciones de estudio en las matrices de varianzas y covarianzas asociadas a los efectos aleatorios y al vector de los errores del modelo, tal como será estudiado en el *Capítulo 3*.

CAPÍTULO 3

Modelo espacio-longitudinal de efectos mixtos

Con el objetivo de ofrecer un enfoque adecuado para el modelamiento de cualquier medición de la contaminación del medio ambiente, para datos que han sido recogidos en el tiempo y que se encuentran espacialmente correlacionados, Militino et al. (2008) proponen un modelo que usa la teoría de *modelos lineales mixtos* (LMM).

Según mencionan, dentro de ese contexto, la posibilidad de obtener un mapa donde se muestren las mediciones globales realizadas en la zona de estudio es de particular preocupación, pues sería de gran ayuda para las agencias ambientales poder estudiar la distribución espacial de la medición de la contaminación, teniendo en cuenta todas las mediciones en el tiempo. Para lograr este propósito, se necesitan predicciones en nuevas ubicaciones, las cuales pueden obtenerse a través de predictores lineales óptimos como los desarrollados para el *kriging* o el *cokriging*, o usando la formulación del LMM expuesta por Militino et al. (2008), tal como se verá a continuación.

3.1. Supuestos y formulación del modelo

Para acomodar la variabilidad en el espacio y el tiempo utilizando un modelo lineal mixto, puede definirse una matriz de covarianza adecuada $\mathbf{G}$ en los efectos aleatorios $\mathbf{b}$, para indicar la covariación común entre las observaciones tomadas en una única ubicación y una matriz de covarianza adecuada $\mathbf{R}$ en el término del error, la cual explica la estructura de dependencia espacial, que puede ser diferente en cada período de tiempo considerado. Así, Militino et al. (2008) definen el modelo *espacio-longitudinal de efectos mixtos* como en (2.1), cuyo equivalente matricial, para un proceso balanceado, se presenta en (3.1); asumen explícitamente que $\mathbf{b} \sim \mathbf{N}(\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \mathbf{R})$ y ambos son independientes.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \mathbf{e} \tag{3.1}$$

En el modelo espacio-longitudinal de (3.1), $\mathbf{Y}$ es el vector $(N \times 1)$ de datos recopilados en las n ubicaciones para t diferentes momentos del tiempo. $\mathbf{X}$ es la matriz de diseño $\mathbf{X} \in \mathcal{M}_{\{N \times p\}}$, donde p es el número de covariables. La matriz $\mathbf{X}$ incluye las covariables que representan las coordenadas x e y para cada ubicación y también puede incluir otra información relacionada con los datos, como el día o la temporada, en el momento de la recopilación de los mismos. $\boldsymbol{\beta}$ es el vector $(p \times 1)$ de efectos fijos.

De este modo, la parte fija del modelo, $\mathbf{X}\boldsymbol{\beta}$, desempeña el mismo papel que el término de tendencia $\boldsymbol{\mu}(s)$ en el modelo lineal espacial. Los efectos aleatorios pueden tomar diferentes formas dependiendo del objetivo del estudio, sin embargo, una forma simple y útil, según proponen Militino et al. (2008), es considerar $\mathbf{b}$ como un vector de dimensión $(n \times 1)$, asociado al número de ubicaciones y $\mathbf{Z}$ como su matriz de diseño correspondiente.

Se supone además que la matriz de covarianza $\mathbf{G}$ es diagonal con elementos iguales σ_b^2; b_i es un efecto específico del sujeto que induce la correlación entre dos mediciones tomadas en la misma ubicación. El término del error, representado por el vector $\mathbf{e}$ de dimensión $(N \times 1)$, modela la estructura espacial. Ahora, cuando se intenta modelar las mediciones recopiladas en diferentes momentos del tiempo, puede ser apropiado usar diferentes estructuras de covarianza para cada uno de los momentos, ya que las dependencias espaciales entre las mediciones pueden variar en el tiempo. Esto se puede lograr fácilmente usando una matriz $\mathbf{R}$ apropiada. (Militino et al., 2008)

3.2. Predictor y error cuadrático medio de predicción

Si la matriz $\mathbf{V}$ presentada en la sección 2.2 fuera conocida, el predictor natural del término de error en la ubicación $\mathbf{x}_{ij*}$ será el mejor predictor lineal insesgado

$$\widehat{e}_{ij*} = E(e_{ij*}|\mathbf{e}) = \mathbf{r}_{ij*}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}),$$

donde $\mathbf{r}_{ij*}$ es el vector fila de covarianzas entre e_{ij*} y el vector de los términos de error $\mathbf{e} = (e_1, \ldots, e_N)'$. La predicción de una nueva observación Y_{ij*} en la ubicación $\mathbf{x}_{ij*}$, donde $\mathbf{x}_{ij*} = (1, x_{i1}, x_{i2})'$ representa el intercepto y las covariales asociadas a la *longitud* y *latitud*, viene dada por

$$\begin{aligned} \widehat{Y}_{ij*} = E(Y_{ij*}|\mathbf{Y}) &= \mathbf{x}'_{ij*}\widehat{\boldsymbol{\beta}} + \widehat{e}_{ij*} \\ &= \mathbf{x}'_{ij*}\widehat{\boldsymbol{\beta}} + \mathbf{r}_{ij*}\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \end{aligned}$$

En consecuencia, el modelo espacio-longitudinal también es un interpolador exacto, pues la predicción $\widehat{\mathbf{Y}}$ de $\mathbf{Y}$ coincide con las observaciones muestreadas:

$$\begin{aligned} \widehat{\mathbf{Y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}} + \widehat{\mathbf{e}} \\ &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{ZGZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) + \mathbf{RV}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{X}\widehat{\boldsymbol{\beta}} + (\mathbf{ZGZ}' + \mathbf{R})\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{VV}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{Y}. \end{aligned}$$

Ahora, expresando $\widehat{e}_{ij*}$ de forma más detallada, se obtiene

$$\begin{aligned} \widehat{e}_{ij*} &= (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{V}^{-1}\mathbf{r}'_{ij*} \\ &= (\mathbf{Y} - \mathbf{XAX}'\mathbf{V}^{-1}\mathbf{Y})'\mathbf{V}^{-1}\mathbf{r}'_{ij*} \\ &= \mathbf{Y}'(\mathbf{I} - \mathbf{V}^{-1}\mathbf{XAX}')\mathbf{V}^{-1}\mathbf{r}'_{ij*} \\ &= \mathbf{Y}'(\mathbf{V}^{-1}\mathbf{r}'_{ij*} - \mathbf{V}^{-1}\mathbf{XAX}'\mathbf{V}^{-1}\mathbf{r}'_{ij*}), \end{aligned}$$

y de manera similar

$$\begin{aligned} \widehat{Y}_{ij*} &= \mathbf{x}'_{ij*}\widehat{\boldsymbol{\beta}} + \widehat{e}_{ij*} \\ &= \mathbf{Y}'[\mathbf{V}^{-1}\mathbf{r}'_{ij*} + \mathbf{V}^{-1}\mathbf{XA}(\mathbf{x}_{ij*} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{r}'_{ij*})]. \end{aligned} \tag{3.2}$$

Por lo tanto, el error cuadrático medio de predicción está dado por

$$E[(\widehat{Y}_{ij*} - Y_{ij*})^2] = \sigma_*^2 + \mathbf{r}_{ij*}\mathbf{V}^{-1}\mathbf{r}'_{ij*} - (\mathbf{x}'_{ij*} - \mathbf{r}_{ij*}\mathbf{V}^{-1}\mathbf{X})\mathbf{A}(\mathbf{x}'_{ij*} - \mathbf{r}_{ij*}\mathbf{V}^{-1}\mathbf{X})', \tag{3.3}$$

donde σ_*^2 es la varianza del proceso y ésta se estima a través de

$$\widehat{\sigma}_*^2 = \frac{(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})'\widehat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})}{N - (p+1)}.$$

Finalmente, Militino et al. (2008) mencionan que maximizar la función de log-verosimilitud restringida dada en (2.6) con respecto a $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\mathbf{b}$ y $\mathbf{e}$ no es recomendable, debido a que hay demasiados parámetros involucrados en la función objetivo (Militino, 1998). Así, en la práctica se realiza el proceso de estimación en dos pasos: en el primer paso, se estiman los parámetros de la matriz de covarianza $\mathbf{R}$ y en el segundo paso, se estiman σ_b^2, $\boldsymbol{\beta}$ y $\mathbf{b}$ utilizando el método *REML* y las expresiones (2.2) y (2.3).

3.3. Aplicación: modelo espacio-longitudinal del ozono como contaminante del aire en España

Haciendo uso de información filtrada para España en la *AirBase* descrita en la sección 2.5, se ajusta ahora un modelo espacio-longitudinal como el de Militino et al. (2008), el cual ha sido expuesto en el presente capítulo.

Así, usando los datos de las $n = 296$ estaciones de monitoreo de la calidad del aire, dispuestas como en *Figura 2.2* sobre España y que alojan información anual, desde el 2007 hasta el 2010, de la concentración de ozono (O_3) en microgramos por metro cúbico ($\mu g/m^3$), se ajustará un modelo espacio-longitudinal de la forma $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \mathbf{e}$, de forma que pueda acomodarse tanto la variabilidad en el espacio como la naturaleza longitudinal de los datos.

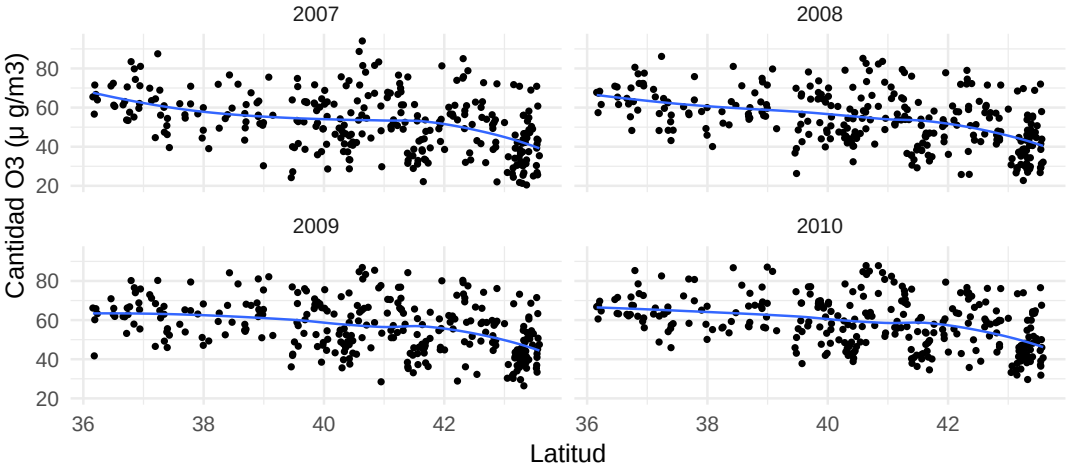
El modelo espacio-longitudinal por ajustar puede particularizarse también como en (2.15), lo cual conlleva a que cada Y_{ij} indica la concentración de O_3 registrada por la estación de monitoreo i ($i = 1, \ldots, 296$) en el j-*ésimo* año, con j desde el 2007 hasta el 2010. Las dos covariables x_{i1} y x_{i2} consideradas corresponden a las coordenadas geográficas de cada estación, en longitud y latitud, respectivamente. Además, $b_i \sim N(0, \sigma_b^2)$ es el efecto aleatorio común para los registros tomados en la i-*ésima* estación.

Los e_{ij} son términos de error para la i-*ésima* estación en el j-*ésimo* año, de modo que $\mathbf{e} \sim \mathbf{N}_{(\mathbf{296\times 4})}(\mathbf{0}, \mathbf{R})$, donde $\mathbf{R} \in \mathcal{M}_{\{(296\times 4)\times(296\times 4)\}}$ es una matriz que refleja la dependencia espacial entre las estaciones de monitoreo, cuya estructura puede ser diferente para cada año de estudio. El vector de parámetros necesarios para determinar $\mathbf{R}$ se denota por $\boldsymbol{\theta}_1$ y consta de diferentes parámetros pertenecientes a los modelos teóricos de semivariogramas ajustados por año. Por lo tanto, el vector de componentes de varianza viene dado por $\boldsymbol{\theta} = (\sigma_b^2, \boldsymbol{\theta}_1')'$. Tal como se discutió en la sección 2.1, el vector de efectos aleatorios $\mathbf{b} \sim \mathbf{N_{296}}(\mathbf{0}, \sigma_b^2\mathbf{I})$ y $\mathbf{e}$ son independientes.

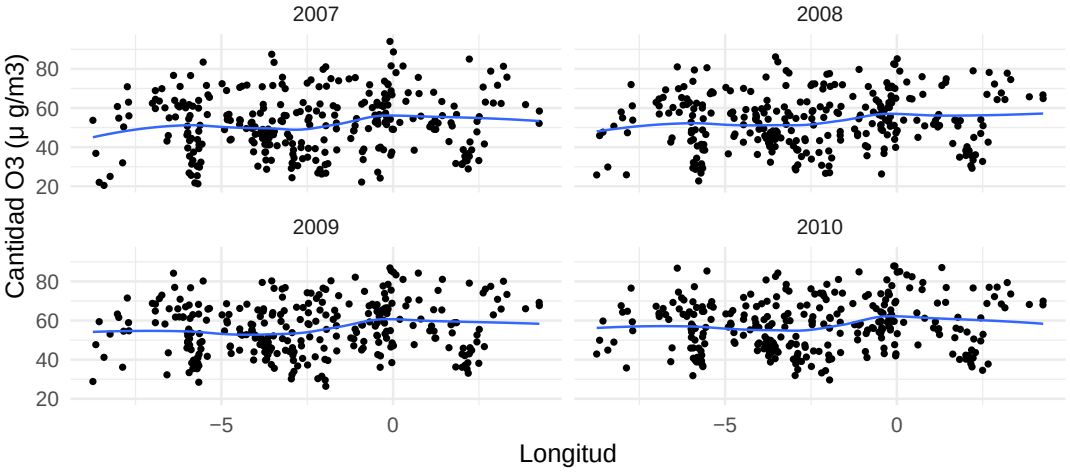
La modelación de la estructura de dependencia espacial, la cual puede ser diferente para cada año, implica contar con el supuesto de estacionariedad en la media. Para verlo, en la *Figura 3.1* se presenta el gráfico de dispersión de la concentración de O_3 en función de la latitud (*Fig. 3.1(a)*) y de la longitud (*Fig. 3.1(b)*). En general, se observa un proceso espacial estacionario en media, pues se presenta una tendencia nula en latitud y longitud para cada año de estudio.

Debe verificarse también que la concentración de ozono sobre España sea un proceso isotrópico, tal como se describió en la sección 1.2. Esto debe hacerse pues, para poder caracterizar la dependencia espacial del O_3 a través de modelos de semivariogramas teóricos, ésta debe depender únicamente de la distancia entre estaciones de monitoreo y no de la dirección de las mismas. De este modo, para verificar ausencia de anisotropía, se han graficado semivariogramas experimentales en cuatro direcciones diferentes y por cada año de estudio, obteniendo lo presentado en la *Figura 3.2*.

Tal como se observa, hay presencia de anisotropía geométrica (ver 1.2.3) con dirección principal de 0^o, esto es, hacia el norte de España. Schabenberger & Gotway (2005) señalan que la anisotropía geométrica es común en los procesos que se desarrollan en direcciones particulares; por ejemplo, la contaminación en el aire probablemente mostrará una anisotropía en la dirección predominante del viento y perpendicular a ella. Así, y debido a que la dirección predominante del viento en España es justamente hacia el norte del país (ver *Fig. 2.1*), la presencia de anisotropía en esa dirección no es de extrañar.



(a) Concentración de O_3 en función de la latitud.



(b) Concentración de O_3 en función de la longitud.

FIGURA 3.1. Modelo de tendencia suavizado referente a la contaminación por O_3 en las direcciones de la latitud y la longitud.

Para corregir la anisotropía geométrica presentada, se realizó una rotación del sistema de coordenadas, de modo que los ejes mayor y menor de los contornos elípticos quedaran alineados, como también una compresión del eje mayor para hacer que los contornos sean esféricos (ver *Fig. 1.3*). Esto se logró calculando la razón de anisotropía λ y definiendo la dirección del eje mayor como $\phi = 0^o$. La escogencia de ϕ obedece a que, tal como se aprecia en la *Figura 3.2*, la dirección 0^o describe el eje mayor de la elipse (variabilidad más lenta) y sus semivariogramas direccionales asociados presentan un rango alrededor de 4 unidades en *coordenadas*[1]; además, el eje menor de la elipse se puede asociar a las demás dirrecciones y tiene un rango de alrededor de 2 unidades en *coordenadas*, así la razón de anisotropía se calcula como $\lambda = 4/2 = 2$.

[1]La distancia en *km* a la que equivale un grado de longitud depende de la latitud. A medida que la latitud aumenta, hacia el Norte o Sur, disminuyen los kilómetros por grado. Para el paralelo del Ecuador, sabiendo que la circunferencia que corresponde al Ecuador mide 40'075.017 *km*, 1^o equivale a 111.319 *km* (resultado de dividir el perímetro del ecuador entre los 360^o de longitud).

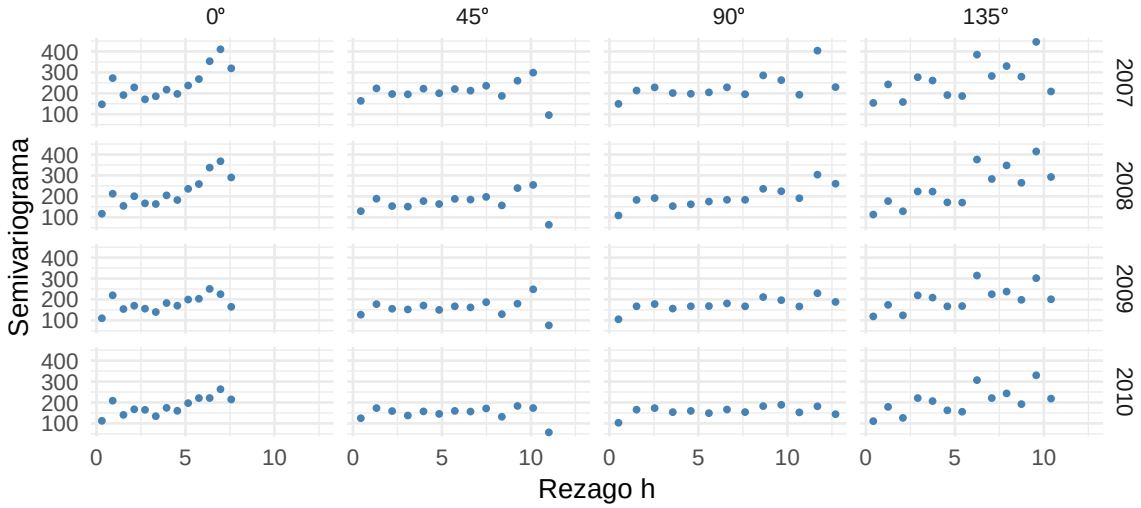


FIGURA 3.2. Semivariogramas experimentales direccionales para cada año de estudio.

Sin embargo, al aplicar la corrección con $\phi = 0^o$, la anisotropía seguía presentándose; así, se eligió $\phi = 90^o$, siguiendo a Schabenberger & Gotway (2005) en su afirmación de presencia de anisotropía en dirección perpendicular a la del viento, logrando la isotropía deseada. Los semivariogramas direccionales del proceso isotrópico se encuentran la *Figura 3.3*. Bajo isotropía,

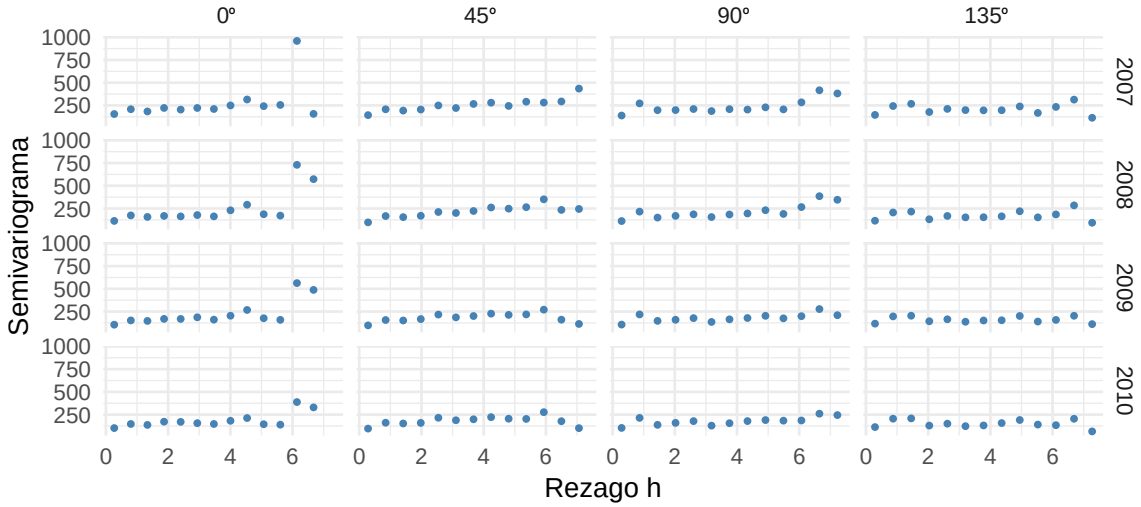


FIGURA 3.3. Semivariogramas experimentales direccionales isotrópicos para cada año de estudio.

modelos teóricos *potencia exponencial* (ver *Tabla 1.1*) fueron ajustados para cada año de estudio ($j = 7, 8, 9, 10$), así

$$\gamma_j(h) = \begin{cases} 0 & \text{si } h = 0 \\ c_{0j} + c_{1j}\left[1 - \exp\left\{-\left(\frac{||h||}{\alpha_j}\right)^{\omega}\right\}\right] & \text{si } ||h|| > 0 \\ \text{donde } \omega = 0.5,\ \alpha_j > 0. & \end{cases} ;$$

donde c_{0j} es el efecto *nugget* por año, $c_{0j} + c_{1j}$ representa la silla para el año j, $\omega = 0.5$ es el parámetro de suavizamiento para la función de covarianza, parámetro común para todos los años (corresponde al valor por defecto empleado por el *software*) y α_j es función del rango de los

semivariogramas por año. De este modo, $\boldsymbol{\theta}_1 = (c_{0j}, c_{1j}, \alpha_j)$ y $\widehat{\boldsymbol{\theta}}_1$ se obtiene ajustando los modelos de semivariogramas teóricos por año, tal como en la *Tabla 3.1.* Los semivariogramas experimentales y ajustados se muestran en la *Figura 3.4.*

TABLA 3.1. Parámetros estimados de los modelos de semivariogramas potencia exponenciales.

	$\widehat{c}_0$	$\widehat{c}_1$	$\widehat{\alpha}$
2007	134.16	9308.08	29632.47
2008	78.71	7780.32	13075.64
2009	113.96	482.00	148.29
2010	116.10	1222.64	1405.23

Tal como se observa, para el 2007 y el 2008 se presenta un mayor grado de dependencia espacial de la concentración de ozono sobre el territorio español. Ahora, la matriz de covarianza $\widehat{\mathbf{R}} \in \mathcal{M}_{\{(296\times4)\times(296\times4)\}}$ toma la forma presentada en (3.4), donde $\widehat{\sigma}_j^2 = c_{0j} + c_{1j}$ para $j = 7,8,9,10$ y $\widehat{\sigma}_{i,lj} = \widehat{c}_{1j}\left[\exp\left\{-\left(\frac{||h_{il}||}{\widehat{\alpha}_j}\right)^{0.5}\right\}\right]$ para cualquier par de estaciones de monitoreo $i \neq l$ con una distancia de separación de h_{il}, en cada año considerado.

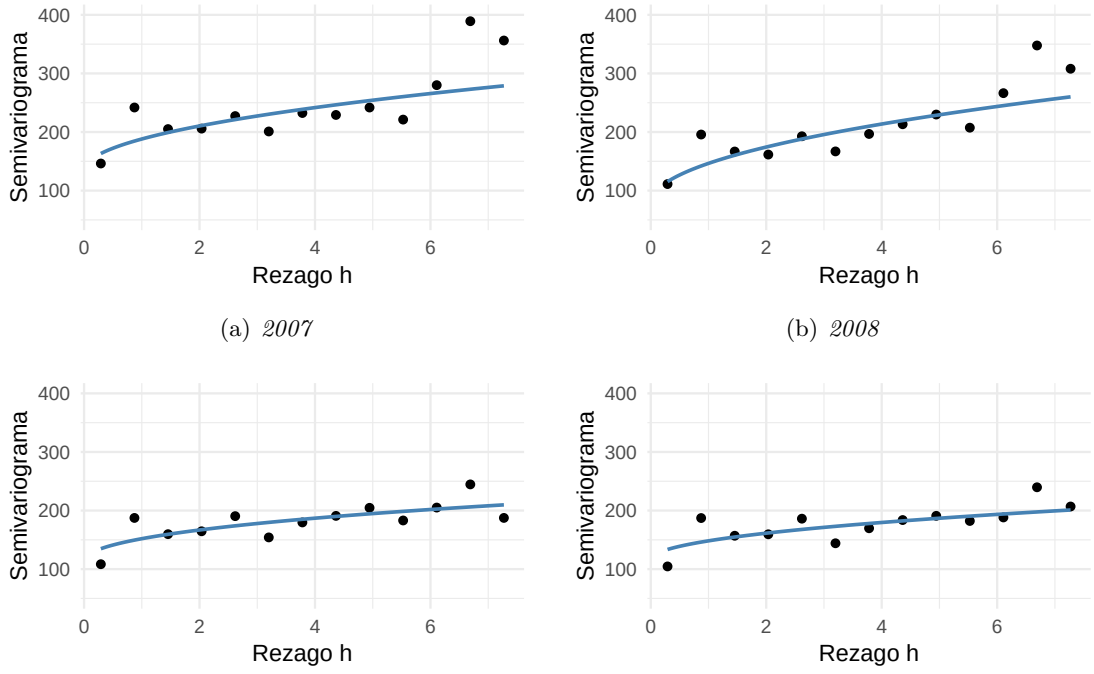
$$\widehat{\mathbf{R}} = \left(\begin{array}{ccc:ccc:c:ccc} \widehat{\sigma}_7^2 & 0 & 0 & \widehat{\sigma}_{1,27} & 0 & 0 & \cdots & \widehat{\sigma}_{1,2967} & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 & \cdots & 0 & \ddots & 0 \\ 0 & 0 & \widehat{\sigma}_{10}^2 & 0 & 0 & \widehat{\sigma}_{1,210} & \cdots & 0 & 0 & \widehat{\sigma}_{1,29610} \\ \hdashline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \hdashline \widehat{\sigma}_{296,17} & 0 & 0 & \widehat{\sigma}_{296,27} & 0 & 0 & \cdots & \widehat{\sigma}_7^2 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 & \cdots & 0 & \ddots & 0 \\ 0 & 0 & \widehat{\sigma}_{296,110} & 0 & 0 & \widehat{\sigma}_{296,210} & \cdots & 0 & 0 & \widehat{\sigma}_{10}^2 \end{array}\right) \tag{3.4}$$

Luego, usando el método *REML* y las expresiones en (2.2) y (2.3), se obtienen las estimaciones $\widehat{\sigma}_b^2$, $\widehat{\boldsymbol{\beta}}$ y $\widehat{\mathbf{b}}$, respectivamente. De este modo, $\mathbf{V}$ se estima a través de $\widehat{\mathbf{V}} = \mathbf{Z}\widehat{\mathbf{G}}\mathbf{Z}' + \widehat{\mathbf{R}}$, donde $\mathbf{Z}\widehat{\mathbf{G}}\mathbf{Z}'$ es una matriz diagonal de bloques, formada por 296 de éstos. Cada bloque $\mathbf{B}_i \in \mathcal{M}_{(4\times4)}$, es una matriz completa de valores iguales a $\widehat{\sigma}_b^2$. La estimación del componente de varianza para los efectos aleatorios, $\widehat{\sigma}_b^2$, es igual a 136.6928, con un error estándar de estimación de 11.5985.

De lo anterior, el vector de componentes de varianza, $\boldsymbol{\theta}$, se estima a través de $\widehat{\boldsymbol{\theta}} = (\widehat{\sigma}_b^2, \widehat{\boldsymbol{\theta}}_1')'$, con $\widehat{\sigma}_b^2 = 136.6928$ y $\widehat{\boldsymbol{\theta}}_1$ como en la *Tabla 3.1.* Una vez explicitado $\widehat{\boldsymbol{\theta}}$, el modelo espacio-longitudinal de efectos mixtos asociado a la contaminación por ozono sobre España entre los años 2007 y 2010 queda definido. Los efectos fijos ajustados pueden consultarse en la *Tabla 3.2* y la varianza del proceso estimada corresponde a $\widehat{\sigma}_*^2 = 0.2791$.

TABLA 3.2. Efectos fijos ajustados para el modelo espacio-longitudinal de efectos mixtos.

Parámetro	Estimación	Std. Error	Estadístico	*p-valor*
β_0	164.6990	41.8321	3.9371	0.0000
β_1	2.6223	0.9457	2.7729	0.0055
β_2	0.9726	1.1124	0.8744	0.3819



(a) *2007* (b) *2008*

(c) *2009* (d) *2010*

FIGURA 3.4. Semivariogramas omnidireccionales, experimentales y ajustados, para la concentración de O_3 en España 2007-2010.

De lo anterior, el incremento de un grado de longitud implica un incremento de la concentración de ozono en 2.62 $\mu g/m^3$ hacia el oriente de España. Que la *latitud* deje de ser significativa se debe a que la información que se acumula en la matriz de covarianza que explica la dependencia espacial, absorve parte de la información de latitud, por lo que esa covariable deja de ser significativa.

Con el objetivo de validar el supuesto de multinormalidad sobre el modelo ajustado, se efectúa el *test* de Shapiro - Wilk multivariado sobre los errores obteniendo una estadística de prueba $W = 0.989$ y un *p-valor* de 0.2763. Así, existe evidencia estadística para garantizar el cumplimiento de dicho supuesto, pues la hipótesis de multinormalidad no es rechazada, empleando un nivel de significancia del 5 %.

Al igual que en la sección 2.5, es posible caracterizar el comportamiento medio de una estación de monitereo particular a través de los cuatro años considerados, haciendo uso del $BLUP$ presentado en (2.4).

Sin embargo, para realizar la predicción de la concentración de ozono para lugares en España diferentes a las estaciones de monitoreo, debe usarse el predictor definido en (3.2), tal como se verá en el *Capítulo 4*.

CAPÍTULO 4

Aplicación: comparación de predicciones de O_3 en España 2007-2010

El interés del presente capítulo recae en obtener predicciones de la concentración de ozono (O_3) en $\mu g/m^3$ sobre el territorio español, para lugares diferentes a las $n = 296$ estaciones de monitoreo de la calidad del aire tenidas en cuenta al estimar los modelos de las subsecciones 2.5 y 3.3. Para tal fin, se usarán los predictores definidos en (2.11) y en (3.2), junto con sus respectivos errores cuadráticos medios de predicción y los modelos estimados, de modo que las predicciones logradas por los mismos puedan ser comparadas.

Para realizar las predicciones, tanto del modelo logitudinal, como del espacio-longitudinal, se definió una red de 6615 puntos sobre España, situados a 0.2 unidades de *coordenadas* de distancia, tal como puede apreciarse en la *Figura 4.1.* Así, para cada punto de la red se obtuvo su predicción y su respectivo error de predicción, como se verá a continuación.

4.1. Predicciones de O_3 a través del LMM longitudinal

De lo desarrollado en la subsección 2.5, se sabe que el modelo *LMM* longitudinal quedó definido por $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$, donde

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 164.9861 \\ 0.5746 \\ -2.6662 \end{pmatrix}$$

y $\widehat{\mathbf{b}}$ se obtiene a través de (2.3). De este modo, el predictor de nuevas observaciones a emplear corresponde a $\widehat{Y}_{ij*} = \mathbf{x}'_{ij*}\widehat{\boldsymbol{\beta}}$, donde cada $\mathbf{x}'_{ij*}$ es un vector de la forma $\mathbf{x}'_{ij*} = (1, x_{i1}, x_{i2})$ de covariables asociadas a la *longitud* y *latitud* de las 6615 nuevas observaciones, $i = 1, \ldots, 6615$. Aplicando el predictor a cada nueva observación se obtiene el mapa de predicción para la concentración de ozono en España de la *Figura 4.2(a).*

De este modo, los niveles de ozono son más elevados hacia el sureste de España y de menor nivel hacia el noroeste del país, presentando una concentración máxima de alrededor de 65 $\mu g/m^3$. Así, ciudades como Málaga, Murcia, Valencia y las islas Baleares fueron los lugares de España con un mayor grado de contaminación en el aire debida al O_3 en el periodo 2007-2010, mientras que la región de Santiago de Compostela, con su abundante vegetación (ver *Fig. 2.2*) fue la menos afectada por dicha contaminación en el mismo periodo, registrando niveles de ozono cercanos a los 45 $\mu g/m^3$. El centro del país, esto es, cerca de la capital, registra una concentración de ozono del orden de los 55 $\mu g/m^3$.

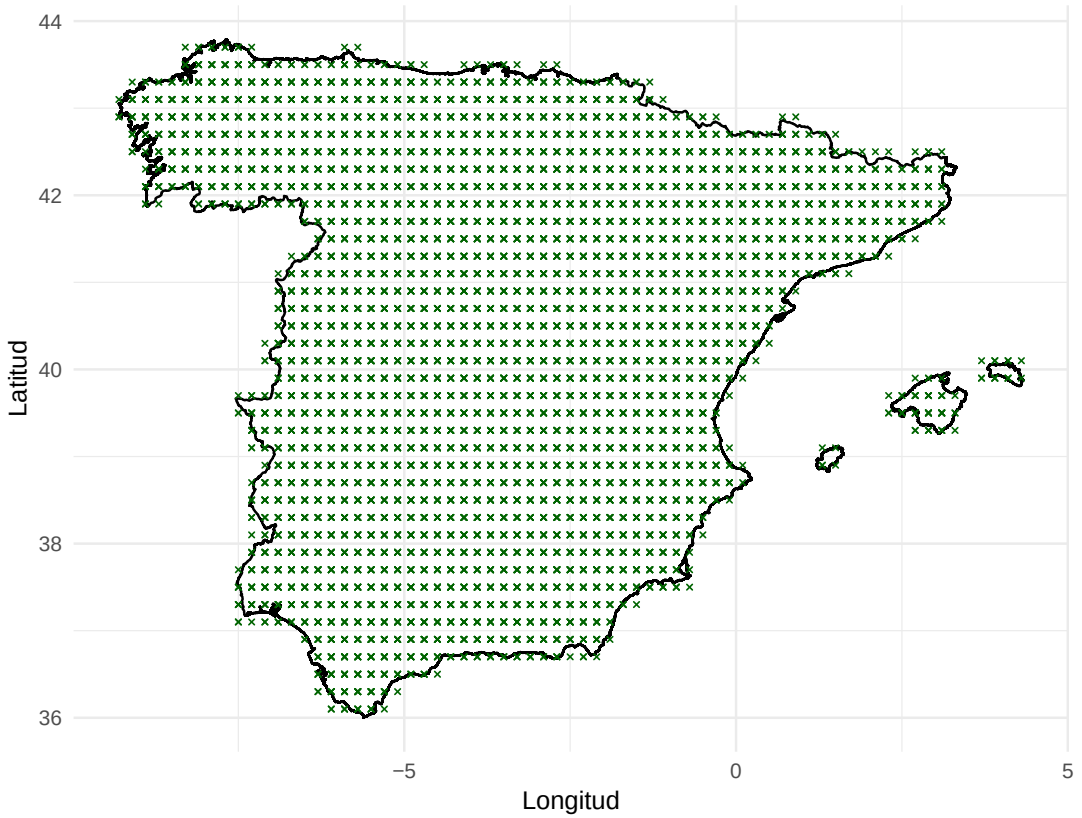


FIGURA 4.1. Localizaciones de los puntos sobre España para realizar predicción a través del modelo longitudinal y del espacio-longitudinal.

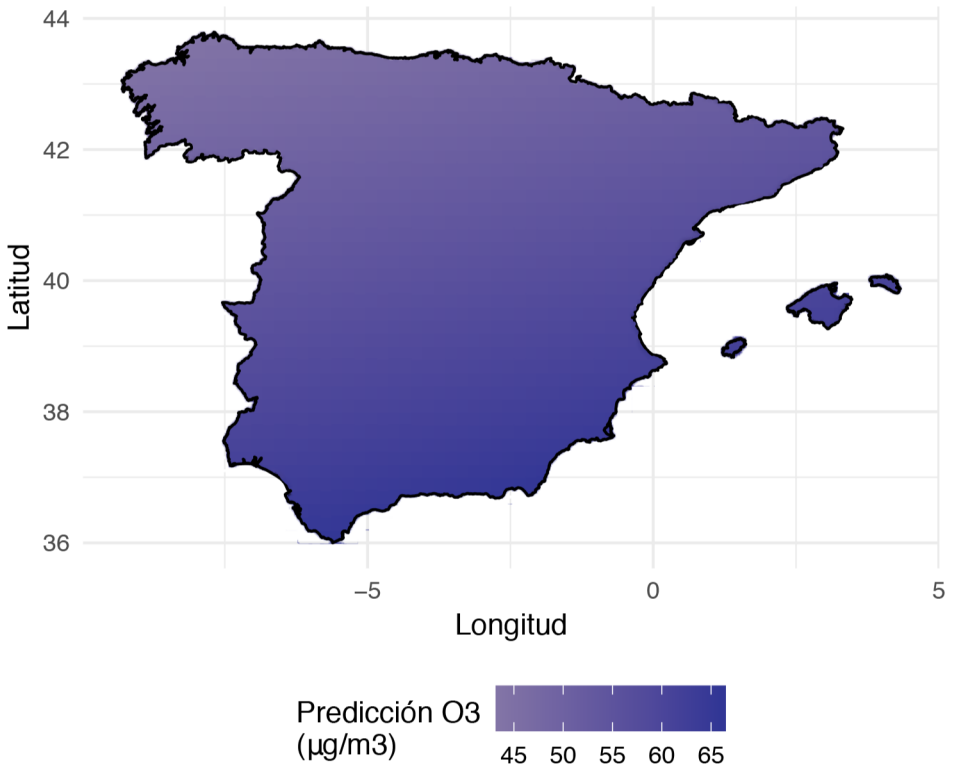
En general, la predicción señala niveles de ozono controlados, por debajo del límite máximo de 180 $\mu g/m^3$ estipulado por la *UE*. De este modo, los daños más graves sobre la población humana que pudieron presentarse fueron irritaciones iniciales de la conjuntiva ocular (ver *Tabla 2.2*).

Respecto al error cuadrático medio de predicción, éste fue calculado a través (2.12), sabiendo, por lo presentado en 2.5, que $\widehat{\sigma}^2_* = 27.5053$. Con los errores de predicción asociados a cada nueva observación, se construye el mapa presentado en la *Figura 4.2(b)*. De lo que se observa, el error de predicción se mantiene entre 5.25 $\mu g/m^3$ y 5.26 $\mu g/m^3$ sobre todo el territorio español, lo cual podría implicar una concentración un poco superior a los 70 $\mu g/m^3$, manteniéndose aún dentro de los límites normales.

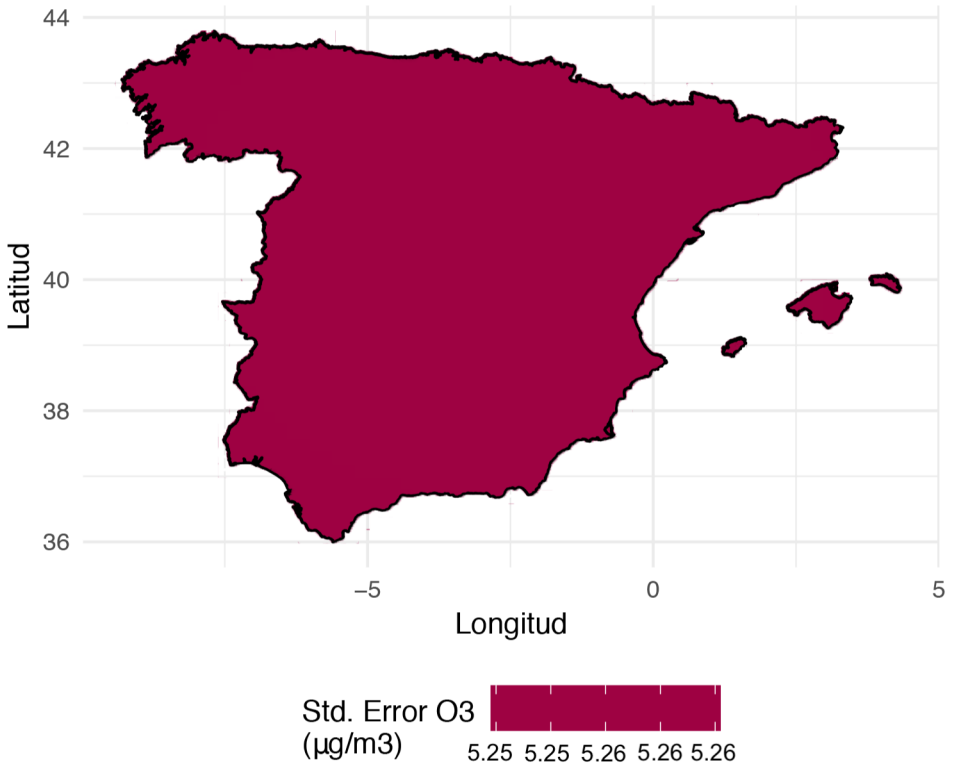
4.2. Predicciones de O_3 a través del LMM espacio - longitudinal

La estimación del modelo espacio-longitudinal tomó la forma $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$, donde $\widehat{\boldsymbol{\beta}}$ y $\widehat{\mathbf{b}}$ se obtuvieron desde (2.2) y (2.3), respectivamente, tal como fue presentado en la sección 3.3. Así, desde la *Tabla 3.2*, el vector de efectos fijos estimados fue

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 164.6990 \\ 2.6223 \\ 0.9726 \end{pmatrix};$$



(a) Predicción global empleando el modelo lineal mixto longitudinal.



(b) Error estándar de predicción empleando el modelo lineal mixto longitudinal.

FIGURA 4.2. Predicción y errores estándar de predicción asociados al predictor longitudinal.

que junto con el vector de efectos aleatorios estimado, $\widehat{\mathbf{b}}$, determinan por completo al *LMM* espacio-longitudinal.

Siguiendo a Militino et al. (2008), el predictor que permite conocer el comportamiento de la concentración de ozono estimada para las 6615 nuevas lozalizaciones a partir del modelo espacio-longitudinal ajustado, fue presentado en (3.2) y toma la forma

$$\widehat{Y}_{ij*} = \mathbf{Y}'[\widehat{\mathbf{V}}^{-1}\mathbf{r}'_{ij*} + \widehat{\mathbf{V}}^{-1}\mathbf{X}\mathbf{A}(\mathbf{x}_{ij*} - \mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{r}'_{ij*})];$$

con $\mathbf{A} = (\mathbf{X}'\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$; $\mathbf{r}_{ij*}$ el vector fila de covarianzas entre e_{ij*}, los errores de la nuevas observaciones, y el vector de los términos de error asociados a las 296 estaciones de monitoreo de la calidad del aire y $\mathbf{x}_{ij*}$ el vector con un uno en la primera posición y las coordenadas planas de ubicación para las nuevas observaciones, en *longitud* y *latitud*, en sus siguientes dos posiciones. De este modo, las predicciones fueron efectuadas, llegando a una distribución del ozono sobre España como la de la *Figura 4.3(a)*.

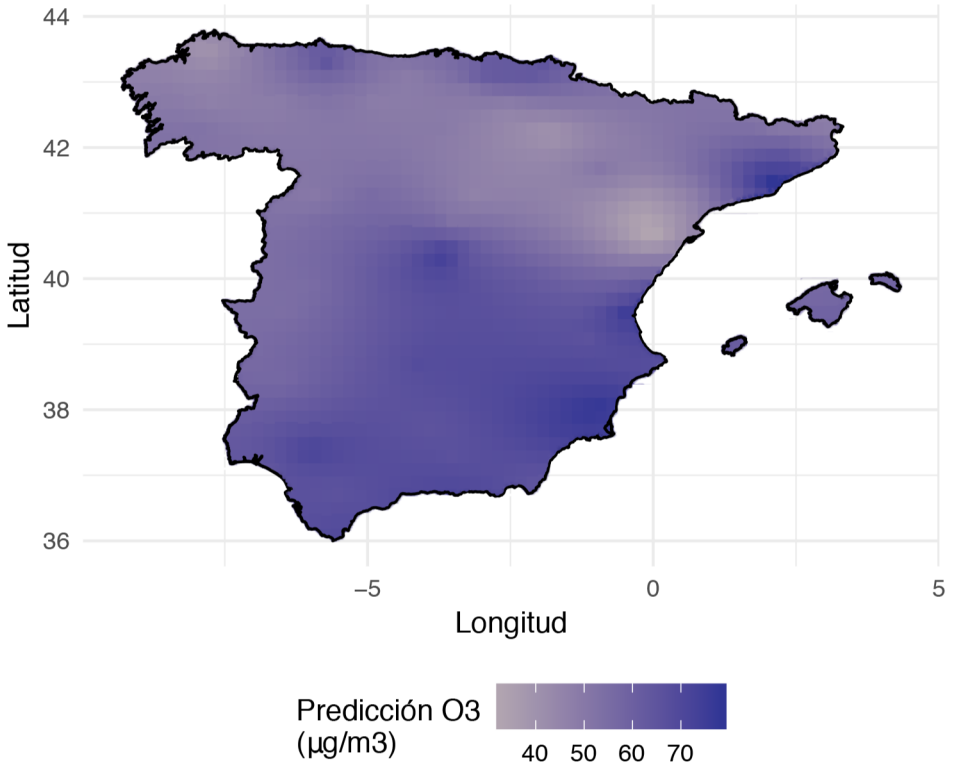
Tal como se observa, el predictor espacio-longitudinal logra una caracterización más detallada de la distribución del O_3 en España para el 2007-2010, presentando focos de mayor contaminación en ciudades principales como Madrid, Sevilla, Valencia, Murcia, Barcelona, Bilbao y la región de Oviedo, cuyos niveles de ozono se encontraron cercanos a los 80 $\mu g/m^3$. Se visualizan también zonas de baja contaminación, hacia el noreste y oeste del país, por las regiones de Castellón de la Plana, Santiago de Compostela, Valladolid y la ciudad de Zaragoza, localizaciones en las que hay presencia de vegetación y un menor número de vías principales de acceso; sus niveles de O_3 se encuentran por el orden de los 40 $\mu g/m^3$ (ver *Fig. 2.2*).

El error cuadrático medio de predicción asociado a cada nueva observación es calculado usando (3.3), con una varianza del proceso de $\widehat{\sigma}^2_* = 0.2791$, obteniendo como resultado el mapa de la *Figura 4.3(b)*. De este modo, se presentan errores de estimación de entre 4 $\mu g/m^3$ y 16 $\mu g/m^3$; el centro del país, que coincide con la localización de Madrid, se estima con el máximo error cuadrático medio de predicción, lo cual se debe al hecho de que aunque el ozono se degrade con mayor facilidad en presencia de otros contaminantes (Shirk, 2000), siguen presentándose niveles elevados del mismo en Madrid (ver *Fig. 4.3(a)*), luego existe una gran variabilidad entre lo que podría ser la concentración de ozono en el aire y su degradación al interactuar con otros contaminantes.

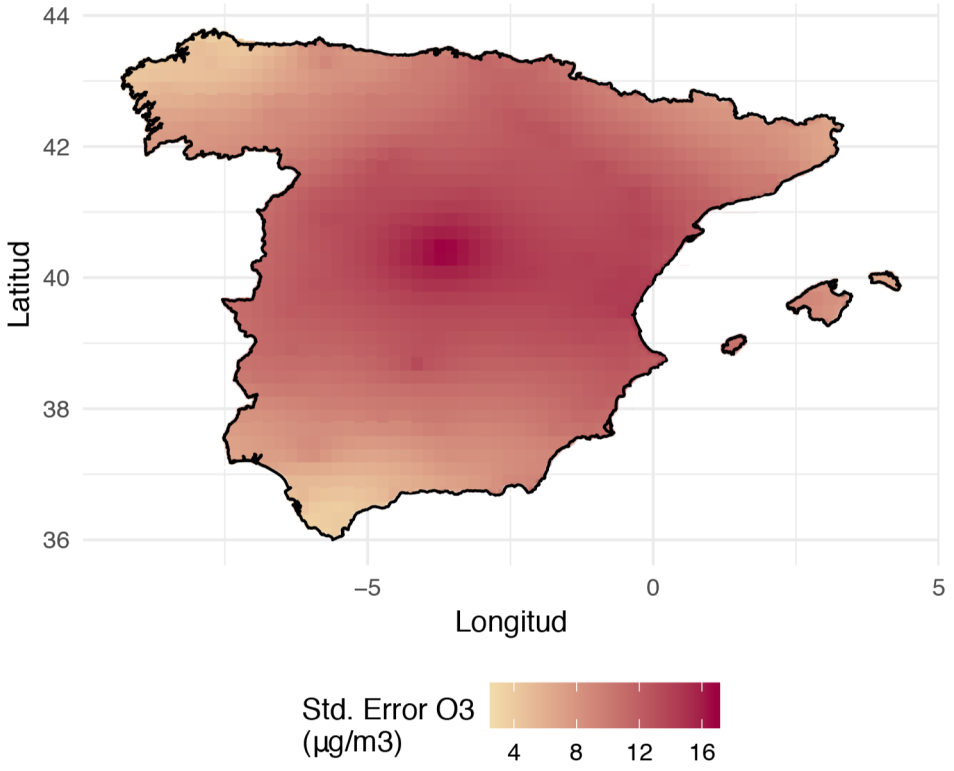
El predictor espacio-longitudinal sobreestima los niveles de contaminación por ozono, en comparación con las estimaciones alcanzadas por el predictor longitudinal, asociando también mayores errores de predicción a cada nueva localización. Sin embargo, ambos señalan niveles de O_3 dentro de los límites normales establecidos por la *UE*, además los posibles efectos de interactuar con aire contaminado por ozono a los niveles predichos por el predictor espacio-longitudinal, incluyen irrataciones iniciales de la conjuntiva ocular y probabilidad de jaquecas (ver *Tabla 2.2*).

4.3. Comparación de predictores

De lo presentado anteriormente, tanto para el predictor longitudinal, como para el espacio-longitudinal, se observan diferencias en los niveles medios de ozono predichos, así como en el error cuadrático medio de predicción generado. Sin embargo, con el objetivo de determinar cuál proporciona la mejor predicción, se realizará una comparación de los mismos a través del error cuadrático medio de predicción (*MSE*, en inglés), definido como $E\left[(\widehat{Y}_i - Y_i)^2\right]$ y el sesgo de predicción calculado a través de $E\left(\widehat{Y}_i - Y_i\right)$, únicamente para los registros de *AirBase* en el año 2011 y para estaciones de monitoreo de la calidad del aire diferentes a las contempladas en los ajustes de los modelos, tal como se presenta a continuación.



(a) Predicción global empleando el modelo lineal mixto espacio-longitudinal.



(b) Error estándar de predicción empleando el modelo lineal mixto espacio-longitudinal.

FIGURA 4.3. Predicción y errores estándar de predicción asociados al predictor espacio-longitudinal.

Metodología y resultados

De la información contenida en *AirBase*, se seleccionaron los registros asociados a la contaminación media de ozono en $\mu g/m^3$ para el año 2011 y en estaciones de monitoreo diferentes a las empleadas para el ajuste de modelos. El filtro arrojó 149 estaciones, cuya distribución sobre España, se presenta en la *Figura 4.4*.

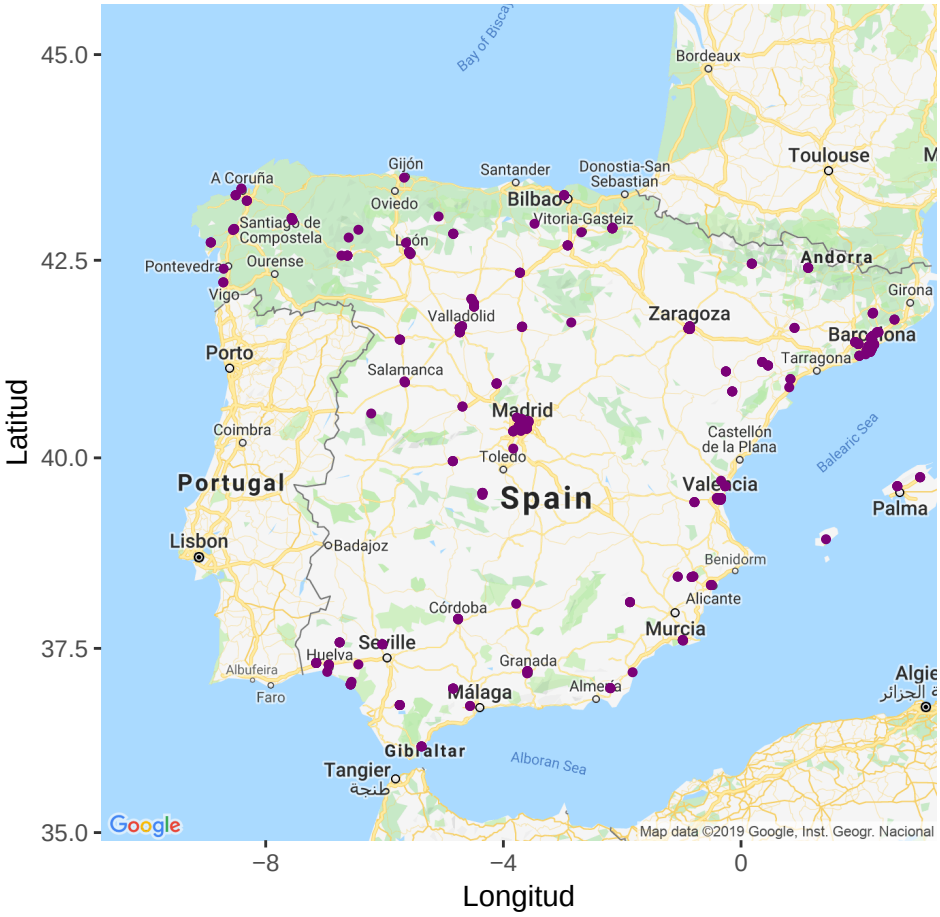


FIGURA 4.4. Configuración de la estaciones de monitoreo de la calidad del aire para comparación de predictores sobre el territorio español. Fuente: Google Maps 2019, `maps.google.com`.

Se realizaron las predicciones con la información de dichas estaciones de monitoreo, usando el predictor de (2.11) y el de (3.2), obteniendo los $\widehat{Y}_i$, $i = 1, \ldots, 149$ para cada predictor. Los valores de Y_i corresponden a los niveles promedio de ozono registrados por cada estación de la *Fig. 4.4* para el año 2011 y que fueron almacenados en *AirBase*.

Una vez explicitados los valores de $\widehat{Y}_i$ y empleando sus Y_i asociados, se calcula el sesgo de predicción y el *MSE* para ambos predictores, llegando a lo expuesto en la *Tabla 4.1*. A partir del sesgo de los predictores se observa que el predictor longitudinal subestima su predicción de ozono mucho más que el predictor espacio-longitudinal, sin embargo, el error cuadrático medio de predicción asociado con el predictor longitudinal es inferior que el error cuadrático medio de predicción para el modelo espacio-longitudinal.

Tener que estimar un mayor número de componentes de varianza en el modelo espacio-longitudinal puede estar generando el aumento en la variabilidad de predicción evidenciado en el *MSE* del predictor, pero si se define MSE_1 como el error cuadrático medio de predicción asociado al predictor longitudinal y MSE_2 el asociado al predictor espacio-longitudinal y se computa

TABLA 4.1. Comparación de predictores a través del sesgo de predicción y del *MSE* de predicción.

Predictor	Sesgo	*MSE*
Longitudinal	-0.0863	185.5511
Espacio - Longitudinal	-0.0086	233.8854

el cociente entre éstos, el resultado es

$$\frac{MSE_2}{MSE_1} = \frac{233.8854}{185.5511} = 1.26 \approx 1.$$

Luego, para definir cuál predictor es estadísticamente más eficiente al reducir significamente el *MSE* de predicción, deben estudiarse las distribuciones teóricas asociadas a los *MSE* de cada predictor y efectuar la prueba estadística adecuada.

Conclusiones

Dos metodologías para la representación y predicción de procesos longitudinales y espacio-longitudinales fueron presentadas. La primera de éstas hizo uso de la teoría de los modelos lineales mixtos, para explicar un proceso con presencia de correlación temporal y la segunda, también empleó el enfoque de los *LMM*, pero contemplando, además de la correlación temporal, la dependencia espacial inherente al proceso de estudio.

Se definieron dos predictores, junto con sus respectivos errores cuadráticos medios de predicción, implementando la predicción de los niveles promedio de contaminación por ozono en España, para el periodo 2007-2010. Además, se realizó una comparación de predictores, usando la información asociada a los niveles de ozono promedio para el año 2011.

De lo desarrollado se puede concluir:

- El predictor espacio-longitudinal identifica focos de contaminación elevada, esto es, niveles promedio de ozono de alrededor de 80 $\mu g/m^3$ en las principales ciudades y baja contaminación en zonas naturales o con poca malla vial, registrando niveles de ozono promedio de alrededor de los 30 $\mu g/m^3$; mientras que el predictor longitudinal, a pesar de ser el resultado de ajustar un modelo cuyos parámetros fueron significativos y de mínimo AIC, sólo permite ver la dirección en la que la polución aumenta o disminuye, siendo está mayor hacia el sur del país y superior hacia el noroeste del mismo.
- Aunque el ozono se degrada con mayor facilidad ante la presencia de otros contaminantes del aire (Shirk, 2000), sus niveles son elevados en ciudades principales como Madrid; de este modo, el predictor espacio-longitudinal logra identificar la variabilidad entre los verdaderos niveles de contaminación por ozono y la posible degradación de la misma debida a otros contaminantes, a través de su error cuadrático medio estándar de predicción.
- En general, los niveles promedio de ozono se mantuvieron dentro de los límites de especificación regulados por la *UE*, siendo inferiores a los 180 $\mu g/m^3$ para el periodo 2007-2010. Sin embargo, la exposición a la máxima concentración de ozono predicha, esto es, alrededor de los 80 $\mu g/m^3$, pudo generar irritaciones iniciales de la conjuntiva ocular y probabilidad de jaquecas. (Shirk, 2000)
- Para el modelo longitudinal ajustado la *latitud* resulta ser una covariable significativa, mientras que para el caso del modelo espacio-longitudinal no lo es. Esto se debe a que al modelar la estructura de dependencia espacial, ésta también explica la información asociada a la coordenada de *latitud*, incluyendo su comportamiento en la matriz de covarianza del modelo.
- El error estándar del predictor espacio-longitudinal toma valores sobre el territorio español, de acuerdo a la estructura de dependencia modelada para las 296 estaciones de monitoreo de la calidad del aire inicialmente estudiadas; mientras que para el predictor longitudinal, al asumir independencia entre estaciones de medición, éste sólo toma en cuenta la estructura de covarianza de las medidas repetidas, de este modo, el mapa del error estándar asociado al predictor longitudinal se comporta de manera constante para toda España. Sin embargo, este

último hecho representa un inconveniente ya que este modelo puede enmascarar la verdadera variabilidad del error en determinadas zonas de interés, lo que hace que la predicción obtenida a través del modelo espacio-longitudinal presente mayor consistencia en el error y tenga una mejor aproximación a la realidad.

- Los sesgos de predicción, tanto para el predictor longitudinal, como para el espacio-longitudinal son pequeños; en este sentido, los predictores no subestiman ni sobreestiman la predicción del valor real. Sin embargo, se presenta una mayor subestimación por parte del predictor longitudinal, que por parte del predictor espacio-longitudinal, para el cual se obtuvo un menor sesgo de predicción.
- El error cuadrático medio de predicción fue superior para el predictor espacio-longitudinal. Tal aumento se debe a que se introducen componentes de varianza en el error del predictor, al modelar la dependencia espacial inherente al proceso de estudio.
- La razón entre los errores cuadráticos medios de predicción asociados a ambos predictores fue aproximadamente de 1, lo cual implica que la determinación del mejor predictor (el que minimiza el *MSE* de predicción), involucra un desarrollo teórico adicional, pues deben establecerse las distribuciones teóricas de ambos *MSE* y realizar la prueba estadística adecuada. Así, de lo observado, la predicción lograda por ambos predictores resulta adecuada, pero de mayor especificidad para el predictor espacio-longitudinal.
- La inclusión de covariables adicionales a la longitud y la latitud, como la temporada del año, ayudaría a obtener mejores predicciones de la concentración de ozono sobre España.
- En términos generales, la aplicación de cualquiera de los modelos ajustados arrojaría resultados similares en cuanto a predicción, sin embargo, usar el predictor espacio-longitudinal otorga la ventaja de detectar focos de mayor concentración de la variable en observación, mientras que el predictor longitudinal sólo permite evidenciar la dirección de aumento o disminución de la misma variable.

Trabajo futuro

Para el ajuste del modelo longitudinal y del espacio-longitudinal se asumió normalidad sobre la variable respuesta, sobre los errores y sobre el vector de efectos aleatorios, así como un conjunto de datos balanceados, esto es, con información para todos los tiempos de medición y en todos las unidades de estudio. Podría desarrollarse una metodología que generalice lo aquí presentado a conjuntos de datos desbalanceados y para los que el supuesto de normalidad no se tiene, a través, por ejemplo, de los *modelos lineales mixtos generalizados* (*GLMM*, por sus siglas en inglés).

También podría pensarse en un enfoque bayesiano de la aplicación, siguiendo, por nombrar alguno, el trabajo desarrollado por Cepeda (2011).

Podría pensarse además en desarrollar la forma teórica de las distribuciones asociadas a los errores cuadráticos medios del predictor longitudinal y del predictor espacio-longitudinal, de modo que pueda establecerse, a través de la prueba estadística adecuada, cuál predictor es significativamente más eficiente, al ser el de menor error cuadrático medio de predicción.

Por último, con el objetivo de obtener mejores predicciones, deberían incluirse covariables adicionales a la longitud y la latitud, que ayuden a caracterizar con mayor detalle el proceso de estudio.

Bibliografía

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle,[w:] proceedings of the 2nd international symposium on information, bn petrow, f, *Czaki, Akademiai Kiado, Budapest* .

Angulo, J., Gonzalez-Manteiga, W., Febrero-Bande, M. & Alonso, F. (1998). Semi-parametric statistical approaches for space-time process prediction, *Environmental and Ecological Statistics* **5**(4): 297–316.

Bi, J., Xu, T., Chen, C.-M. & Johannesen, J. (2015). Spatio-temporal modeling of eeg data for understanding working memory, *ICML Workshop on Statistics, Machine Learning and Neuroscience (Stamlins 2015)*.

Bogaert, P. & Christakos, G. (1997). Spatiotemporal analysis and processing of thermometric data over belgium, *Journal of Geophysical Research: Atmospheres* **102**(D22): 25831–25846.

Bohorquez, C. M. (2009). Estadística espacial, *Notas de Clase. Departamento de Estadística, Universidad Nacional de Colombia* .

Cepeda, E. (2011). Generalized spatio-temporal models, *SORT: statistics and operations research transactions* **35**(2): 0165–178.

Chamberlain, G. (1984). Panel data. In Z. griliches and M. D. Intriligator (eds.), *Handbook of econometrics* **2**: 1247–1318.

Chen, Q. (2013). Spatial-temporal modeling of active layer thickness. *Department of Statistics The George Washington University*.

Clark, I. (1979). *Practical geostatistics*, Vol. 3, Applied Science Publishers London.

Cressie, N. A. (1993). *Statistics for spatial data: Wiley series in probability and mathematical statistics*.

Cressie, N. & Majure, J. J. (1997). Spatio-temporal statistical modeling of livestock waste in streams, *Journal of Agricultural, Biological, and Environmental Statistics* pp. 24–47.

Davidian, M. (2019). ST 732, Longitudinal Data Analysis, Spring 2019 - North Carolina State University.
URL: *https://www4.stat.ncsu.edu/ davidian/st732/*

Davis, C. S. (2002). *Statistical methods for the analysis of repeated measurements*, Springer Science & Business Media.

de Castro González, F. V. (2012). *La contaminación en España: los efectos del ozono y del cambio climático*, Editorial Club Universitario.

Design, G. (1995). Gs+: Geostatistical software for the agronomic and biological sciences, *Plainwell, Michigan* .

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S. et al. (2002). *Analysis of longitudinal data*, Oxford University Press.

Diggle, P. & Ribeiro Jr, P. (2007). Model based geostatistics springer, *New York* .

Duong, Q. P. (1984). On the choice of the order of autoregressive models: a ranking and selection approach, *Journal of Time Series Analysis* **5**(3): 145–157.

Edward Vonesh, V. M. C. (1996). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*, Statistics: Textbooks and Monographs, har/dskt edn, Marcel Dekker.
URL: *http://gen.lib.rus.ec/book/index.php?md5=71E2CDD9E0915B853547059B2CAC6452*

EEA (2019). European environment agency. AirBase - The European air quality database.
URL: *https://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-7#tab-figures-produced*

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation, *The American Statistician* **37**(1): 36–48.

Fernández-Casal, R., González-Manteiga, W. & Febrero-Bande, M. (2003). Flexible spatio-temporal stationary variogram models, *Statistics and Computing* **13**(2): 127–136.

Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. (2008). *Longitudinal data analysis*, CRC Press.

Giraldo, H. R. (2011). Estadística espacial, *Notas de Clase. Departamento de Estadística, Universidad Nacional de Colombia* .

Giraldo, R. (2002). Introducción a la geoestadística: Teoría y aplicación, *Bogotá: Universidad Nacional de Colombia* .

Gneiting, T. (2002). Compactly supported correlation functions, *Journal of Multivariate Analysis* **83**(2): 493–508.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model, *Journal of the American Statistical Association* **57**(298): 369–375.

Gotway, C. A. & Cressie, N. (1993). Improved multivariate prediction under a general linear model, *Journal of multivariate analysis* **45**(1): 56–72.

Gotway, C. A. & Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction, *Journal of Agricultural, Biological, and Environmental Statistics* pp. 157–178.

Henderson, C. R. (1953). Estimation of variance and covariance components, *Biometrics* **9**(2): 226–252.

Homish, G. G., Edwards, E. P., Eiden, R. D. & Leonard, K. E. (2010). Analyzing family data: a gee approach for substance use researchers, *Addictive behaviors* **35**(6): 558–563.

Isaaks, E. H. & Srivastava, R. M. (1989). An introduction to applied geostatistics, *Technical report*, Oxford university press.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices., *Biometrics* **42**(4): 805–820.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*, Springer Science & Business Media.

Jones, R. H. (1993). *Longitudinal data with serial correlation: a state-space approach*, Chapman and Hall/CRC.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand, *Journal of the Southern African Institute of Mining and Metallurgy* **52**(6): 119–139.

Kyriakidis, P. C. & Journel, A. G. (1999). Geostatistical space–time models: a review, *Mathematical geology* **31**(6): 651–684.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22.
URL: *http://dx.doi.org/10.1093/biomet/73.1.13*

Mardia, K. V., Goodall, C., Redfern, E. J. & Alonso, F. J. (1998). The kriged kalman filter, *Test* **7**(2): 217–282.

Matheron, G. (1963). Principles of geostatistics, *Economic geology* **58**(8): 1246–1266.

McCullagh, P. (1983). Quasi-likelihood functions, *The Annals of Statistics* pp. 59–67.

Militino, A. (1998). Strategies for dynamic space-time statistical modeling: discussion of "the kriged kalman filter"by mardia et al.

Militino, A. F., Ugarte, M. D. & Ibáñez, B. (2008). Longitudinal analysis of spatially correlated data, *Stochastic Environmental Research and Risk Assessment* **22**(1): 49–57.

Minasny, B. & McBratney, A. B. (2005). The matérn function as a general model for soil variograms, *Geoderma* **128**(3-4): 192–207.

Molina, J. J. C. (2015). La presión atmosférica y los vientos en la Península Ibérica. Reflexiones sobre el Monzón Ibérico, *NIMBUS nº 4* **4**(4): 5–60.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Longitudinal analysis of spatially correlated data, *Journal of the Royal Statistical Society, Series A* (135): 370–384.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**(3): 545–554.

Petitgas, P. (1996). Geostatistics and their applications to fisheries survey data, *Computers in fisheries research*, Springer, pp. 113–142.

Samper, F. & Carrera, J. (1990). Geoestadística, *Aplicaciones a la Hidrogeología Subterránea. Centro Internacional de Métodos Numéricos en Ingeniería. Universitat Politécnica de Catalunya. Barcelona* .

SAS-Institute (1999). *SAS Procedures Guide: Version 8*, Vol. 1, Sas Inst.

Schabenberger, O. & Gotway, C. A. (2005). *Statistical methods for spatial data analysis*, CRC press.

Schwarz, G. (1978). Estimating the dimension of a model, *The annals of statistics* **6**(2): 461–464.

Shirk, O. (2000). Las mediciones del ozono, *Dräger Sicherheitstechnik GmbH. Dräger Hispania Mapfre Seguridad* **77**: 17–21.

Stein, M. L. (2005). Space–time covariance functions, *Journal of the American Statistical Association* **100**(469): 310–321.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region, *Economic geography* **46**(sup1): 234–240.

Warrick, A. & Myers, D. (1987). Optimization of sampling locations for variogram calculations, *Water Resources Research* **23**(3): 496–500.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**(3): 439–447.