Universidade Vigo

Escuela Internacional de Doctorado

Sergio Alberto Castillo Páez

TESIS DOCTORAL

Aportaciones a la Geoestadística no paramétrica

Dirigida por los doctores:

Pilar García Soidán y Rubén Fernández Casal

Año:

2017

Índice general

Prefacio

A lo largo de las últimas décadas, la estadística espacial se ha consolidado como un importante campo de investigación sobre todo debido a sus múltiples aplicaciones. Aunque en sus inicios el estudio de la predicción espacial estuvo dedicado a la minería y meteorología (ver Cressie, 1990, para más detalles), actualmente sus avances y desarrollos prácticos abarcan distintas áreas, por ejemplo en agricultura (Steere *et al.*, 2016), contaminación ambiental (Antunes y Albuquerque, 2013), salud pública (Hanna-Attisha *et al.*, 2016), finanzas (Da Barrosa *et al.*, 2016), procesamiento de imágenes (Chiang *et al.*, 2014), entre otras.

La Estadística Espacial proporciona métodos para el análisis de la información contenida en datos espaciales, es decir, en datos asociados a las posiciones geográficas en las que han sido recogidos. Estos datos espaciales pueden corresponder a tres distintos tipos de procesos (ver p.e. Gelfand *et al.*, 2010): (1) procesos espaciales continuos (o geoestadísticos), (2) procesos espaciales discretos (o *lattices*) y, (3) procesos o patrones puntuales. El presente documento se centrará exclusivamente en el primer caso, en el cual la variable aleatoria de estudio está definida sobre un dominio espacial continuo.

Si bien la modelización de estos procesos geoestadísticos se puede realizar de distintas maneras, a lo largo de este estudio se considera la siguiente descompo-

sición en escalas de variación espacial:

$$\begin{aligned} \text{variación en datos espaciales} &= \text{variación a gran escala} \\ &\quad + \text{variación de pequeña escala} \end{aligned}$$

donde el primer término se encuentra relacionado con la media o tendencia del proceso, la cual está presente en toda la región de estudio, mientras que la variación a pequeña escala recoge la dependencia espacial. Sería de esperar que los valores observados entre dos localizaciones cercanas fuesen próximos entre sí y que su efecto disminuyese conforme su separación aumenta. Este tipo de descomposición es similar a la que se presenta en (Cressie, 1993, Sección 3.1, p. 113), en la cual se podría considerar adicionalmente un proceso de ruido debido, entre otros factores, a errores de medida (en ese caso, el objetivo sería realizar inferencias sobre el proceso libre de ruido).

Para un proceso geoestadístico bajo el modelo anteriormente presentado, la caracterización de su tendencia espacial guarda estrecha relación con la estimación de la función de regresión bajo el supuesto de errores correlacionados. Por tanto, varios resultados teóricos y técnicas disponibles para este tipo de modelos pueden ser extendidas al caso espacial. Sin embargo, y de forma similar a lo que sucede en el caso de la estimación de la función de regresión, la aproximación no paramétrica permite obtener estimaciones más flexibles de la tendencia, evitando problemas de mala especificación de modelos paramétricos. Los métodos no paramétricos se consideran además valiosas herramientas de analisis exploratorio que reflejan de mejor forma el comportamiento de la información muestral, brindando adecuadas estimaciones piloto para el posterior ajuste de modelos de mayor complejidad (Härdle, 1990, pp. 7-14). Sin embargo, estas estimaciones no paramétricas dependen de una apropiada selección de ventanas de suavizado y

no están exentas de sesgos debido a la presencia de la correlación espacial.

De manera similar, es factible obtener estimaciones no paramétricas de la dependencia espacial. El procedimiento habitual de estimación recurre al uso directo de residuos, los cuales se obtienen al eliminar de los datos muestrales el efecto de la tendencia estimada. Sin embargo, es un hecho conocido que la variabilidad de estos residuos subestiman la variabilidad de pequeña escala, y por tanto los estimadores derivados de estos presentan sesgos respecto a su correspondiente dependencia teórica.

Frente a las limitaciones anteriores, en este trabajo se introducen varias aportaciones relacionadas a la estimación no paramétrica de procesos geoestadísticos, tomando como punto de partida una ligera modificación del método de corrección del sesgo del estimador de la dependencia espacial basado en residuos, propuesto por Fernández-Casal y Francisco-Fernández (2014). A partir de este proceso de corrección se pueden obtener estimaciones conjuntas no paramétricas de la tendencia y de la dependencia espacial. El uso del estimador corregido del variograma influye notablemente en las inferencias sobre el proceso, por ejemplo en los criterios de la selección de la ventana para la estimación de la tendencia. A partir de esta misma idea, se diseñó un método bootstrap no paramétrico el cual reproduce de manera adecuada la variabilidad espacial, permitiendo realizar inferencias sobre el proceso geoestadístico. La versatilidad de esta nueva técnica permitió su aplicación en la construcción de mapas de riesgo. Finalmente, este enfoque de corrección de sesgos se extendió al caso de procesos geoestadísticos heterocedásticos, planteándose un nuevo método de estimación conjunta no paramétrica para este tipo de modelos.

Este documento se encuentra organizado de la siguiente manera:

Capítulo 1. Introducción a la geoestadística. Dentro de este capítulo se exponen los fundamentos teóricos comunes y las técnicas tradicionales que se

utilizan para el estudio de los procesos geoestadísticos, así como las distintas aproximaciones paramétricas para la caracterización de la tendencia y dependencia espacial, y sus correspondientes limitaciones que motivaron el presente estudio.

Capítulo 2. Estimación no paramétrica de la tendencia espacial. En este capítulo, nos centraremos en la estimación lineal local de la tendencia bajo correlación espacial, para lo cual se presenta un método de estimación conjunta no paramétrica de la tendencia y la dependencia espacial, basado en el método de corrección de sesgo debido al uso de residuos. Luego, se proponen nuevos criterios para la selección de la ventana para la estimación de la tendencia. El comportamiento de estos selectores se analiza mediante estudios de simulación, y se verifica su utilidad mediante su aplicación en datos reales. Parte de los resultados que se describen en este capítulo se han incluido en el trabajo realizado por Castillo-Páez *et al.* (2017a).

Capítulo 3. Métodos bootstrap para datos espaciales. En las primeras secciones de este capítulo se revisan algunas de las técnicas bootstrap disponibles para el estudio de datos espaciales, en especial bajo la presencia de una tendencia espacial determinística. Luego, se introduce un nuevo método bootstrap no paramétrico, y se comprueba mediante simulación su buen funcionamiento en la aproximación de la variabilidad de diversos estimadores utilizados para la caracterización de la dependencia espacial. Algunos estudios sobre el comportamiento de este método bootstrap propuesto se presentan en el trabajo de Castillo-Páez *et al.* (2017b). A continuación, se realiza una aplicación del método propuesto para la construcción de mapas de riesgo. El comportamiento de estos métodos para la estimación de la probabilidad incondicional fueron analizados mediante estudios numéricos. Finalmente, se aplicaron las técnicas propuestas en este capítulo a un conjunto de datos de mediciones de la precipitación total mensual en diversas localizaciones de EEUU. Las principales aportaciones propuestas en este capítulo

para la construcción de mapas de riesgo han sido publicadas en el trabajo de Fernández-Casal *et al.* (2017a).

Capítulo 4. Estimación no paramétrica de procesos espaciales heterocedásticos. Un elemento común en los modelos considerados en los capítulos anteriores, es la hipótesis de estacionariedad del proceso de error subyacente. Sin embargo, en ciertos casos esta hipótesis no es razonable. Por ejemplo en el caso espacio-temporal es habitual que la variabilidad cambie con el tiempo. En este capítulo se considera la estimación en procesos geoestadísticos heterocedásticos, en los cuales aparece una nueva componente que debe ser aproximada, como es la función varianza. Con este objetivo, en las primeras secciones se analizan las características de este tipo de procesos, y en especial, los métodos no paramétricos usuales para estimar la función varianza. Siguiendo la idea del método de corrección del sesgo debido a los residuos (descrita en el Capítulo 2), se propone un nuevo método iterativo para la estimación conjunta de la tendencia y de la variabilidad de pequeña escala. Para verificar la validez de esta técnica, se realizaron distintos estudios de simulación además de su aplicación a un conjunto de datos reales. Algunos de los resultados indicados anteriormente han sido publicados en Fernández-Casal *et al.* (2017b).

Esta memoria finaliza con la presentación de las conclusiones generales de los distintas propuestas presentadas a lo largo de este trabajo, así como con algunas sugerencias y posibles líneas de trabajo futuro derivadas de los resultados obtenidos en cada uno de los capítulos de esta memoria.

Capítulo 1

Introducción a la Geoestadística

En el presente capítulo, se realiza en primer lugar una breve revisión de los fundamentos teóricos que rigen a los procesos geoestadísticos, en especial en lo referente a la definición y propiedades del semivariograma (Sección 1.2). Luego, en la Sección 1.3 se presentan los métodos clásicos de predicción espacial (técnicas kriging). Posteriormente, en la Sección 1.4 se presenta el procedimiento habitual para modelar la dependencia espacial en procesos estacionarios. Se hace mención especial a los estimadores no paramétricos y a los modelos flexibles de variograma. La Sección 1.5 trata acerca de la inferencia en el contexto de procesos espaciales cuando se admite la presencia de una función tendencia determinística. Se presenta el método paramétrico tradicional basado en residuos para caracterizar la dependencia espacial en estos casos, y se evidencia la presencia de problemas circulares entre la estimación de la tendencia y el variograma, así como el efecto de sesgos debido al uso directo de estos residuos.

1.1. Un ejemplo introductorio

La geoestadística tiene como objetivo primordial el estudio de variables aleatorias definidas sobre un dominio espacial continuo. Desde el punto de vista práctico, esto significa que los datos muestrales de la variable de estudio pueden ser obtenidos en cualquier localización dentro de la región espacial de observación. Tomaremos como ejemplo un conjunto de datos muy conocido en el contexto espacial, relacionado a mediciones de contaminación por metales pesados en la capa superficial de las riberas del río Meuse, ubicado en la localidad de Stein (Países Bajos) (Burrough y McDonnell, 1998). Esta base de datos se encuentran disponible en el paquete estadístico `gstat` del software `R` (Pebesma, 2004). Específicamente, en este ejemplo se consideraron las mediciones de concentración de zinc (medidas en ppm) tomados en 155 ubicaciones espaciales a lo largo de la ribera del río, tal como se muestra en la Figura 1.1.

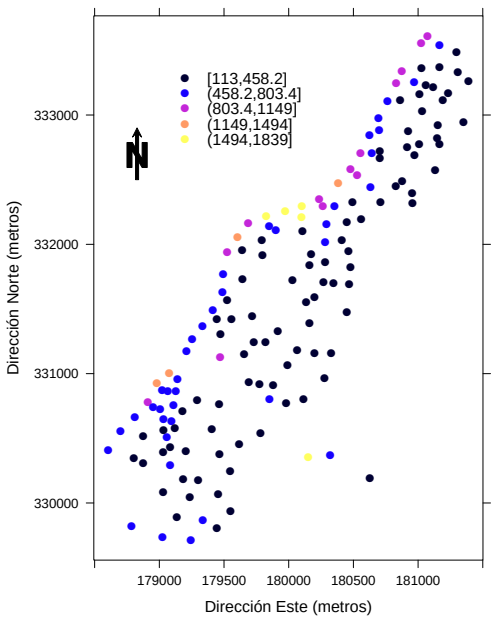
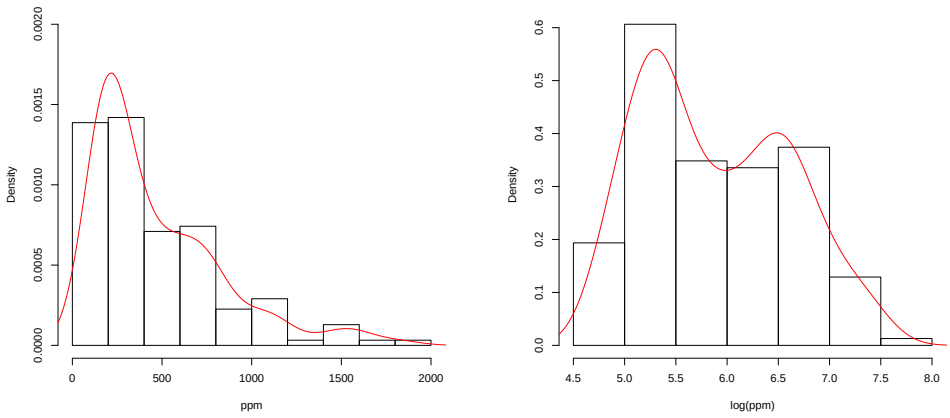


Figura 1.1: Distribución espacial de la concentración de zinc (en ppm) medida en 155 posiciones espaciales ubicadas en las riberas del Río Meuse.

Un análisis descriptivo básico de estos datos nos permite verificar que es necesario realizar una transformación de los mismos para que estos sean aproximadamente simétricos (ver Figura 1.2(a)). En este caso, el efecto de realizar una transformación logarítmica de los datos se puede observar en el histograma de la Figura 1.2(b).



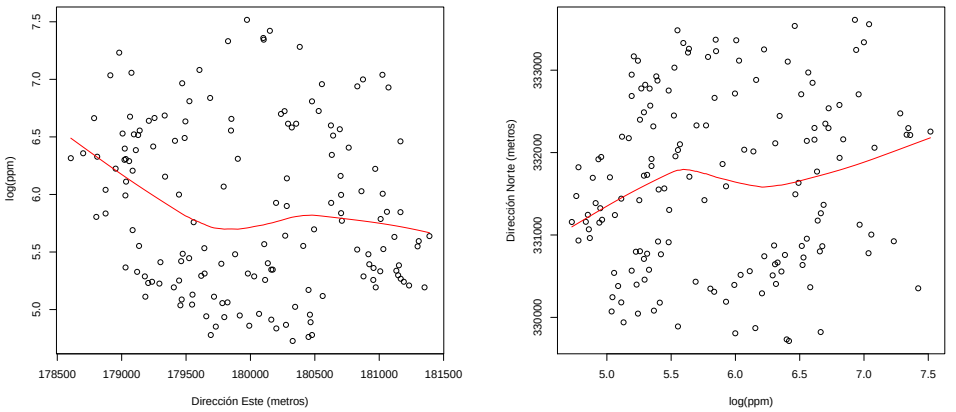
(a) (b)

Figura 1.2: Histograma de los datos de concentración de zinc en escala (a) original y (b) logarítmica, y curvas de estimación no paramétrica de la densidad (en color rojo).

Es factible suponer que la concentraciones de zinc medidas en ubicaciones cercanas sean similares entre sí, y difieran de las mediciones realizadas en zonas más distantes. Se puede apreciar de forma exploratoria si la variable de estudio presenta algún tipo de dependencia con su posición espacial, mediante gráficos de dispersión que relacionen la concentración de zinc y las coordenadas geográficas, tal como se muestran en las Figuras 1.3(a) y 1.3(b). Aquí se observa que la concentración de zinc presenta un comportamiento respecto el eje X (Dirección Este) que resulta diferente cuando se analiza considerando la dirección Norte.

Para tratar de caracterizar la dependencia espacial en las mediciones de concentración de zinc, se puede recurrir a técnicas geoestadísticas. Para esto, es necesario asumir ciertas hipótesis sobre el proceso aleatorio subyacente, en especial

aquellas relacionadas a la estacionariedad del mismo.



(a) (b)

Figura 1.3: Gráfico de dispersión de la concentración de zinc (en log(ppm)) respecto a la dirección (a) Este y (b) Norte y curvas de regresión no paramétrica (en color rojo).

1.2. Procesos geoestadísticos estacionarios

Un proceso geostadístico (o variable regionalizada) se define como un proceso aleatorio definido sobre un dominio espacial. De manera más formal, si consideramos una región D con volumen positivo dentro del espacio d-dimensional $\mathbb{R}^d$, entonces el proceso geostadístico se representará como el conjunto de variables aleatorias $\{Y(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d\}$, siendo $\mathbf{x}$ cada una de las localizaciones espaciales dentro de la región de observación D.

Luego, si se consideran n localizaciones $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de observación, entonces el conjunto $\{y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n)\}$ corresponde a una realización del conjunto de variables aleatorias $\{Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n)\}$.

Consideraremos de ahora en adelante, la siguiente notación:

- *Función de distribución de probabilidad:* $\mathbb{P}\left[Y(\mathbf{x}_1) \leq y_1, \ldots, Y(\mathbf{x}_n) \leq y_n\right] = F_{\mathbf{x}_1,\ldots,\mathbf{x}_n}(y_1, \ldots, y_n)$.

- *Media o tendencia del proceso espacial:* $\mathbb{E}[Y(\mathbf{x})] = \mu(\mathbf{x})$.
- *Varianza:* $Var[Y(\mathbf{x})] = \mathbb{E}\left[(Y(\mathbf{x}) - \mu(\mathbf{x}))^2\right] = \sigma^2(\mathbf{x})$.
- *Covarianza:* $Cov[Y(\mathbf{x}_i), Y(\mathbf{x}_j)] = \mathbb{E}[(Y(\mathbf{x}_i) - \mu(\mathbf{x}_i))(Y(\mathbf{x}_j) - \mu(\mathbf{x}_j))]$.

En el análisis de datos espaciales se suele recurrir a ciertas hipótesis de estacionariedad, algunas de las cuales se introducen a continuación.

1.2.1. Tipos de procesos estacionarios

Un proceso geostadístico se dice que es *estrictamente estacionario* (o fuertemente estacionario), si su función de distribución es invariante a cualquier traslación respecto a un vector o salto $\mathbf{u}$, es decir:

$$F_{\mathbf{x}_1+\mathbf{u},\ldots,\mathbf{x}_n+\mathbf{u}}(y_1,\ldots,y_n) = F_{\mathbf{x}_1,\ldots,\mathbf{x}_n}(y_1,\ldots,y_n), \text{ para todo } \mathbf{x}_1,\ldots\mathbf{x}_n \in D,$$
$$\forall \mathbf{u} \in \mathbb{R}^d, \forall n \in \mathbb{N}.$$

Por otra parte, en los procesos geostadísticos se suele suponer que los momentos de primer y segundo orden existen. Esto permite definir un tipo de estacionariedad menos restrictiva que la anterior, en función de los momentos de ambos órdenes.

Un proceso $Y(\cdot)$ se dice que es *estacionario de segundo orden*, cuando:

- $\mathbb{E}[Y(\mathbf{x})] = \mu, \forall \mathbf{x} \in D$.
- $Cov[Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})] = C(\mathbf{u}), \forall \mathbf{u} \in D$.

La función $C(\cdot)$ se denomina *covariograma* o función de covarianza, la cual depende exclusivamente del vector de salto $\mathbf{u}$ y no de la localización espacial $\mathbf{x}$. En ciertos casos, el covariograma puede depender únicamente de la magnitud del salto, y no de su dirección, es decir, $C(\mathbf{u}) = C(\|\mathbf{u}\|)$. Cuando esto sucede, se

dice que el covariograma es *isotrópico*, y en caso contrario se trata de un proceso estacionario de segundo orden *anisotrópico*.

A partir del covariograma y suponiendo $C(0) > 0$ es factible definir el *correlograma* como:

$$\rho(\mathbf{u}) = \frac{C(\mathbf{u})}{C(0)} \in [-1, 1]\,.$$

Este tipo de estacionariedad implica que la varianza del proceso está definida, es finita y no depende de la posición espacial $\mathbf{x}$, pues se verifica que:

$$Var\left[Y(\mathbf{x})\right] = C(0), \forall \mathbf{x} \in D.$$

Por otra parte existen procesos aleatorios en los que la varianza no está definida, pero donde sus incrementos o diferencias tienen varianza finita. Estos procesos se incluyen en una clase más general, los procesos *intrínsecos o intrínsicamente estacionarios*, que se caracterizan por:

- $\mathbb{E}\left[Y(\mathbf{x}+\mathbf{u}) - Y(\mathbf{x})\right] = 0, \forall \mathbf{x} \in D.$
- $Var\left[Y\left(\mathbf{x}\right) - Y\left(\mathbf{x}+\mathbf{u}\right)\right] = 2\gamma(\mathbf{u}), \forall \mathbf{u} \in D.$

La función $2\gamma(\cdot)$ se denomina *variograma*, y $\gamma(\cdot)$ recibe el nombre de *semivariograma*. A lo largo de la presente memoria (y abusando de la notación), nos referiremos a ambas funciones como *variograma*. De forma similar al covariograma, el variograma depende exclusivamente del vector $\mathbf{u}$ y se considera isotrópico si depende solo de la magnitud de dicho salto, es decir, cuando $\gamma(\mathbf{u}) = \gamma(\|\mathbf{u}\|)$. De lo contrario, cuando también depende de la dirección, se trataría de un proceso intrínseco anisotrópico.

Todo proceso estrictamente estacionario con momentos de primer y segundo orden finitos es a su vez estacionario de segundo orden. Luego, en el caso de procesos gaussianos ambas propiedades son equivalentes. Por otra parte, la clase

de los procesos intrínsicamente estacionarios es más general que la clase de los procesos estacionarios de segundo orden. Si el proceso $Y(\cdot)$ es estacionario de segundo orden con covariograma $C(\cdot)$, entonces:

$$\begin{aligned} Var\left[Y\left(\mathbf{x}\right) - Y\left(\mathbf{x}+\mathbf{u}\right)\right] &= Var\left[Y\left(\mathbf{x}\right)\right] + Var\left[Y\left(\mathbf{x}+\mathbf{u}\right)\right] \\ &\quad -2Cov\left[Y\left(\mathbf{x}\right), Y\left(\mathbf{x}+\mathbf{u}\right)\right] \\ &= 2C(0) - 2C(\mathbf{u}). \end{aligned}$$

Luego, se cumple que $\gamma(\mathbf{u}) = C(0) - C(\mathbf{u})$. Además, si se considera que $\sigma^2 = C(0)$, entonces:

$$\gamma(\mathbf{u}) = \sigma^2 - C(\mathbf{u}).$$

Sin embargo, la relación recíproca no puede garantizarse. Por ejemplo, como se pone de manifiesto en Cressie (1993, Sección 2.3.2), si $Y(\mathbf{x})$ es un movimiento Browniano isotrópico $d-$dimensional, su variograma correspondiente es $Var\left[Y\left(\mathbf{x}\right) - Y\left(\mathbf{x}+\mathbf{u}\right)\right] = ||\mathbf{u}||$, $\mathbf{u} \in \mathcal{R}^d$ (y por tanto el variograma no es acotado). En ese caso se tiene que $Cov\left[Y\left(\mathbf{x}\right), Y\left(\mathbf{x}+\mathbf{u}\right)\right] = \frac{1}{2}\left(||\mathbf{x}|| + ||\mathbf{x}+\mathbf{u}|| - ||\mathbf{u}||\right)$, que no depende exclusivamente del salto $\mathbf{u}$.

En ocasiones, la estacionariedad intrínseca solo se verifica de forma local, de modo que las hipótesis relativas a la media y al variograma son válidas únicamente en un entorno de cada localización espacial de la región de observación. Este tipo de procesos reciben el nombre de *quasi-intrínsecos*.

Por otra parte, consideramos que la hipótesis de estacionariedad de segundo orden no es muy restrictiva en la práctica, ya que en general se puede suponer que hay independencia a partir de una distancia mayor que el máximo de las distancias consideradas en el análisis. Por ejemplo, se pueden seleccionar dos variogramas, uno acotado y otro no, que toman los mismos valores hasta un determinado salto (como por ejemplo, un semivariograma lineal y un semivariograma lineal con um-

bral - también denominado modelo triangular - con la misma pendiente y efecto nugget, ver p.e. Chilès y Delfiner, 2012, Sección 2.5). En ese caso, las predicciones kriging utilizando ambos modelos coincidirán si no es necesario evaluar el semivariograma a distancias mayores que el rango del variograma acotado (ver p.e. Chilès y Delfiner, 2012, Sección 4.6.2).

También podría ocurrir que no fuese realista suponer ningún tipo de estacionariedad, como en los casos donde no es posible asumir que la media es constante en todas las localizaciones espaciales (modelos no estacionarios en media). En cambio, si es el covariograma (o variograma) quien difiere de forma significativa al cambiar la posición espacial, entonces se puede recurrir a los modelos con estructura de dependencia no estacionaria. En la presente memoria se abordará la caracterización de funciones asociadas a procesos estacionarios y no estacionarios.

Finalmente, cabe mencionar que los conceptos de variograma y covariograma (así como algunos de los métodos que se presentan en los capítulos posteriores), tienen relación con ciertas nociones y técnicas desarrolladas para el estudio de series temporales. Por ejemplo, la función variograma fue introducida en este último contexto como *diferencias de medias cuadradas* por Jowett (1952), mientras que el concepto de covariograma (correlograma) también es conocido en el caso temporal como *función de autocovarianza* (*autocorrelación*) (ver p.e. Cressie, 1993, Sec. 2.3.2). Otra analogía importante está relacionada con el concepto de *ergodicidad*. De manera similar a lo que sucede en el análisis de series de tiempo, en la práctica se dispone únicamente de una realización parcial del proceso espacial $Y(\mathbf{x})$ para poder realizar inferencias sobre parámetros de interés de dicho proceso. Para esto, se suele recurrir a las propiedades de ergodicidad, las cuales han sido generalizadas al caso espacial. Estas propiedades garantiza la convergencia en media cuadrática de los promedio muestrales respecto a sus correspondientes valores teóricos. Existen distintos tipos de ergodicidad en el contexto espacial,

como la ergodicidad en media, en la covarianza o en el variograma, y también se han determinado condiciones necesarias y suficientes para que estas propiedades se cumplan (ver p.e. Yaglom, 1986, Cap. 3, Secciones 16 y 17 para mayores detalles).

1.2.2. Propiedades del variograma y covariograma

La condición necesaria y suficiente para que la función $C(\cdot)$ sea el covariograma de un proceso estacionario de segundo orden, es que sea semidefinida positiva:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{x}_i - \mathbf{x}_j) \geq 0, \forall n \in \mathbb{N}, \forall \mathbf{x}_1, \ldots, \mathbf{x}_n \in D, \forall a_1, \ldots, a_n \in \mathbb{R}.$$

Adicionalmente, el covariograma $C(\cdot)$ tiene las siguientes propiedades:

- $C(0) = Var(\mathbf{u}) \geq 0$.
- $C(\mathbf{u}) = C(-\mathbf{u}), \forall \mathbf{u} \in \mathbb{R}^d$ (función simétrica).

En un proceso intrínsicamente estacionario, una condición necesaria (más no suficiente) para que la función $\gamma(\cdot)$ sea un semivariograma, es que sea condicionalmente semidefinida negativa:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\mathbf{x}_i - \mathbf{x}_j) \leq 0, \forall n \in \mathbb{N}, \forall \mathbf{x}_1, \ldots, \mathbf{x}_n \in D, \forall a_1, \ldots, a_n \in \mathbb{R};$$

tales que $\sum_{i=1}^{m} a_i = 0$. Luego, el semivariograma $\gamma(\cdot)$ cumple las siguientes propiedades:

- $\gamma(0) = 0$.
- $\gamma(\mathbf{u}) \geq 0, \forall \mathbf{u} \in \mathbb{R}^d$ (función no negativa).
- $\gamma(\mathbf{u}) = \gamma(-\mathbf{u}), \forall \mathbf{u} \in \mathbb{R}^d$ (función simétrica).

El variograma presenta ciertas características geométricas importantes que pueden ser de interés a la hora de realizar su estimación o en el posterior ajuste de un modelo paramétrico. Las más importantes se comentan a continuación.

- **Efecto nugget:** La forma del variograma cerca del origen es de especial interés pues está relacionada con la continuidad y la regularidad espacial del proceso. Aunque el variograma siempre es nulo en el origen, existen casos en los que existe discontinuidad conforme los saltos se hacen más pequeños. Esto se conoce como *efecto nugget o pepita* y se define como $c_0 = \lim_{\|\mathbf{u}\|\to 0} \gamma(\mathbf{u})$. Este efecto se suele producir por varias causas: errores de medida, escala espacial del muestreo, variabilidad a microescala, entre otros.

  Por otra parte, existen casos en los que el variograma puede ser modelado por un *efecto nugget puro*, es decir, que la dependencia permanece constante en todos los saltos considerados. Esto puede ser debido a que los datos son efectivamente independientes, o debido a que la dependencia espacial se manifiesta a una escala mucho menor que la considerada en el diseño de muestreo. Cabe indicar, que la correcta estimación del variograma cerca del origen es de vital importancia a la hora de realizar predicciones del proceso espacial (Stein, 1988).

- **Umbral:** Si el variograma está acotado y además su límite existe, entonces este valor recibe el nombre de **umbral**. De manera más formal, el umbral viene dado por:

$$\sigma^2 = \lim_{\|\mathbf{u}\|\to\infty} \gamma(\mathbf{u}).$$

  Si el proceso $Y(\cdot)$ es un proceso estacionario de segundo orden tal que $\lim_{\|\mathbf{u}\|\to\infty} C(\mathbf{u}) = 0$, entonces $\sigma^2 = C(0)$. El comportamiento del variograma

en saltos grandes puede servir para detectar la presencia de "deriva" o tendencia espacial (ver p.e. Armstrong, 1998, pp. 27-28).

- **Umbral parcial:** Cuando un variograma presenta efecto nugget, la diferencia $c_1 = \sigma^2 - c_0$ se denomina *umbral parcial*. De manera general, se podría considerar al parámetro c_1 como una medida del grado de dependencia espacial presente en los datos

- **Rango:** Si el umbral existe se puede definir el rango del semivariograma $\gamma(\mathbf{u})$ en la dirección $\mathbf{e}_0 = \mathbf{u}_0/||\mathbf{u}_0||$ como:

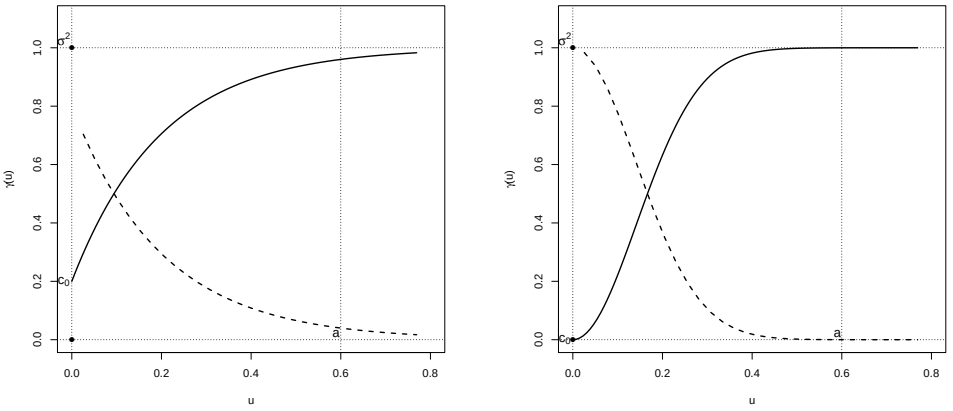
$$a = min\,\{u : \gamma(u(1+\epsilon)\mathbf{e}_0) = \sigma^2, \forall \epsilon > 0\}.$$

  Cabe mencionar que el umbral no siempre existe, por tanto es factible tener variogramas no acotados (ver comentarios en la sección anterior). Asimismo, el rango no siempre es el mismo en todas las direcciones, como sucede en los variogramas anisotrópicos (Armstrong, 1998, pp. 26).

  Cuando el variograma alcanza el umbral de forma asintótica, se suele redefinir a como el *rango práctico*, correspondiente a la distancia en la cual el valor del variograma alcanza el 95 % del umbral parcial.

- **Anisotropía:** Si la forma del variograma $\gamma(\mathbf{u})$ cambia dependiendo de la dirección del salto $\mathbf{u}$, se dice que el variograma es *anisotrópico*. Existen casos en que la anisotropía se puede corregir mediante una transformación lineal de las coordenadas del salto $\mathbf{u}$ (anisotropía geométrica). En ese caso, los variogramas direccionales tienen el mismo umbral, pero distintos rangos. Otro caso se presenta cuando el variograma depende solamente de alguna dirección o componente del salto (anisotropía zonal o estratificada). Lo habitual en esta situación es descomponer el variograma en una parte isotrópica más otro variograma que actúa únicamente en dicha dirección. Ejemplos más de-

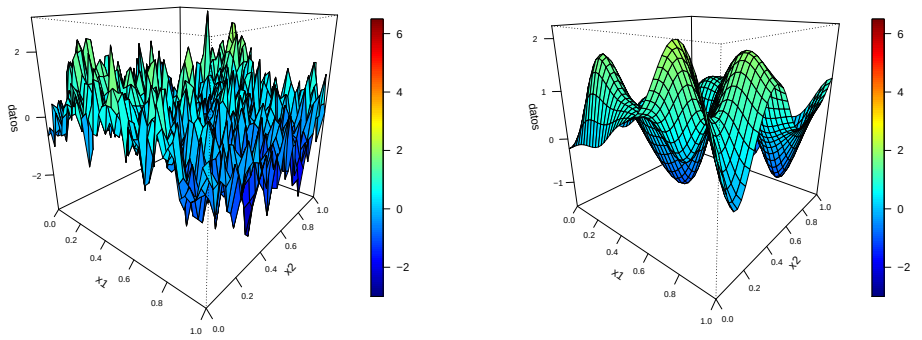
tallados sobre anisotropía se pueden encontrar en Goovaerts (1997, Sección 4.2.2) y Chilès y Delfiner (2012, Sección 2.5.2).



(a) (b)

Figura 1.4: Ejemplos de variogramas isotrópicos (líneas continuas) y covariogramas correspondientes (líneas discontinuas) (a) exponencial con efecto nugget $c_0 = 0{,}2$ y (b) gaussiano con $c_0 = 0$, ambos con $\sigma^2 = 1$ y $a = 0{,}6$ y $c_0 = 0{,}2$

Un resumen detallado de las propiedades y características del variograma y covariograma, se puede encontrar en Fernández-Casal (2003, Cap. 2). Las Figuras 1.4(a) y 1.4(b) presentan dos variogramas isotrópicos (líneas contínuas) con parámetros $\sigma^2 = 1$, $a = 0{,}6$, correspondientes a un modelo exponencial con efecto nugget $c_0 = 0{,}2$ y a un modelo gaussiano con efecto nugget nulo, respectivamente (en la Sección 1.4.2 se detallan estos modelos de variograma). Los covariogramas relativos a cada uno de ellos se muestran mediante líneas discontinuas. Estos variogramas comparten ciertas características comunes, por ejemplo en ambos casos la varianza del proceso estacionario es la misma ($\sigma^2 = 1$), y se puede considerar que observaciones ubicadas a una distancia mayor que 0.6 son incorreladas entre sí ($a = 0{,}6$). Sin embargo, estas figuras también presentan algunas diferencias significativas. En el primer variograma, se evidencia la discontinuidad en el origen, lo cual no sucede en el caso del variograma con nugget nulo. El comportamiento



(a) (b)

Figura 1.5: Errores estacionarios de media cero simulados en $D = [0,1]^2$ con $\gamma(u)$ isotrópico (a) exponencial y $c_0 = 0{,}2$ y (b) gaussiano y $c_0 = 0$.

cerca del origen difiere además por el modelo de variograma utilizado. De hecho, se muestra que el grado de dependencia es más fuerte en valores muy cercanos entre sí para el caso de variograma exponencial. A medida que el salto u aumenta, el primer variograma tarda en alcanzar el umbral correspondiente, a diferencia de lo que sucede con el variograma gaussiano, en el cual la curva del variograma crece lentamente en saltos u cercanos a cero, para luego alcanzar rápidamente valores cercanos a la varianza.

Estos efectos se muestran también en las Figuras 1.5(a) y 1.5(b) en las cuales se representan $n = 40 \times 40$ datos simulados con cada variograma anterior, en una rejilla regular bidimensional definida sobre en la región $D = [0,1]^2 \subset \mathbb{R}^2$. Aquí se observa que aunque ambos procesos tienen la misma media y varianza, el segundo proceso es más "suave" respecto al primero, debido precisamente a las distintas estructuras de dependencia consideradas.

1.3. Predicción espacial

Supongamos que el vector $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^t$ representa una muestra de n observaciones de un proceso espacial $Y(\cdot)$ en las posiciones espaciales muestrales $\mathbf{x}_1, \ldots, \mathbf{x}_n$, y se desea obtener la predicción de dicho proceso en la localización no observada $\mathbf{x}_0$, denotada por $\hat{Y}(\mathbf{x}_0)$.

Los métodos de predicción espacial o mejor conocidos como *métodos kriging* son algoritmos que permiten construir predictores lineales óptimos teniendo en cuenta la estructura de segundo orden del proceso espacial $Y(\cdot)$ (ver p.e. Cressie, 1990). Estos predictores son lineales pues se obtienen como una combinación lineal de las observaciones muestrales:

$$\hat{Y}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i Y(\mathbf{x}_i) + \lambda_0. \tag{1.1}$$

Para que este predictor sea uniformemente insesgado, los valores $\lambda_i \in \mathbb{R}$ (también llamados *pesos kriging*) se asignan de forma que la media del error de predicción sea cero, es decir:

$$\mathbb{E}\left[\hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_0)\right] = 0. \tag{1.2}$$

El predictor kriging es óptimo en el sentido de que, además de ser insesgado, este trata de minimizar el error de predicción en media cuadrática:

$$\mathbb{E}\left[\left(\hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_0)\right)^2\right]. \tag{1.3}$$

Los métodos kriging suponen de forma general que el proceso espacial $Y(\cdot)$

admite una descomposición de la siguiente forma:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1.4}$$

donde, $\mu(\mathbf{x})$ representa la tendencia determinística espacial, y el proceso de error $\varepsilon(\mathbf{x})$ es un proceso estacionario (de segundo orden o intrínseco) de media cero. Dependiendo de la suposición que se haga sobre la media del proceso, se distinguen tres método principales:

- *Kriging Simple*(KS): Cuando $\mu(\mathbf{x})$ es conocida.
- *Kriging Ordinario*(KO): Cuando $\mu(\mathbf{x})$ es desconocida pero constante.
- *Kriging Universal*(KU): Cuando $\mu(\mathbf{x})$ es desconocida no constante pero que se puede expresar como una combinación lineal de funciones conocidas.

Considerando que el método KO puede entenderse como un caso particular del Kriging Universal, a continuación se revisan brevemente los métodos KS y KU. Además, aunque anteriormente se ha supuesto que tanto el semivariograma como el covariograma dependen del salto, esta consideración no es necesaria para los métodos kriging, y por tanto, se utilizará una notación más general no estacionaria (la cual también será de utilidad en capitulos posteriores):

$$C(\mathbf{x}_1, \mathbf{x}_2) = Cov(Y(\mathbf{x}_1), Y(\mathbf{x}_2)), \;\; 2\gamma(\mathbf{x}_1, \mathbf{x}_2) = Var(Y(\mathbf{x}_1) - Y(\mathbf{x}_2))$$

1.3.1. Kriging Simple

Suponiendo que el proceso espacial puede ser modelado por (1.4) y que la función tendencia $\mu(\mathbf{x})$ es conocida para toda la región espacial de interés, el KS trata de construir un predictor óptimo de la forma (1.1). Para que este predictor

sea insesgado, la condición (1.2) se traduce como:

$$\mathbb{E}\left[\sum_{i=1}^{n}\lambda_i Y(\mathbf{x}_i)+\lambda_0-Y(\mathbf{x}_0)\right]=0,$$

de donde se deduce que $\lambda_0=\mu(\mathbf{x}_0)-\sum_{i=1}^{n}\lambda_i\mu(\mathbf{x}_i)$.

Entonces, el predictor KS tiene la forma:

$$\hat{Y}(\mathbf{x}_0)=\mu(\mathbf{x}_0)+\sum_{i=1}^{n}\lambda_i(Y(\mathbf{x}_i)-\mu(\mathbf{x}_i))=\mu(\mathbf{x}_0)+\sum_{i=1}^{n}\lambda_i\varepsilon(\mathbf{x}_i). \tag{1.5}$$

Los pesos kriging λ_i, se obtienen de forma que minimicen el error cuadrático medio dado por (1.3), también denominado *varianza kriging*:

$$\begin{aligned}\sigma^2_{KS} &= \mathbb{E}\left[\left(\mu(\mathbf{x}_0)+\sum_{i=1}^{n}\lambda_i\varepsilon(\mathbf{x}_i)-Y(\mathbf{x}_0)\right)^2\right]\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(\mathbf{x}_i,\mathbf{x}_j)-2\sum_{i=1}^{n}\lambda_i C(\mathbf{x}_i,\mathbf{x}_0)+C(\mathbf{x}_0,\mathbf{x}_0).\end{aligned}$$

Calculando las derivadas parciales de la última expresión respecto a los pesos kriging e igualando a cero, se obtiene la siguiente ecuación matricial:

$$\boldsymbol{\Sigma\lambda}=\mathbf{c},$$

siendo $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_n)^t$, $\mathbf{c}=(C(\mathbf{x}_1,\mathbf{x}_0),\ldots,C(\mathbf{x}_n,\mathbf{x}_0))^t$ y $\boldsymbol{\Sigma}$ es la matriz de orden $n\times n$ de varianzas y covarianzas de los datos, tal que $\Sigma_{ij}=C(\mathbf{x}_i,\mathbf{x}_j)$. En base a esta notación, el predictor kriging simple (1.5) se obtiene de la siguiente manera:

$$\hat{Y}_{KS}(\mathbf{x}_0)=\mu(\mathbf{x}_0)+\mathbf{c}^t\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu}), \tag{1.6}$$

siendo $\boldsymbol{\mu}=(\mu(\mathbf{x}_1),\ldots,\mu(\mathbf{x}_n))^t$ el vector de tendencias conocido.

Luego, la varianza kriging en este caso es igual a:

$$\sigma^2_{KS}(\mathbf{x}_0) = C(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{c}^t \boldsymbol{\Sigma}^{-1} \mathbf{c}. \tag{1.7}$$

Suponiendo normalidad, es factible construir intervalos de predicción a partir de las ecuaciones (1.6) y (1.7):

$$\left(\hat{Y}_{KS}(\mathbf{x}_0) \pm z_{(1-\alpha)/2} \cdot \sigma_{KS}(\mathbf{x}_0) \right)$$

1.3.2. Kriging Universal

En este caso, se considera que la tendencia del proceso espacial $Y(\cdot)$ puede expresarse como una combinación lineal de funciones conocidas:

$$\mu(\mathbf{x}) = \sum_{j=0}^{p} f_j(\mathbf{x}) \beta_j, \tag{1.8}$$

donde el vector de coeficientes $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^t \in \mathbb{R}^{p+1}$ es desconocido. Se supondrá además que $f_0(\cdot) = 1$, lo cual permite expresar las ecuaciones en función del variograma (y permite definir al KO como un caso particular del kriging universal). De forma matricial, se puede escribir:

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.9}$$

siendo $\boldsymbol{\varepsilon} = (\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n))^t$ y $\boldsymbol{X}$ es una matriz de orden $n \times (p+1)$ con $X_{ij} = f_{j-1}(\mathbf{x}_i)$; además:

$$Y(\mathbf{x}_0) = \boldsymbol{x}^t \boldsymbol{\beta} + \varepsilon(\mathbf{x}_0),$$

donde $\boldsymbol{x} = (f_0(\mathbf{x}_0), \ldots, f_p(\mathbf{x}_0))^t$. En este caso, el predictor (1.1) debe satisfacer la condición (1.2):

$$\mathbb{E}\left[\sum_{i=1}^{n} \lambda_i Y(\mathbf{x}_i) + \lambda_0 - Y(\mathbf{x}_0)\right] = \boldsymbol{\lambda}^t \boldsymbol{X}\boldsymbol{\beta} + \lambda_0 - \boldsymbol{x}^t\boldsymbol{\beta} = 0.$$

Luego, una condición necesaria y suficiente para que el predictor KU sea uniforme insesgado es que:

$$\boldsymbol{\lambda}^t \boldsymbol{X} = \boldsymbol{x}^t,\ \lambda_0 = 0. \tag{1.10}$$

Lo anterior implica que si $f_0(\cdot) = 1$, entonces $\sum_{i=1}^{n} \lambda_i = 1$, que es la misma condición sobre los pesos en el caso del kriging ordinario. Luego, los pesos kriging se determinan minimizando el error cuadrático medio de predicción (1.3) sujeto a las restricciones (1.10), para lo cual se recurre al método de los multiplicadores de Lagrange. Igualando a cero las derivadas parciales de la expresión resultante, se obtiene la siguiente ecuación matricial:

$$\boldsymbol{\Gamma}_U \boldsymbol{\lambda}_U = \boldsymbol{\gamma}_U,$$

con:

$$\boldsymbol{\Gamma}_U = \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{X} \\ \boldsymbol{X}^t & 0 \end{pmatrix},\ \boldsymbol{\lambda}_U = \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{m} \end{pmatrix},\ \boldsymbol{\gamma}_U = \begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{x} \end{pmatrix},$$

donde $\boldsymbol{\gamma} = (\gamma(\mathbf{x}_1, \mathbf{x}_0), \ldots, \gamma(\mathbf{x}_n, \mathbf{x}_0))^t$, $\mathbf{m} = (m_0, \ldots, m_p)^t$, y $\boldsymbol{\Gamma}$ es una matriz de orden $n \times n$ con $\Gamma_{ij} = \gamma(\mathbf{x}_i, \mathbf{x}_j)$. Luego, los pesos kriging así como la varianza de predicción del kriging universal se obtienen mediante:

$$\boldsymbol{\lambda}_U = \boldsymbol{\Gamma}_U^{-1} \boldsymbol{\gamma}_U,$$

$$\sigma^2_{KU}(\mathbf{x}_0) = \boldsymbol{\lambda}_U^t \boldsymbol{\gamma}_U.$$

1.3.3. Relación entre kriging simple y kriging universal

Si se supone que en el modelo (1.9) el vector $\boldsymbol{\beta}$ es conocido (es decir, suponiendo KS), el predictor lineal óptimo (1.6) se reescribiría de la siguiente manera:

$$\hat{Y}_{KS}(\mathbf{x}_0) = \mathbf{c}^t\boldsymbol{\Sigma}^{-1}\mathbf{Y} + (\boldsymbol{x} - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{c})^t\boldsymbol{\beta}.$$

Sin embargo, si $\boldsymbol{\beta}$ es desconocido, se puede recurrir a su estimador lineal óptimo obtenido por mínimos cuadrados generalizados $\hat{\boldsymbol{\beta}}_{mcg} = (\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{Y}$, obteniéndose el predictor:

$$\hat{Y}^*(\mathbf{x}_0) = \mathbf{c}^t\boldsymbol{\Sigma}^{-1}\mathbf{Y} + (\boldsymbol{x} - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{c})^t\hat{\boldsymbol{\beta}}_{mcg},$$

que coincide con el predictor del KU. Además, como:

$$\hat{Y}_{KU}(\mathbf{x}_0) = \hat{Y}_{KS}(\mathbf{x}_0) + (\boldsymbol{x} - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{c})^t(\hat{\boldsymbol{\beta}}_{mcg} - \boldsymbol{\beta}),$$

entonces:

$$\sigma^2_{KU}(\mathbf{x}_0) = \sigma^2_{KS}(\mathbf{x}_0) + (\boldsymbol{x} - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{c})^t(\boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{x} - \boldsymbol{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{c}),$$

donde el segundo término cuantifica la precisión en la estimación de la tendencia. Estas relaciones de aditividad permiten obtener la predicción espacial bajo tendencia desconocida mediante un proceso de dos etapas: en la primera se estima la tendencia desconocida, y en la segunda se realiza la predicción lineal óptima con media supuestamente conocida. Este tipo de aproximación se utiliza también para el cálculo de predicciones kriging con tendencias no lineales.

1.3.4. Consideraciones sobre los métodos kriging

Los predictores kriging presentados en esta sección cumplen las siguientes propiedades:

- Son BLUP (Best Linear unbiased predictor). Además, si el proceso es gaussiano, entonces el predictor kriging coincide con el mejor predictor posible
- Los predictores kriging son interpoladores exactos, es decir $\hat{Y}(\mathbf{x}_i) = Y(\mathbf{x}_i)$ para $i = 1, \ldots, n$ (suponiendo que no hay error de medida), y en ese caso la varianza de predicción es 0.
- Una práctica frecuente a la hora de obtener la predicción kriging $\hat{Y}(\mathbf{x}_0)$ es la elección de un vecindario de la posición espacial $\mathbf{x}_0$ en lugar de utilizar los n datos muestrales. Este vecindario puede seleccionarse considerando los datos más próximos (ver algunas directrices para ello en Webster y Oliver, 2007, Sección 8.5), o puede utilizarse un radio de búsqueda en torno a la posición considerada.
- Los métodos kriging admiten la predicción de agregaciones espaciales en lugar de valores puntuales (e incluso cuando los valores observados son también agregaciones). Por ejemplo puede ser de interés la estimación del valor medio de la respuesta en una determinada región, para lo que se puede emplear el denominado block kriging (ver p.e. Isaaks y Srivastava, 1989, Cap. 13).

Por otra parte, los métodos kriging asumen que la estructura de dependencia es conocida (función variograma o covariograma), pero en la práctica estas funciones deben ser estimadas a partir de los datos muestrales. En este sentido, es necesario tener en cuenta las siguientes consideraciones:

- Para garantizar la existencia de solución del sistema de ecuaciones kriging,

el estimador de la función variograma (o del covariograma) debe ser válido, en el sentido de que debe verificar la propiedad de ser condicionalmente semidefinido negativo (o definido positivo, respectivamente).

- Se ha demostrado que el predictor kriging es asintóticamente eficiente, siempre y cuando la estimación del variograma en saltos cercanos al origen sea adecuada (Stein, 1988).

- Sin embargo, el uso de variogramas estimados en las ecuaciones kriging inciden en la varianza de predicción kriging, la cual subestima al error final de la predicción. (ver p.e. Cressie, 1993, Sección 5.3.3). En general se tiene que:

$$\sigma_K^2(\mathbf{x}_0) \leq \mathbb{E}\left[\left(\hat{Y}(\mathbf{x}_0) - Y(\mathbf{x}_0)\right)^2\right]. \tag{1.11}$$

El método kriging puede ser modificado para responder a distintas situaciones, como por ejemplo, en los casos del kriging log-normal y trans-normal, en los cuales es necesario transformar el proceso original para garantizar la linealidad del predictor (ver p.e. Cressie, 1993, Sección 3.2.2 para más detalles).

1.4. Modelización de la dependencia espacial en procesos estacionarios

En el estudio de datos geoestadísticos, es necesario caracterizar la estructura de dependencia para poder realizar inferencia sobre el proceso espacial, y en especial para obtener las predicciones kriging. En el caso de procesos estacionarios, esto puede reducirse a la estimación de la función de covarianza o del variograma, si se supone que el proceso es estacionario de segundo orden o intrínseco, respectivamente. Sin embargo en la práctica, en lugar de estimar el covariograma se

prefiere utilizar el variograma, por ser este último más general, además de existir ciertas ventajas de orden teórico en su estimación (ver Sección 1.4.1).

En geoestadística, el mecanismo para la obtención de un modelo válido de variograma que describa adecuadamente la variabilidad espacial de los datos se denomina *análisis estructural*. El procedimiento habitual se basa en el método de mínimos cuadrados, que consta de los siguientes pasos:

1. Aproximación inicial del variograma, utilizando algún tipo de estimador piloto no paramétrico.

2. Selección y ajuste de un modelo (paramétrico) válido de variograma a estas estimaciones piloto.

3. Diagnosis del modelo de variograma ajustado, utilizando por ejemplo el método de validación cruzada.

De forma alternativa, se puede recurrir a métodos de máxima verosimilitud, o de máxima verosimilitud restringida para obtener un modelo válido de variograma. De manera general, estos métodos asumen que la distribución de los datos es normal y obtienen estimaciones de los parámetros de dicho modelo seleccionando aquellos valores que maximizan la función de verosimilitud. Una revisión general sobre estos procedimientos se puede encontrar en Cressie (1993, Sección 2.6). Si bien estos métodos presentan ciertas ventajas (p.e. no requiere de estimaciones piloto del variograma), en la práctica resulta difícil verificar la normalidad de los datos a partir de una realización parcial del proceso espacial. Además, estos métodos por lo general tienen un gran coste computacional, debido a los cálculos matriciales involucrados.

En el presente trabajo nos centraremos en el método basado en mínimos cuadrados dado que requiere menos suposiciones acerca de la distribución de los

datos. Además el enfoque de estimación no paramétrico que se verá en capítulos posteriores guarda algunas similitudes con este procedimiento.

1.4.1. Estimación piloto del variograma:

Supongamos que el proceso espacial $Y(\cdot)$ es intrínsicamente estacionario. Considerando que en este tipo de procesos la media es constante, la idea general para obtener estimadores piloto del variograma se basa en aproximar la relación:

$$2\gamma(\mathbf{u}) = \mathbb{E}\left[(Y(\mathbf{x}) - Y(\mathbf{x}+\mathbf{u}))^2\right]. \tag{1.12}$$

Semivariograma empírico

En términos simples, este estimador intenta promediar los cuadrados de las diferencias entre todos los pares de observaciones disponibles, que se encuentran separadas a un salto $\mathbf{u} \in \mathbb{R}^d$ determinado (o dentro una zona de tolerancia de dicho salto). El semivariograma empírico o clásico (Matheron, 1962), se expresa por:

$$\hat{\gamma}(\mathbf{u}) = \frac{1}{2\,|N(\mathbf{u})|} \sum_{i,j \in N(\mathbf{u})} (Y(\mathbf{x}_i) - Y(\mathbf{x}_j))^2, \tag{1.13}$$

siendo $|N(\mathbf{u})|$ la cardinalidad del conjunto $N(\mathbf{u}) = \Big\{(i,j) : \mathbf{x}_i - \mathbf{x}_j \in Tol(\mathbf{u})\Big\}$, donde $Tol(\mathbf{u}) \subset \mathbb{R}$ es una región de tolerancia prefijada, admitiendo un margen de desviación respecto al salto $\mathbf{u}$ en cuanto a su dirección y distancia.

De forma análoga es factible definir el estimador clásico del covariograma para el caso de procesos estacionarios de segundo orden:

$$\hat{C}(\mathbf{u}) = \frac{1}{2\,|N(\mathbf{u})|} \sum_{i,j \in N(\mathbf{u})} \left(Y(\mathbf{x}_i) - \bar{Y}\right)\left(Y(\mathbf{x}_j) - \bar{Y}\right), \tag{1.14}$$

donde $\bar{Y}$ es la media muestral de los datos observados $\mathbf{Y}$.

Comparando los estimadores (1.13) y (1.14), se puede verificar que el principal inconveniente del estimador del covariograma, es que este requiere estimar la media μ del proceso espacial. Por otro lado, se ha demostrado que el estimador empírico del variograma es insesgado bajo la hipótesis intrínseca, mientras que para procesos estacionarios de segundo orden, el estimador clásico $\hat{C}(\mathbf{u})$ tiene sesgo $O(1/n)$ (Para más detalles, ver Cressie, 1993, Section 2.4.1.).

Si los datos no siguen un comportamiento cercano a la distribución normal o si existe la presencia de datos atípicos, se puede recurrir al uso de estimadores más robustos, como los propuestos por Cressie y Hawkings (Cressie, 1993, Sección 2.4.3 p. 75), entre los que se encuentran:

$$2\hat{\gamma}(\mathbf{u}) = \frac{1}{B(\mathbf{u})} \left(mediana \left\{ |Y(\mathbf{x}_i) - Y(\mathbf{x}_j)|^{1/2} : (i,j) \in N(\mathbf{u}) \right\}\right)^4.$$

donde $B(\mathbf{u})$ es un término de corrección de sesgo y asintóticamente se tiene que $B(\mathbf{u}) = 0{,}457$.

Otros estimadores no paramétricos

La relación (1.12) se puede plantear como una media ponderada, de la forma siguiente:

$$\hat{\gamma}(\mathbf{u}) = \frac{1}{2\sum\limits_{i,j} w_{ij}(\mathbf{u})} \sum_{i,j} w_{ij}(\mathbf{u}) \left(Y(\mathbf{x}_i) - Y(\mathbf{x}_j)\right)^2, \tag{1.15}$$

donde $w_{ij}(\mathbf{u}) \geq 0, \forall i,j$. Dependiendo de la asignación de los pesos $w_{ij}(\mathbf{u})$, se obtienen diferentes estimadores no paramétricos del variograma:

- **Estimador Nadaraya-Watson:** En este caso, los pesos vienen dados por:

$$w_{ij}(\mathbf{u}) = K_{\mathbf{G}}(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}),$$

donde $K_{\mathbf{G}}(\mathbf{u}) = |\mathbf{G}|^{-1} K(\mathbf{G}^{-1}\mathbf{u})$, siendo K es una función tipo núcleo d-dimensional, $\mathbf{G}$ es la matriz ventana $d \times d$ simétrica no singular y $|\mathbf{G}|$ es el determinante de $\mathbf{G}$.

- **Estimador Lineal Local:** Tomando como base los trabajos sobre regresión polinómica local de Wand y Jones (1995, Cap. 2) y Fan y Gijbels (1996, Cap. 3), García-Soidán *et al.* (2003) proponen el estimador lineal local del variograma para un salto $\mathbf{u}$ determinado, como la solución para α del siguiente problema de mínimización:

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i,j} \left\{ \frac{1}{2} \left(Y(\mathbf{x}_i) - Y(\mathbf{x}_j)\right)^2 - \alpha - \boldsymbol{\beta}^t(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}) \right\}^2 K_{\mathbf{G}}(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}), \tag{1.16}$$

Bajo la suposición de isotropía, los pesos vienen dados por:

$$w_{ij}(u) = K\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\| - u}{g}\right) \times$$

$$\sum_{k<l} K\left(\frac{\|\mathbf{x}_k - \mathbf{x}_l\| - u}{g}\right) \left(\|\mathbf{x}_k - \mathbf{x}_l\| - u\right) \left(\|\mathbf{x}_k - \mathbf{x}_l\| - \|\mathbf{x}_i - \mathbf{x}_j\|\right).$$

Los estimadores mencionados anteriormente dependen de una adecuada selección de la matriz de suavizado $\mathbf{G}$ (o ventana g en el caso isotrópico). Estas ventanas se suelen obtener mediante técnicas de validación cruzada, minimizando el error cuadrático o cuadrático relativo de la estimaciones del variograma respectivo.

Varios estudios han demostrado buenas propiedades teóricas para el estimador lineal local, tales como su insesgadez asintótica y su consistencia, (ver p.e. García-Soidán *et al.*, 2003). Una ventaja importante es la reducción de los efectos frontera, lo cual es de especial importancia en la estimación del variograma cerca del origen. Cabe indicar, que el procedimiento para obtener el estimador

lineal local del variograma se encuentra implementado, por ejemplo en la función `np.svar` del paquete `npsp` del software R (Fernández-Casal, 2014).

1.4.2. Selección y ajuste de modelos válidos de variogramas

Los estimadores del variograma presentados en el apartado anterior no cumplen necesariamente la condición de ser condicionalmente semidefinidos negativos y por tanto no pueden aplicarse directamente en las ecuaciones kriging. Por tal razón, normalmente se selecciona y ajusta un módelo válido de variograma a dichas estimaciones piloto.

Modelos paramétricos isotrópicos

A continuación se presentan algunos modelos paramétricos de variograma conocidos, bajo el supuesto de isotropía, pues estos constituyen la base para la construcción de modelos más complejos. En la notación utilizada en todos estos modelos, $\boldsymbol{\theta} = (c_0, c_1, a)$ es el vector de parámetros, cuyos elementos corresponden a las características del variograma presentadas en la Sección 1.2.2, donde $c_0 \geq 0$ y $c_1 \geq 0$ representan el efecto nugget y el umbral parcial respectivamente, tales que la varianza $\sigma^2 = c_0 + c_1$, mientras que a representa el rango práctico.

- **Modelo exponencial:** Representa una dependencia espacial que crece exponencialmente con la distancia.

$$\gamma(\mathbf{u}|\boldsymbol{\theta}) = \begin{cases} 0 & \text{si} \quad \mathbf{u} = \mathbf{0} \\ c_0 + c_1 \left(1 - \exp\left(-3\frac{\|\mathbf{u}\|}{a}\right)\right) & \text{si} \quad \mathbf{u} \neq \mathbf{0} \end{cases} \tag{1.17}$$

- **Modelo gaussiano:**

$$\gamma(\mathbf{u}|\boldsymbol{\theta}) = \begin{cases} 0 & \text{si} \quad \mathbf{u} = \mathbf{0} \\ c_0 + c_1 \left(1 - \exp\left(-3\frac{\|\mathbf{u}\|^2}{a^2}\right)\right) & \text{si} \quad \mathbf{u} \neq \mathbf{0} \end{cases} \tag{1.18}$$

- **Modelo de Matérn:**

$$\gamma(\mathbf{u}|\boldsymbol{\theta}) = \begin{cases} 0 & \text{si} \quad \mathbf{u} = \mathbf{0} \\ c_0 + c_1 \left(1 - \frac{1}{2^{v-1}\Gamma(v)} \left(\frac{||u||}{a}\right)^v K_v \left(\frac{||u||}{a}\right)\right) & \text{si} \quad \mathbf{u} \neq \mathbf{0} \end{cases} \tag{1.19}$$

donde K_v denota la función de Bessel modificada de tercera clase de orden v (ver p.e. Abramowitz y Stegun, 1964, pp. 374-379) y v es un paramétro de suavizado. Un caso especial se presenta cuando $v = 1/2$, donde se obtiene el modelo de variograma exponencial, y para el caso que $v \longrightarrow \infty$ corresponde a un modelo gaussiano.

Métodos de ajuste de modelos válidos

Una vez seleccionado el modelo paramétrico, el siguiente paso es ajustar el mismo a las estimaciones piloto del variograma (usualmente obtenidas mediante el estimador empírico (1.13)) mediante el método de mínimos cuadrados. Supongamos que se desea estimar el variograma teórico $2\gamma(\mathbf{u}, \boldsymbol{\theta_0})$, para lo cual se tiene el vector de estimaciones piloto $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}(\mathbf{u}_1), \ldots, \hat{\gamma}(\mathbf{u}_q))^t$. Para establecer los saltos $\mathbf{u}_i$ se pueden considerar distancias que sean menores o iguales a la mitad del salto máximo y tales que existan al menos 30 pares de datos distintos (Journel y Huijbregts, 1978, p. 194).

La estimación por mínimos cuadrados de los parámetros $\boldsymbol{\theta_0}$ (ver p.e. Cressie,

1993, p. 96-97) se obtiene al minimizar:

$$(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta}))^t \, \mathbf{V}(\boldsymbol{\theta}) \, (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}(\boldsymbol{\theta})) \,, \tag{1.20}$$

siendo $\boldsymbol{\gamma}(\boldsymbol{\theta}) = (\hat{\gamma}(\mathbf{u}_1, \boldsymbol{\theta}), \ldots, \hat{\gamma}(\mathbf{u}_q), \boldsymbol{\theta})^t$ y $\mathbf{V}(\boldsymbol{\theta})$ una matriz de orden $q \times q$ semidefinida positiva que puede depender de $(\boldsymbol{\theta})$ de alguna de las siguientes maneras:

- **Mínimos cuadrados ordinarios (m.c.o.)**: $\mathbf{V}(\boldsymbol{\theta}) = \mathbf{I}_q$ donde $\mathbf{I}_q$ es la matriz identidad de orden q.
- **Mínimos cuadrados ponderados (m.c.p.)**: Se selecciona $\mathbf{V}(\boldsymbol{\theta}) = diag(w_1(\boldsymbol{\theta}), \ldots, w_q(\boldsymbol{\theta}))$ con pesos $w_i(\boldsymbol{\theta}) \geq 0, i = 1, \ldots, q$. Usualmente, estos pesos se toman inversamente proporcionales a $Var(\hat{\gamma}(\mathbf{u}_i))$, por ejemplo haciendo $w_i(\boldsymbol{\theta}) = |N(\mathbf{u}_i)|/\hat{\gamma}(\mathbf{u}_i, \boldsymbol{\theta})$ (ver p.e. Cressie, 1985).
- **Mínimos cuadrados generalizados (m.c.g.)**: $\mathbf{V}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\theta})^{-1}$ donde $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\gamma}}}(\boldsymbol{\theta})$ es la matriz de varianzas y covarianzas de $\hat{\boldsymbol{\gamma}}$ obtenida suponiendo que el variograma teórico es $2\gamma(\mathbf{u}, \boldsymbol{\theta_0})$.

Sin embargo, los últimos métodos conllevan un problema circular, pues la matriz de pesos $\mathbf{V}(\boldsymbol{\theta})$ depende de los parámetros que se desean estimar. Por tal motivo, se suele recurrir a un proceso iterativo, donde en cada etapa k el vector $\boldsymbol{\theta}^k$ se estima por m.c.p o m.c.g. utilizando el vector de pesos $\mathbf{V}(\boldsymbol{\theta}^{k-1})$, siendo $\boldsymbol{\theta}^0$ la estimación obtenida por m.c.o.

Ejemplo de aplicación: Datos Meuse

Retomando el ejemplo de la Sección 1.1, consideraremos que los datos de concentración de zinc (en log(ppm)) corresponden a un proceso estacionario de segundo orden. Luego, asumiendo isotropía, se estimó el variograma respectivo

utilizando las rutinas disponibles en el paquete *gstat* del software *R*, con sus configuraciones por defecto para realizar las distintas operaciones, como por ejemplo, para determinar los saltos u y el método de ajuste de modelos.

En el primer paso, se obtuvieron las estimaciones piloto $\hat{\gamma}(u)$ de dicho variograma para cada salto u, utilizando el estimador empírico (1.13). De acuerdo con el comportamiento de este variograma piloto, se consideró que el modelo $\gamma(\mathbf{u}|\boldsymbol{\theta})$ de variograma exponencial (1.17) es el que mejor se aproxima a estas estimaciones. Por tanto, ajustando este modelo se obtuvieron las siguientes estimaciones de los parámetros: $c_0 = 0$, $a = 450$ y $c_1 = 0{,}719$. En la figura 1.6 se muestran las estimaciones obtenidas para las estimaciones piloto (representado por puntos) y para el variograma exponencial ajustado (línea continua).

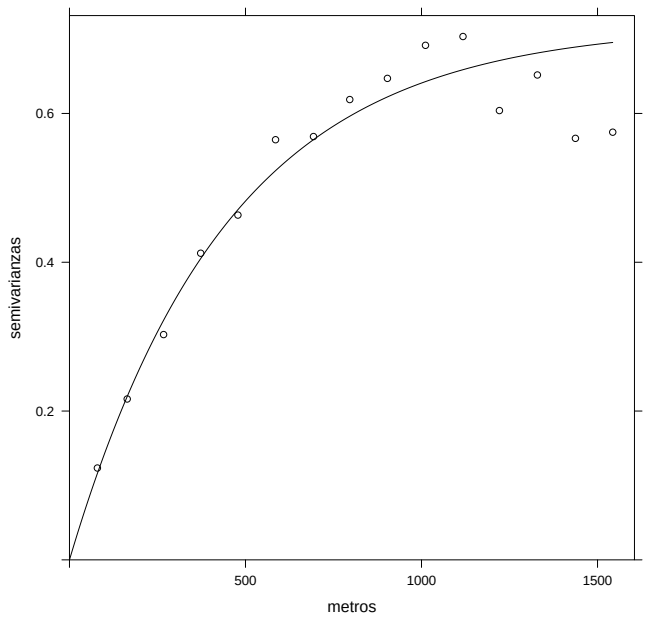
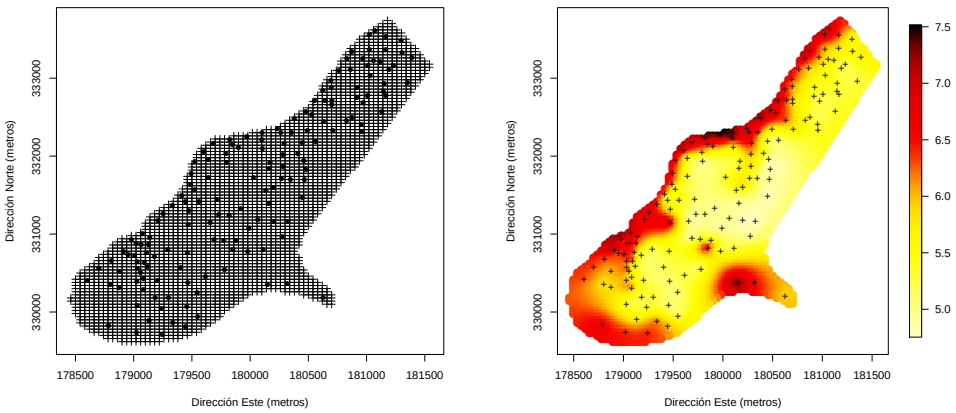


Figura 1.6: Variograma exponencial ajustado (linea continua) al variograma empírico piloto (puntos) para los datos de concentración de zinc (log(ppm)).

De acuerdo con la figura anterior y a los valores obtenidos, podemos observar que el modelo considerado presenta un buen ajuste en especial a saltos cercanos al origen. De acuerdo a las estimaciones, se puede decir que el proceso estacionario

tiene una varianza de $\sigma^2 = 0{,}719$, un efecto nugget nulo (por tanto no hay discontinuidad en el origen del variograma) y que pares de observaciones con distancias mayores a 450 metros se pueden considerar incorreladas entre sí.

A partir de este modelo ajustado, es factible construir predicciones kriging, en este caso mediante KO. Para esto, se ha dividido la zona de observación en una rejilla regular de predicción donde cada celda tiene un área de 40×40 metros, obteniéndose 3103 localizaciones en total (este conjunto de datos se encuentra disponible en el paquete *sp* de *R*.) Esta rejilla de predicción, junto con las predicciones kriging obtenidas se pueden observar en las Figura 1.7(a) y 1.7(b) respectivamente.



(a) (b)

Figura 1.7: (a) Rejilla regular de predicción y (b) predicciones kriging obtenidas por KO y variograma exponencial ajustado con $c_0 = 0$, $a = 450$ y $c_1 = 0{,}719$.

1.4.3. Diagnosis del modelo ajustado

El tercer paso del análisis estructural implica utilizar algún mecanismo para tratar de diagnosticar si el modelo válido ajustado reproduce de manera adecuada la dependencia espacial del proceso $Y(\cdot)$. Habitualmente, la técnica utilizada con estos fines es la *Validación Cruzada* (VC), la cual recurre a las predicciones

espaciales (p.e. mediante kriging) y a las estimaciones del error cuadrático medio de predicción.

De manera general, el método VC consiste en utilizar el modelo de variograma ajustado para obtener el predictor $\hat{Y}_{-j}(\mathbf{x}_j)$ de $Y(\mathbf{x}_j)$, a partir del vector de observaciones $\mathbf{Y}_{-j} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_{j-1}), Y(\mathbf{x}_{j+1}), \ldots, Y(\mathbf{x}_n))$, así como su correspondiente error en media cuadrática de predicción $\sigma^2_{-j}(\mathbf{x}_j)$. Para medir la aproximación de los predictores a los valores observados, se puede proceder de distintas formas:

- Analizar si la media $\frac{1}{n}\sum_{j=1}^{n}(\hat{Y}_{-j}(\mathbf{x}_j) - Y(\mathbf{x}_j))/\sigma_{-j}(\mathbf{x}_j)$ es próxima a 0.
- Analizar si la desviación $\left[\frac{1}{n}\sum_{j=1}^{n}\left((\hat{Y}_{-j}(\mathbf{x}_j) - Y(\mathbf{x}_j))/\sigma_{-j}(\mathbf{x}_j)\right)^2\right]^{1/2}$ es próxima a 1.
- Utilizar criterios gráficos como histogramas o diagramas de tallos y hojas de los valores $(\hat{Y}_{-j}(\mathbf{x}_j) - Y(\mathbf{x}_j))/\sigma_{-j}(\mathbf{x}_j)$ y analizar la presencia de outliers.

1.4.4. Modelos flexibles de variogramas

En la práctica, la selección de un determinado modelo paramétrico se suele realizar mediante procedimientos automáticos, o de acuerdo al criterio del usuario. Esto puede generar problemas de mala especificación del modelo a ajustar. Además, un modelo paramétrico puede no ser lo suficientemente flexible para representar la información obtenida mediante las estimaciones empíricas. Por tales motivos, se han propuesto distintos tipos de modelos flexibles de covariograma (o semivariograma), como por ejemplo, aquellos basados en la representación en medias móviles (Barry *et al.*, 1996), en suavizado tipo spline (Lele, 1995), a través de series de Fourier (García-Soidán *et al.*, 2012) o en la representación espectral del covariograma (Shapiro y Botha, 1991).

En el presente trabajo recurriremos a los modelos de Shapiro-Botha (S-B) para la obtención de modelos válidos, debido a sus buenas propiedades estadísticas, en especial si se utilizan como estimadores pilotos los estimadores no paramétricos de la forma (1.15) (ver p.e. Fernández-Casal *et al.*, 2003b; García-Soidán *et al.*, 2004). Asimismo, Cherry (1996) comprobó el mejor comportamiento de esta técnica en comparación con el enfoque paramétrico tradicional, incluso cuando los datos se habían generado bajo modelos paramétricos. Por otra parte, Carmack *et al.* (2012b) presentaron una versión aún más flexible de este modelo de variograma cerca del origen. Además, los modelos S-B han demostrado ser de fácil implementación en la práctica y han sido extendidos al contexto espacio temporal, incluso considerando anisotropía (Fernández-Casal *et al.*, 2003a,b).

Construcción de modelos Shapiro - Botha

Supongamos que el proceso espacial es estacionario de segundo orden. A partir del teorema de Bochner (Bochner, 1955), la presentación del variograma isotrópico (no necesariamente continuo en el origen) es de la forma:

$$\gamma(u) = \begin{cases} 0 & \text{si} \quad u = 0 \\ v_0 - v(u) & \text{si} \quad u \neq 0 \end{cases} \tag{1.21}$$

donde v_0 es una constante positiva y $v(u)$ es una función semidefinida positiva continua en el origen (es decir, un covariograma), que admite la siguiente representación espectral:

$$v(u) = \int_0^\infty \kappa_d(\omega u) dG(\omega),$$

donde

$$\kappa_d(x) = \left(\frac{2}{x}\right)^{\frac{d-2}{2}} \Gamma\left(\frac{d}{2}\right) J_{\frac{d-2}{2}}(x),$$

siendo J_l una función de Bessel de orden l (ver Abramowitz y Stegun, 1964, pp. 358-359) y $G(\cdot)$ una función positiva, acotada y no decreciente en $[0, \infty)$. Entonces, el problema se reduce a encontrar la constante positiva v_0 y la función G con estas características, de manera que el variograma dado por (1.21) describa la estructura de dependencia del conjunto de datos. Sin embargo, la resolución numérica de este problema resulta ser demasiado complicada, por lo que Shapiro y Botha (1991) propusieron una simplificación del mismo, considerando una discretización de la función G, suponiendo que dG es una medida atómica con saltos finitos positivos z_j en puntos $x_j, j = 1, \ldots, J$:

$$G(x) = \sum_{x_j \leq x} z_j,$$

y tomando posiciones equiespaciadas $x_j = j\phi$, siendo ϕ un número positivo fijado de antemano. Luego, los modelos de variogramas obtenidos por este método son de la forma (1.21), donde:

$$v(u) = \sum_{j=1}^{J} \kappa_d(x_j u) z_j,$$

sujeto a la restricción lineal:

$$v_0 - \sum_{j=1}^{J} z_j \geq 0.$$

Esta última restricción se debe al hecho de que toda función semidefinida positiva cumple la desigualdad de Cauchy–Schwarz, es decir, que el valor absoluto de dicha función está acotada por su valor en el origen. Esto implica además que en (1.21) se verifica:

$$v_0 - v(0) = v_0 - \int_0^{\infty} dG(\omega) \geq 0.$$

Además, bajo la presencia de efecto nugget se tiene que $c_0 = v_0 - v(0) > 0$.

Supongamos que el vector $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}(u_1), \ldots, \hat{\gamma}(u_q))^t$ representa las estimaciones piloto obtenidas con el variograma isotrópico $\gamma(u)$ en los saltos $u_i, i = 1 \ldots, q$. Luego, utilizando m.c.p., el problema de ajuste se reduce a encontrar el vector $\boldsymbol{\theta} = (z_1, \ldots, z_j, v_0)^t$ que minimiza la función:

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{q} w_i \left(\hat{\gamma}(u_i) - v_0 + \sum_{j=1}^{J} \kappa_d(x_j u_i) z_j \right)^2, \tag{1.22}$$

sujeto a las restricciones lineales:

$$v_0 - \sum_{j=1}^{J} z_j \geq 0; \quad z_j \geq 0, \quad j = 1, \ldots, J.$$

La función objetivo (1.22) se puede reescribir como una función cuadrática:

$$Q(\boldsymbol{\theta}) = (\hat{\boldsymbol{\gamma}} - \mathbf{A}\boldsymbol{\theta})^t \, \mathbf{W} \, (\hat{\boldsymbol{\gamma}} - \mathbf{A}\boldsymbol{\theta}),$$

donde $\mathbf{W} = diag(w_1, \ldots, w_q)$ es la matriz de pesos de orden $q \times q$, y $\mathbf{A}$ es una matriz de coeficientes de orden $q \times (J+1)$ tal que $a_{ij} = -\kappa_d(x_j u_i)$ para $i = 1, \ldots, q$, $j = 1, \ldots, J$, y $a_{i,J+1} = 1$ para $i = 1, \ldots, q$. Luego, es factible determinar el vector $\boldsymbol{\theta}$ utilizando técnicas de programación cuadrática.

Los pesos w_i se pueden determinar de forma iterativa, por ejemplo, definiendo en primer lugar $w_i = 1$, y luego se recalculan los pesos en cada iteración hasta la convergencia. Cabe mencionar, que este procedimiento se puede resolver utilizando m.c.g., pues solo se cambiaría la matriz de pesos.

Si $\hat{\boldsymbol{\theta}} = (\hat{z}_1, \ldots, \hat{z}_j, \hat{v}_0)^t$ es la solución de (1.22), el modelo de variograma ajus-

tado S-B tiene la forma:

$$\hat{\gamma}(u) = \hat{v}_0 - \sum_{j=1}^{J} \kappa_d(x_j u)\hat{z}_j, \tag{1.23}$$

donde el efecto nugget se puede estimar por $c_0 = \hat{v}_0 - \sum_{j=1}^{J} \hat{z}_j$.

Es importante mencionar que el procedimiento de ajuste de este tipo de modelos se encuentra implementado en la función `fitsvar.sb.iso` del paquete `npsp` del software R (Fernández-Casal, 2014).

1.5. Estimación en procesos no estacionarios

En geoestadística, la hipótesis de estacionaridad simplifica en gran medida la estimación de las características del proceso, facilitando realizar inferencia sobre el proceso en una localización espacial determinada a partir de las observaciones cercanas a él (Cressie, 1993). Sin embargo, no siempre es factible hacer esta suposición, ya sea debido a la presencia de una tendencia no constante o porque la estructura de dependencia no es estacionaria. En el presente estudio nos centraremos en el primer caso (procesos no estacionarios en media), mientras que en el Capítulo 4 se analizará un modelo más general bajo dependencia no estacionaria (procesos heterocedásticos).

Un proceso $Y(\mathbf{x})$ *no es estacionario en media* cuando su esperanza matemática depende de la posición espacial $\mathbf{x}$, es decir:

$$\mathbb{E}\left[Y(\mathbf{x})\right] = \mu(\mathbf{x}), \forall \mathbf{x} \in D.$$

donde $\mu(\mathbf{x})$ representa la función *tendencia* del proceso en la localización $\mathbf{x}$.

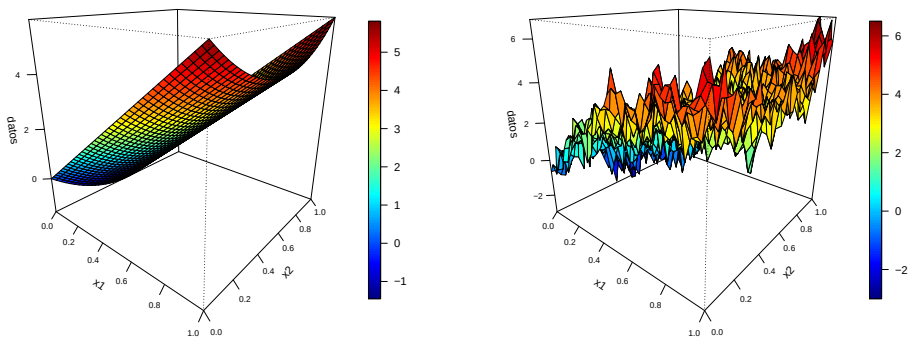
Supongamos que el proceso no estacionario $Y(\mathbf{x})$ se puede representar como

la suma de una componente determinística y un proceso de error aleatorio:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1.24}$$

siendo $\mu(\cdot)$ una función tendencia determinística relacionada con la variabilidad a gran escala, y $\varepsilon(\cdot)$ es un proceso espacial de media nula, en el cual se recoge la variabilidad de pequeña escala. Se podría imponer además, que el proceso de error es estacionario de segundo orden, con covariograma dado por $C(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$, para $\mathbf{u} > 0$. En el caso más general, si se trata de un proceso intrínsicamente estacionario, el variograma del proceso está definido por:

$$2\gamma(\mathbf{u}) = Var(Y(\mathbf{x}) - Y(\mathbf{x}+\mathbf{u})) = Var(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u})). \tag{1.25}$$



(a) (b)

Figura 1.8: (a) Tendencia determinística bidimensional $\mu(x_1, x_2) = 5{,}8(x_1 - x_2 + x_2^2)$, y (b) Datos espaciales simulados sobre $D = [0,1]^2$, bajo el modelo $Y(\mathbf{x}) = \mu(\mathbf{x}) + \varepsilon(\mathbf{x})$, con semivariograma isotrópico exponencial $\gamma(u)$ con $\sigma^2 = 1$, $a = 0{,}6$ y $c_0 = 0{,}2$.

A modo de ejemplo, en la Figura 1.8(b) se presentan $n = 1600$ observaciones simuladas de un proceso espacial gaussiano no estacionario bajo el modelo (1.24), definido sobre la región $D = [0,1]^2 \subset \mathbb{R}^2$, donde $\mu(x_1, x_2) = 5{,}8(x_1 - x_2 + x_2^2)$

corresponde a la tendencia teórica presentada en la Figura 1.8(a), y tomando como $\gamma(\mathbf{u})$ al variograma representado en la Figura 1.4(a). Aquí se puede observar el efecto de la tendencia sobre el comportamiento de los datos espaciales, en comparación con los datos simulados sin tendencia presentados en la Figura 1.5(a).

Supongamos que se dispone del vector $\mathbf{Y}$ correspondiente a n observaciones de una realización parcial del proceso espacial $Y(\cdot)$, el cual puede ser modelado de la forma (1.24). Entonces, para poder realizar inferencia del proceso espacial en una posición espacial $\mathbf{x}_0$, es necesario obtener estimaciones tanto de la función tendencia $\mu(\mathbf{x})$ como del variograma $\gamma(\mathbf{u})$.

Si bien la tendencia puede tener cualquiera forma, es usual suponer que esta puede representarse bajo un modelo lineal:

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{1.26}$$

donde $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))$, $\boldsymbol{\beta}$ es un vector de paramétros y la matriz de diseño $\mathbf{X}$ puede corresponderse con funciones de las posiciones espaciales, como se propone en el kriging universal (ver Sección 1.3.2), o incluir variables auxiliares (ver p.e. Wackernagel, 2003, Cap. 37, acerca de los métodos kriging con "external drift"). Es habitual asumir una estructura polinómica para $\mu(\mathbf{x})$, en la cual las funciones $\{f_j(\cdot) : j = 0, \ldots, p\}$ en la ecuación (1.8) corresponden a los monomios $x_1^{a_1} \ldots x_d^{a_d}$, donde x_i representa la i-ésima componente del vector $\mathbf{x} \in \mathbb{R}^d$, y $a_1, \ldots a_d$ son números naturales cuya suma es menor o igual que k, siendo $k \in \mathbb{N}$. En ese caso, la tendencia $\mu(\cdot)$ corresponde a una superficie de tendencia polinómica de grado k. Para el caso bidimensional, la tendencia tendría la siguiente forma:

$$\mu(\mathbf{x}) = \sum_{0 \leq a_1 + a_2 \leq k} \alpha_{a_1 a_2} x_1^{a_1} x_2^{a_2}; \;\; \mathbf{x} = (x_1, x_2) \in \mathbb{R}^2,$$

donde $p = \{(k+1)(k+2)/2\} - 1$, y:

$$f_0(\mathbf{x}) = 1, f_1(\mathbf{x}) = x_1, f_2(\mathbf{x}) = x_2, \ldots, f_p(\mathbf{x}) = x_2^k.$$

La tendencia representada en la Figura 1.8(a) constituye un ejemplo de una tendencia polinómica de grado $k = 2$.

Por otra parte, de la definición (1.25) resulta obvio que si la tendencia es determinística, el variograma del proceso coincide con el variograma de los errores. Esto podría inducir a utilizar directamente los estimadores presentados en la Sección 1.4.1 para aproximar $\gamma(\cdot)$, sin embargo si se admite la presencia de una tendencia se tiene que:

$$\mathbb{E}\left[(Y(\mathbf{x}) - Y(\mathbf{x}+\mathbf{u}))^2\right] = 2\gamma(\mathbf{x}, \mathbf{x}+\mathbf{u}) + (\mu(\mathbf{x}) - \mu(\mathbf{x}+\mathbf{u}))^2. \tag{1.27}$$

Como los estimadores en el caso estacionario se basan en aproximar la relación (1.12), estos no son adecuados para estimar $\gamma(\cdot)$ cuando la tendencia no es constante.

1.5.1. Estimación paramétrica basada en residuos

El enfoque tradicional de estimación en procesos de tendencia no constante implica utilizar los residuos $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$, donde $\hat{\boldsymbol{\mu}}$ representa la tendencia estimada. La idea general del método basado en residuos es aproximar el variograma del proceso mediante la variabilidad de los residuos.

El método anterior necesita en primer lugar de un estimador de la tendencia

para obtener los residuos, pero la estimación óptima de $\mu(\cdot)$ requiere a su vez tomar en cuenta la estructura de la dependencia espacial, generándose un problema circular. Por ejemplo, suponiendo que la tendencia tiene la forma dada en (1.26), entonces el estimador lineal óptimo de $\boldsymbol{\beta}$ es:

$$\hat{\boldsymbol{\beta}}_{mcg} = (\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

donde $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas de los datos, la cual en la práctica es desconocida. Para solventar este inconveniente, Neuman y Jacobson (1984) propusieron un método iterativo, el cual consta de los siguientes pasos:

1. En primer lugar, se obtiene una estimación inicial de la tendencia $\mu(\cdot)$ del proceso, mediante mínimos cuadrados ordinarios:

$$\hat{\boldsymbol{\beta}}_{mco} = \min_{\boldsymbol{\beta}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

2. Posteriormente, se calculan los residuos o errores estimados: $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$, donde $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{mco}$ es el vector de estimaciones de la tendencia.

3. Se obtiene un variograma paramétrico válido para $\gamma(\mathbf{u})$ construido a partir de $\hat{\boldsymbol{\varepsilon}}$, siguiendo el procedimiento descrito en la Sección 1.4.

4. Se reestima la función $\mu(\cdot)$, teniendo en cuenta la estructura de dependencia de segundo orden. Para esto se utiliza el método de mínimos cuadrados generalizados estimados (m.c.g.e.), donde:

$$\hat{\boldsymbol{\beta}}_{mcge} = (\mathbf{X}^t\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}$$

donde la matriz de varianzas y covarianzas de los datos $\hat{\boldsymbol{\Sigma}}$ se obtiene construyendo el covariograma $\hat{C}(\mathbf{u})$ a partir del variograma estimado en el paso

anterior.

5. Se recalculan los residuos a partir del estimador $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{mcge}$

6. Se repiten los pasos 3 al 5 hasta obtener convergencia.

Ejemplo de aplicación: Datos Meuse

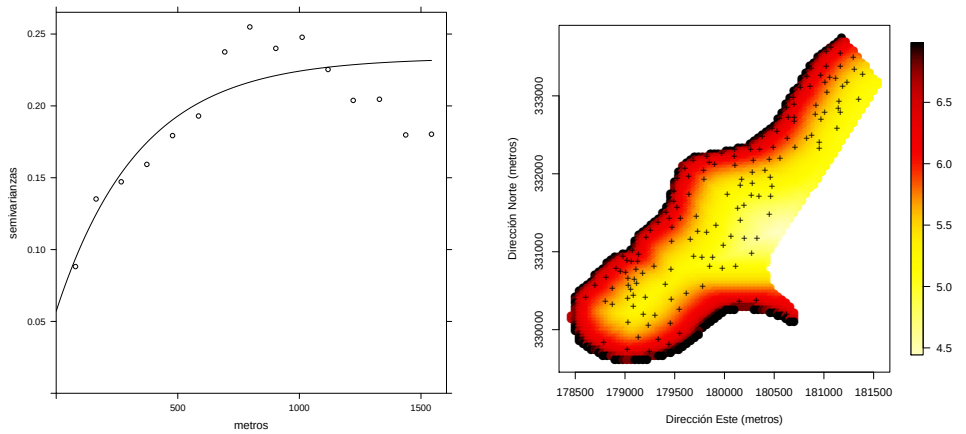
Consideremos ahora que los datos observados de las concentraciones de zinc medidas en las riberas del río Meuse corresponden a una realización parcial de un proceso espacial que admite el modelo no estacionario en media (1.24). Para obtener las estimaciones de los componentes respectivos mediante el método basado en residuos, recurrimos nuevamente a las funciones disponibles en el paquete *gstat*, con sus opciones por defecto. Este paquete permite obtener un variograma a partir de los residuos, especificando previamente un modelo paramétrico para la tendencia. Para este caso, asumimos que la tendencia se puede expresar como el siguiente modelo lineal:

$$\mu = \beta_0 + \beta_1\sqrt{dist}$$

donde la variable *dist* representa la distancia a la orilla del río.

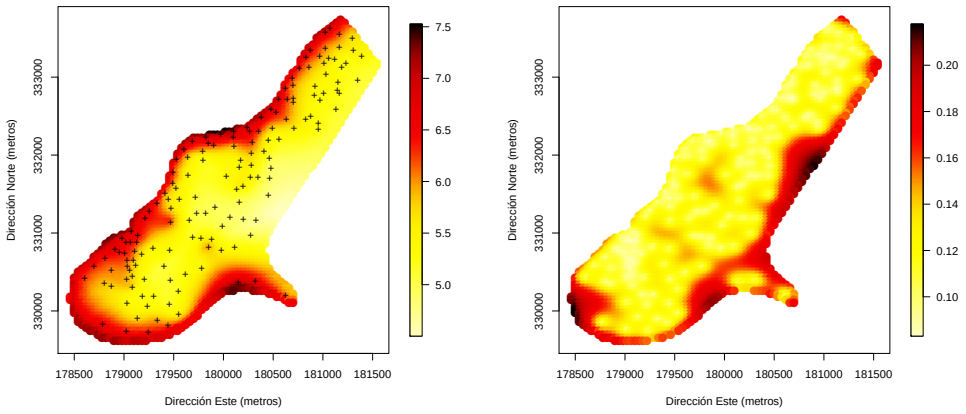
Una vez estimada la tendencia (inicialmente por m.c.o.) y calculados los residuos, se construye el variograma respectivo, utilizando el ajuste paramétrico considerando nuevamente el modelo de variograma exponencial, y obteniéndose finalmente los siguientes valores $c_0 = 0{,}0571$, $c_1 = 0{,}1764$ y $a = 340$. Estas estimaciones del variograma se presentan en la Figura 1.9(a). A partir del variograma ajustado, se obtiene la estimación de la tendencia por mínimos cuadrados generalizados estimados que se observa en la Figura 1.9(b).

Es importante observar que a diferencia del variograma estimado obtenido suponiendo estacionariedad (ver Figura 1.6), en este último variograma existe



(a) (b)

Figura 1.9: (a) Variograma exponencial con $c_0 = 0{,}0571$, $c_1 = 0{,}1764$ y $a = 340$ (línea contínua) ajustado al variograma empírico (puntos) estimado a partir de residuos, y (b) Tendencia estimada por m.c.g.e., para los datos de concentración de zinc (log(ppm)).



(a) (b)

Figura 1.10: (a) Predicciones y (b) Varianzas kriging de la concentración de zinc (log(ppm)) a partir de la tendencia estimada por m.c.g.e. y variograma exponencial con $c_0 = 0{,}0571$, $c_1 = 0{,}1764$ y $a = 340$ estimado a partir de los residuos.

discontinuidad en el origen. Además, como cabría esperar, la varianza del proceso de error es $\sigma^2 = c_0 + c_1 = 0{,}2335$ es mucho menor que la obtenida en el caso estacionario ($\sigma^2 = 0{,}719$), pues parte de la variabilidad del proceso original está recogida en la variabilidad a gran escala representada por la tendencia esti-

mada. Asimismo, se observa que el rango del variograma es de 340 metros, por lo que ahora los errores correspondientes a pares de observaciones separadas a una distancia superior a dicho rango se pueden considerar incorrelados (no así las observaciones originales, pues entre ellas persiste la variabilidad a gran escala).

A partir de estas estimaciones de $\mu(\cdot)$ y $\gamma(\cdot)$ es factible construir las predicciones kriging, tomando la misma rejilla de predicción utilizada en el ejemplo anterior. En este caso, las predicciones se obtienen realizando kriging simple sobre los residuos y posteriormente añadiendo la tendencia estimada. En las Figuras 1.10(a) y 1.10(b) se presentan las predicciones y varianzas kriging obtenidas mediante este procedimiento. Si se compara estas predicciones con las obtenidas mediante kriging ordinario suponiendo estacionariedad (ver Figura 1.7(b)), no se observan mayores diferencias excepto en los valores cercanos a la orilla del río, lo cual también parece evidenciarse con la variabilidad kriging, la cual es más alta en esas zonas.

1.5.2. Limitaciones de la estimación basada en residuos y propuestas alternativas

Aunque el método de estimación basada en residuos permite obtener estimaciones paramétricas conjuntas de la tendencia y la varianza de un proceso con tendencia no constante, es conocido que el uso de los residuos genera sesgos en la estimación del variograma (ver p.e. Cressie, 1993; Chilès y Delfiner, 2012; Wackernagel, 2003). Incluso estimando la tendencia de la manera más eficiente (utilizando m.c.g. y suponiendo conocida la matriz de varianzas y covarianzas teórica de los datos $\mathbf{\Sigma}$) se verifica que:

$$Var(\hat{\boldsymbol{\varepsilon}}) = Var(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{mcg}) = \mathbf{\Sigma} - \mathbf{X}(\mathbf{X}^t\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}^t$$

Debido a la presencia de este sesgo, los estimadores del variograma obtenidos mediante el procedimiento anterior generalmente subestiman a su correspondiente valor teórico. Por lo general, el sesgo del variograma es menor en saltos cercanos al origen, y va aumentando conforme aumentan los saltos (Matheron, 1971, pp. 152-155), tal como se puede observar en los resultados simulados representados en la Figura 1.11. Desde el punto de vista de la predicción espacial, si el estimador del variograma se obtiene mediante ajuste por m.c.p o m.c.g., el efecto del sesgo puede tener menor influencia en la predicción kriging, pues estos métodos otorgan mayor peso a los saltos pequeños, sin embargo, la varianza kriging si se verá afectada, subestimando a la varianza de predicción teórica (ver Cressie, 1993, pp. 166-168).

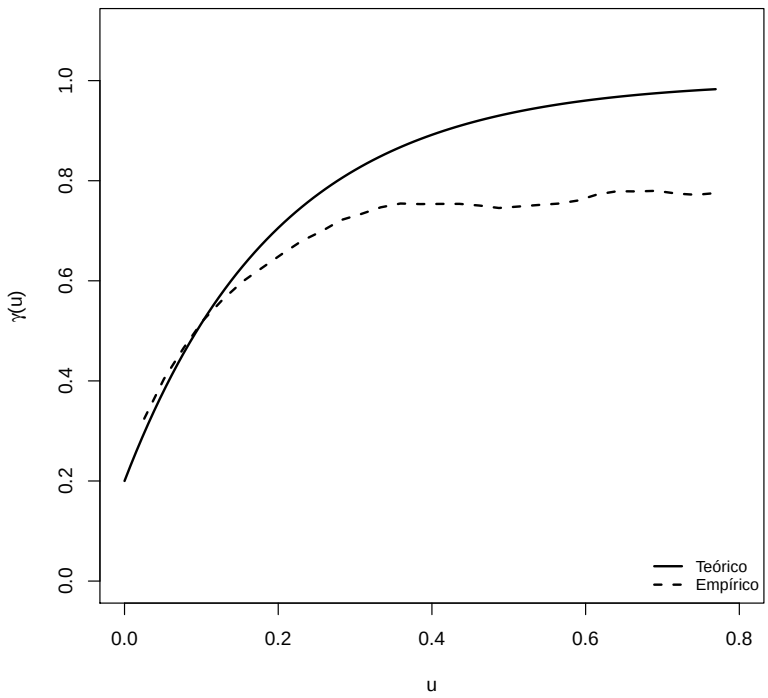


Figura 1.11: Variograma teórico (línea continua) y Variograma Empírico (línea discontinua) de los residuos obtenidos con la tendencia $\mu(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$, estimada por m.c.o. a partir de los datos simulados presentados en la Figura 1.8(b).

Debido a estos inconvenientes, se han propuesto otros métodos de estimación,

como los modelos de Funciones Aleatorias Intrínsecas de orden k, IRF-k (ver p.e. Chilès y Delfiner, 2012, Cap. 4), los cuales asumen una tendencia teórica polinómica de grado k (el cual es desconocido), e intenta filtrarla de los datos espaciales mediante *incrementos generalizados de orden k*. Sin embargo, este método supone imponer condiciones adicionales sobre el proceso espacial que suelen ser difíciles de comprobar en la práctica, como por ejemplo, para seleccionar el grado k de la tendencia lineal. También se puede recurrir a métodos basados en máxima verosimilitud o máxima verosimilitud restringida REML (ver p.e. Gelfand *et al.*, 2010, Cap. 4). No obstante estos procedimientos necesitan suponer que la distribución de los datos es normal, se encuentran expuestos a problemas de mala especificación de los modelos paramétricos y por lo general requieren alto coste computacional.

Alternativamente, otros estudios sugieren aproximar el sesgo a partir de correcciones en las estimaciones piloto del variograma. Por ejemplo, Beckers y Bogaert (1998) proponen corregir el sesgo calculando los residuos suponiendo una tendencia lineal, y estimando los parámetros de un modelo válido de variograma mediante un procedimiento iterativo no lineal que minimiza la diferencia entre la esperanza matemática de dicho modelo y el estimador empírico del variograma. Un enfoque similar es propuesto por Kim y Boos (2004), en el cual se propone una corrección multiplicativa del variograma empírico de los residuos (mediante una tendencia estimada por m.c.o.), para obtener una estimación monótona piloto del variograma.

Independientemente de los resultados obtenidos por cualquiera de los procedimientos paramétricos anteriores, estos se encuentran expuestos a problemas de mala especificación y pueden generar estimaciones poco flexibles del variograma y de la tendencia espacial. Considerando estas limitaciones, en el presente trabajo se plantean distintos algoritmos de estimación no paramétrica basados en la

estimación tipo núcleo, diseñados especialmente para la estimación de las componentes del proceso no estacionario en media. Estos métodos propuestos tienen la ventaja de que permiten obtener estimaciones flexibles de las componentes del modelo, que tienen en cuenta el efecto del sesgo debido al uso de residuos, y que adicionalmente no se encuentran expuestos a problemas de mala especificación de modelos.

Capítulo 2

Estimación no paramétrica de la tendencia espacial

El modelo (1.24) ha sido estudiado desde una perspectiva no paramétrica por varios autores, con el principal objetivo de estimar la función tendencia $\mu(\cdot)$ con errores correlacionados. Los distintos enfoques utilizados con este fin incluyen técnicas como los estimadores tipo núcleo, regresión tipo spline, el uso de wavelets o los desarrollos en series de Fourier. Una revisión de estos métodos se puede encontrar en Opsomer *et al.* (2001).

Dentro de los estimadores tipo núcleo para datos correlacionados, se han propuesto varios alternativas, por ejemplo, el estimador de Nadaraya-Watson, el estimador Priestley-Chao, o el estimador lineal local, el cual tiene como base la regresión polinómica local para datos independientes (ver p.e. Fan y Gijbels, 1996). A lo largo del presente estudio se propone utilizar el estimador lineal local para estimar la tendencia espacial, debido a sus propiedades teóricas que se explican en detalle en la Sección 2.1.

Por otra parte, esta estimación no paramétrica de la tendencia depende de la adecuada selección de una ventana de suavizado. En la Sección 2.2 se presentan

algunos criterios para seleccionar la matriz ventana en este contexto. Sin embargo, estos selectores de ventana deben considerar la estructura de correlación de los datos, por lo que en la Sección 2.3 se propone un método no paramétrico de estimación conjunta de la tendencia y el variograma para solventar este problema circular, y que a la vez corrige el sesgo en la estimación del variograma debido al uso directo de los residuos.

Otro aportación reseñable se encuentra en la Sección 2.4, en el cual se proponen nuevos criterios para la selección de la ventana del estimador de la tendencia bajo dependencia espacial. El comportamiento de las ventanas resultantes se compara con otros selectores de ventana mediante estudios de simulación, cuyos resultados se presentan en la Sección 2.5. Finalmente, en la Sección 2.6 se presenta una aplicación a datos reales de las técnicas no paramétricas presentadas en este capítulo.

Las principales contribuciones de este capítulo, así como algunos de los resultados, se encuentran en Castillo-Páez *et al.* (2017a).

2.1. Estimador lineal local de la tendencia

Suponiendo el modelo (1.24), el estimador lineal local para $\mu(\cdot)$ en la posición espacial $\mathbf{x}$, se obtiene como la solución para α del siguiente problema de minimización:

$$\min_{\alpha,\beta} \sum_{i=1}^{n} \left\{ Y(\mathbf{x}_i) - \alpha - \beta^T(\mathbf{x}_i - \mathbf{x}) \right\}^2 K_{\mathbf{H}}(\mathbf{x}_i - \mathbf{x}),$$

donde $Y(\mathbf{x}_i)$ corresponde a la i-ésima componente del vector $\mathbf{Y}$, observado en las posiciones $\mathcal{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ tal que $\mathbf{x}_i \in D \subset \mathbb{R}^d$, y donde $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, siendo K es una función tipo núcleo d-dimensional, y $\mathbf{H}$ es la

matriz ventana $d \times d$ simétrica no singular.

Este estimador puede escribirse explicitamente como:

$$\hat{\mu}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} = s_\mathbf{x}^t \mathbf{Y}, \tag{2.1}$$

donde $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{X}_\mathbf{x}$ es la matriz cuya i-ésima fila es igual a $(1, (\mathbf{x}_i - \mathbf{x})^t)$, $\mathbf{W}_\mathbf{x} = diag\left\{K_\mathbf{H}(\mathbf{x}_1 - \mathbf{x}), \ldots, K_\mathbf{H}(\mathbf{x}_n - \mathbf{x})\right\}$, Esta ecuación permite expresar al estimador lineal local de la tendencia como un suavizado lineal de los datos $(\mathbf{x}_i, Y(\mathbf{x}_i))$, es decir,

$$\hat{\boldsymbol{\mu}} = \mathbf{S}\mathbf{Y}, \tag{2.2}$$

donde la matriz de suavizado $\mathbf{S}$ es aquella cuya i-ésima fila corresponde a $s_\mathbf{x}^t$.

Respecto a las propiedades del estimador (2.1), Rupert y Wand (1994) estudiaron las propiedades asintóticas del estimador lineal local multivariante para el caso de datos incorrelados bajo diseño aleatorio. Por otra parte, la consistencia de estimadores tipo núcleo de la tendencia bajo dependencia ha sido ampliamente estudiada en el contexto de las series temporales. En este sentido, Francisco-Fernández y Vilar-Fernández (2001) obtuvieron expresiones para el sesgo y la varianza del estimador lineal local univariante, evidenciando el efecto de la dependencia sobre la variabilidad de dicho estimador. Siguiendo un enfoque similar al considerado en series de tiempo, el trabajo de Opsomer *et al.* (2001) estableció las propiedades asíntoticas de este tipo de estimadores para el caso bidimensional. Sin embargo, fue el estudio realizado por Liu (2001, Cap. 2), donde se analizaron las propiedades asintóticas de los estimadores multidimensionales de Nadaraya-Watson, Priestley-Chao y lineal local para datos correlacionados, y en el cual se observó que el estimador lineal local es asintóticamente insesgado y de diseño adaptativo (es decir, que su sesgo no depende de la densidad del diseño

muestral) a diferencia de los otros estimadores. Algunos de los resultados teóricos presentados por este autor, se resumen a continuación.

Denotaremos por $\rho_n(\cdot)$ a la función de correlación del término de error, tal que si $n \longrightarrow \infty$ entonces $\rho_n(\cdot) \longrightarrow 0$. Se impondrán las siguientes hipótesis:

(H1) La función $\mu(\cdot)$ es dos veces diferenciable sobre el conjunto compacto D, la varianza del error $\sigma^2 > 0$ y la función de diseño de las localizaciones $f_x(\cdot) > 0$.

(H2) $K(\cdot)$ es una función Lipschitz continua, simétrica, tal que $\int K(\mathbf{u})d\mathbf{u} = 1$, $\int \mathbf{u}K(\mathbf{u})d\mathbf{u} = 0$ y $\int \mathbf{u}\mathbf{u}^t K(\mathbf{u})d\mathbf{u} = m_2(K)\mathbf{I}$ con $m_2(K) \neq 0$.

(H3) La matriz ventana $\mathbf{H}$ es simétrica y definida positiva. $\mathbf{H} \longrightarrow 0$ cuando $n \longrightarrow \infty$. La razón $\lambda_{max}(\mathbf{H})/\lambda_{min}(\mathbf{H})$ está acotada superiormente, y $n|\mathbf{H}|\lambda^2_{min}(\mathbf{H}) \longrightarrow \infty$ conforme $n \longrightarrow \infty$, donde $\lambda_{max}(\mathbf{H})$ y $\lambda_{min}(\mathbf{H})$ corresponden a los valores propios máximo y mínimo de $\mathbf{H}$ respectivamente.

(H4) Existen constantes ρ_I, C_ρ tales que $n\int \rho_n(\mathbf{u})d\mathbf{u} \longrightarrow \rho_I$ y además se cumple que $n\int |\rho_n(\mathbf{u})|d\mathbf{u} \leq C_\rho$. Para una sucesión $\epsilon_n > 0$ la cual satisface $n^{1/d}\epsilon_n \longrightarrow \infty$, entonces $n \int\limits_{||\mathbf{u}||\geq\epsilon_n} |\rho_n(\mathbf{u})|d\mathbf{u} \longrightarrow 0$ cuando $n \longrightarrow \infty$.

La hipótesis **(H1)** impone ciertas condiciones de regularidad sobre el proceso subyacente (tendencia suficientemente suave y varianza estrictamente positiva), así como la selección aleatoria de las localizaciones muestrales. Además, dado que se aplicará la metodología no paramétrica de tipo núcleo, se asumirán ciertas propiedades habituales en este contexto sobre la función K, que se presentan en **(H2)**. Para la matriz ventana $\mathbf{H}$ se deben verificar las condiciones establecidas en **(H3)**, que son similares a las impuestas en Rupert y Wand (1994) para datos incorrelados, aunque algo más restrictivas (la extensión al caso de datos dependientes requiere que todos los elementos de H converjan a cero). La última

hipótesis implica que si $n \longrightarrow \infty$ entonces la integral de $\rho_n(\mathbf{u})$ debe converger a cero con un orden de convergencia no menor a $O(1/n)$. Esta suposición también implica que la integral $|\rho_n(\mathbf{u})|$ está esencialmente dominada por los valores de $\rho_n(\mathbf{u})$ cercanos al origen. Desde la perspectiva espacial, se puede considerar que esta situación corresponde a un contexto de dominio creciente (Cressie, 1993, p. 100), en el cual la función correlación ρ_n se mantiene fija respecto al tamaño muestral, a la vez que el soporte espacial D (y por tanto la función tendencia $\mu(\cdot)$) de las localizaciones $\mathbf{x}$ se expande.

Bajo las hipótesis (H1) a (H4), y siendo $\mathbf{x}$ un punto interior en D, (Liu, 2001, Sección 2.3.1) demostró que:

$$\mathbb{E}[\hat{\mu}_{\mathbf{H}}(\mathbf{x}) - \mu(\mathbf{x})|\mathcal{X}] = \frac{1}{2} m_2(K) tr(\mathbf{H}^2 \mathcal{H}_\mu(\mathbf{x})) + o_p(tr(\mathbf{H}^2)), \tag{2.3}$$

$$Var[\hat{\mu}_{\mathbf{H}}(\mathbf{x})|\mathcal{X}] = \frac{m_2(K)\sigma^2(1 + f_x(\mathbf{x})\rho_I)}{n|\mathbf{H}| f_x(\mathbf{x})} + o_p\left(\frac{1}{n|\mathbf{H}|}\right), \tag{2.4}$$

donde $\mathcal{H}_\mu(\mathbf{x})$ representa la matriz Hessiana de $\mu(\cdot)$ evaluada en $\mathbf{x}$ y $tr(\mathbf{A})$ es la traza de la matriz $\mathbf{A}$.

Otro aspecto importante a tener en cuenta en la estimación de la tendencia es la apropiada selección de una matriz ventana. Para el caso de datos incorrelados, esta ventana se suele seleccionar mediante técnicas de validación cruzada. Sin embargo, se sabe que cuando los datos son dependientes, estos métodos no resultan apropiados (ver, p.e Liu, 2001; Opsomer *et al.*, 2001). Estos resultados indican de manera implicíta que las ventanas óptimas dependen de la matriz de varianzas y covarianzas de los datos. En la siguiente sección se analizan algunos de los criterios que suelen utilizarse para seleccionar la matriz $\mathbf{H}$ bajo dependencia espacial.

2.2. Selección de la ventana bajo dependencia

Para seleccionar la ventana se podría emplear la aproximación habitual en estadística y tratar de minimizar el error en media cuadrática (*mean squared error*) $\text{MSE}(\hat{\mu}(\mathbf{x}), \mathbf{H}) = Var[\hat{\mu}_{\mathbf{H}}(\mathbf{x})|\mathcal{X}] + \mathbb{E}[\hat{\mu}_{\mathbf{H}}(\mathbf{x}) - \mu(\mathbf{x})|\mathcal{X}]^2$. Sin embargo, para obtener expresiones más simples que permitan analizar el efecto de la ventana, puede ser preferible considerar la aproximación asintótica. El error cuadrático medio asintótico $\text{AMSE}(\hat{\mu}(\mathbf{x}), \mathbf{H}))$ (*asymptotic mean squared error*) se obtendría como la suma del cuadrado del sesgo (2.3) y la varianza (2.4). Liu (2001, p. 46) demostró que la matriz ventana "óptima" local $\mathbf{H}_{opt}(\mathbf{x})$, obtenida minimizando el AMSE respecto a $\mathbf{H}$, es:

$$\mathbf{H}_{opt}(\mathbf{x}) = \left\{ \frac{m(K^2)\sigma^2(1 + f_x(\mathbf{x})\rho_I)|\tilde{\mathcal{H}}_\mu(\mathbf{x})|^{-1/2}}{n\, d\, m_2^2(K^2) f_x(\mathbf{x})} \right\}^{1/(d+4)} (\tilde{\mathcal{H}}_\mu(\mathbf{x}))^{-1/2}, \qquad (2.5)$$

donde $m(K^2) = \int K^2(\mathbf{u})d\mathbf{u}$ y $\tilde{\mathcal{H}}_\mu(\mathbf{x}) = \mathcal{H}_\mu(\mathbf{x})$ si $\mathcal{H}_\mu(\mathbf{x})$ es definida positiva, o si es definida negativa se iguala a $-\mathcal{H}_\mu(\mathbf{x})$. Luego, el término $(\tilde{\mathcal{H}}_\mu(\mathbf{x}))^{-1/2}$ determina la forma y la orientación del vecindario local utilizado para estimar la tendencia en un punto dado. El primer término de (2.5), en cambio, determina el tamaño de la ventana, la cual depende entre otros factores, del tamaño muestral n, la dimensión espacial d y de la función de correlación integrada ρ_I.

Por otra parte, la mayoría de las veces se desea obtener una ventana óptima global. Como criterio para medir globalmente la distancia entre la tendencia estimada y la tendencia teórica se puede emplear el error cuadrático medio integrado (*mean integrated squared error*):

$$\text{MISE}(\mathbf{H}) = \int \text{MSE}(\hat{\mu}(\mathbf{x}, \mathbf{H}))w(\mathbf{x})d\mathbf{x},$$

donde $w(\cdot) \geq 0$ es una función de pesos, normalmente introducida para reducir el efecto frontera. En el caso de diseño aleatorio se suele establecer como función de pesos la función de densidad $f_x(\mathbf{x})$. Por ejemplo, teniendo en cuenta los resultados anteriores, se podría pensar en minimizar el error cuadrático medio asintótico integrado (*asymptotic mean integrated squared error*):

$$\text{AMISE}(\mathbf{H}) = \int \text{AMSE}(\hat{\mu}(\mathbf{x}, \mathbf{H})) f_x(\mathbf{x}) d\mathbf{x}.$$

Aunque a diferencia de la ventana local, no se dispone actualmente de una expresión explícita para esta ventana.

Una aproximación del MISE($\mathbf{H}$) es el denominado MASE (*mean average squared error*), en el cual se selecciona la ventana que minimiza el promedio de los errores cuadráticos de los datos observados:

$$\text{MASE}(\mathbf{H}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[(\hat{\mu}_{\mathbf{H}}(\mathbf{x}_i) - \mu(\mathbf{x}_i))^2 \,|\mathcal{X}] w(\mathbf{x}_i)$$

Como muchos de los criterios normalmente empleados (entre ellos los de validación cruzada) tratan de aproximar este criterio, en el presente trabajo se considerará como ventana óptima la que minimice MASE($\mathbf{H}$) con $w(\cdot) = 1$, y que denotaremos por $\mathbf{H}_{\text{MASE}}$. Expresando al estimador lineal local de la tendencia $\hat{\mu}(\cdot)$, como un suavizador lineal (2.2), entonces la ventana óptima resultante bajo este criterio es aquella que minimiza:

$$\text{MASE}(\mathbf{H}) = \frac{1}{n} \mathbb{E}\left((\mathbf{SY} - \boldsymbol{\mu})^t (\mathbf{SY} - \boldsymbol{\mu}) \right).$$

El segundo término de esta ecuación corresponde a la esperanza matemática de una forma cuadrática respecto al vector ($\mathbf{SY} - \boldsymbol{\mu}$). Por tanto, el criterio anterior

puede reescribirse de la siguiente manera:

$$\text{MASE}(\mathbf{H}) = \frac{1}{n}\left(\mathbf{S}\boldsymbol{\mu} - \boldsymbol{\mu}\right)^t \left(\mathbf{S}\boldsymbol{\mu} - \boldsymbol{\mu}\right) + \frac{1}{n} tr\left(\mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t\right), \tag{2.6}$$

donde $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))^t$ y $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas teórica de los datos.

En la práctica la expresión (2.6) no puede utilizarse, pues tanto $\boldsymbol{\mu}$ como $\boldsymbol{\Sigma}$ son desconocidas. Por este motivo, en el contexto de los métodos de regresión tipo núcleo, se suele recurrir a las criterios de validación cruzada.

El método ordinario de Validación Cruzada (*leave-one-out cross-validation*) selecciona la matriz ventana óptima, escogiendo aquella matriz $\mathbf{H}$ que minimiza la siguiente expresión:

$$CV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{\mu}_{-i}(\mathbf{x}_i)\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i)}{1 - s_{ii}}\right)^2, \tag{2.7}$$

donde $\hat{\mu}_{-i}$ es el estimador de la tendencia que se obtiene al omitir el i-ésimo par $(\mathbf{x}_i, Y(\mathbf{x}_i))$, siendo s_{ii} el i-ésimo elemento de la diagonal de $\mathbf{S}$. La última igualdad en (2.7) se debe al hecho de que la suma de todos los términos en cada fila de la matriz de suavizado $\mathbf{S}$ es igual a 1.

De forma alternativa, también se puede recurrir al criterio de *validación cruzada generalizada* (GCV) propuesto por Craven y Wahba (1978). En este caso, se reemplaza el término s_{ii} de la expresión (2.7) por su promedio, donde $\sum_{i=1}^{n} s_{ii}/n = tr\left(\mathbf{S}\right)/n$. Con este criterio, la matriz ventana se obtiene minimizando:

$$GCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i)}{1 - \frac{1}{n}tr\left(\mathbf{S}\right)}\right)^2.$$

Estos selectores permiten obtener ventanas apropiadas para datos independientes, pero no son selectores eficientes bajo dependencia, puesto que no tienen

en cuenta la correlación entre los errores. Por ejemplo, en Opsomer *et al.* (2001) se observó que para datos bidimensionales correlacionados positivamente (como sucede normalmente en el caso espacial), los métodos basados en CV seleccionan ventanas muy pequeñas para la estimación tipo núcleo de la regresión. Asimismo, Liu (2001) verificó que bajo la presencia de correlación, el criterio GCV es asintóticamente sesgado y por tanto las ventanas obtenidas mediante estos criterios tienden a infrasuavizar la tendencia.

Para resolver el problema de la selección de ventana bajo dependencia, se han planteado distintos mecanismos, como por ejemplo, el propuesto por Chu y Marron (1991) el cual generaliza el criterio CV para el caso unidimensional (y que también fue propuesto posteriormente por Carmack *et al.*, 2009, de forma independiente utilizando el estimador lineal local (2.1) para estimar la tendencia). En el contexto espacial, este criterio de *validación cruzada modificada* (MCV) selecciona como ventana óptima aquella que minimice:

$$MCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{\mu}_{-N(i)}(\mathbf{x}_i)\right)^2, \tag{2.8}$$

donde $\hat{\mu}_{-N(i)}(\mathbf{x}_i)$ es el estimador de la tendencia que se obtiene al omitir las observaciones dentro de un vecindario $N(i)$ del punto de estimación $\mathbf{x}_i$. Cabe mencionar que el criterio MCV coincide con el criterio CV si se considera que $N(i) = \{\mathbf{x}_i\}$. Aunque los estudios numéricos de este tipo de estimadores obtienen buenos resultados para datos bajo dependencia, aún existe el problema de la adecuada elección del tamaño del vecindario $N(i)$. Por ejemplo, si la correlación entre los datos es muy fuerte entonces el tamaño del vecindario a ser omitido debe ser mayor, lo que implica una pérdida de información importante a la hora de estimar la tendencia espacial.

Otra alternativa, propuesta en Francisco-Fernández y Opsomer (2005), propo-

ne una corrección del criterio GCV a partir de su distribución asintótica, tratando de reducir su sesgo al incluir el efecto de la correlación espacial. Este criterio GCV *corregido* ($CGCV$) selecciona la ventana óptima, escogiendo la matriz $\mathbf{H}$ que minimiza:

$$CGCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i)}{1 - \frac{1}{n}tr\,(\mathbf{SR})}\right)^2, \tag{2.9}$$

donde $\mathbf{R}$ es la matriz de correlación tal que $\mathbf{\Sigma} = \sigma^2\mathbf{R}$, siendo $\sigma^2 = C(\mathbf{0})$ la varianza (o umbral del respectivo variograma). Aunque este criterio requiere conocer previamente la matriz de varianzas y covarianzas teórica $\mathbf{\Sigma}$, los autores verificaron el buen comportamiento del criterio CGCV a la hora de obtener ventanas óptimas para la regresión lineal local a partir de estimaciones paramétricas $\hat{\Sigma}$, incluso cuando hay problemas de mala especificación del modelo de correlación (en el caso unidimensional, este criterio también fue propuesto de forma independiente por Carmack *et al.*, 2012a, adaptando las definiciones de los grados de libertad al caso dependiente).

Por otra parte, la forma de la matriz $\mathbf{H}$ se encuentra directamente relacionada con la estructura de la función de tendencia teórica, de manera que si existe interacción entre todos los distintos componentes espaciales que la conforman, entonces es necesario aproximar la estructura completa de la matriz ventana (es decir, sus $d \times d$ elementos). Esta matriz ventana también puede depender de la estructura de dependencia, p.e. en caso de anisotropía la ventana óptima podría ser completa. Sin embargo, Liu (2001) observó que selectores más simples de dicha ventana basados en la estimación de la función de correlación pueden proporcionar resultados eficientes. En ciertos casos, se puede considerar que la matriz ventana es diagonal de orden d, donde cada uno de los elementos de dicha diagonal controla el grado de suavizado en su respectiva dimensión espacial y no hay interacción entre las componentes espaciales de la tendencia. El caso más

simple implica suponer una matriz *esférica* $\mathbf{H} = h\mathbf{I}$, de modo que el problema se reduce a la elección de un solo parámetro h, asumiendo que el grado de suavizado es el mismo en todas las dimensiones.

2.3. Estimación conjunta de la tendencia y la dependencia espacial

Suponiendo que la tendencia fue aproximada mediante el estimador lineal local (2.1), entonces se pueden calcular los *residuos o errores estimados* $\hat{\boldsymbol{\varepsilon}} = (\hat{\varepsilon}(\mathbf{x}_1), \ldots, \hat{\varepsilon}(\mathbf{x}_n))$, donde:

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\boldsymbol{\mu}} = \mathbf{Y} - \mathbf{S}\mathbf{Y} = (\mathbf{I} - \mathbf{S})\mathbf{Y}. \tag{2.10}$$

Generalmente se trata de aproximar el variograma del proceso partir de la variabilidad de estos residuos, sin embargo, es sabido que esta estimación presenta sesgos. En este caso, a partir de (2.10) se tiene que:

$$\begin{aligned} Var(\hat{\boldsymbol{\varepsilon}}) &= Var((\mathbf{I} - \mathbf{S})\mathbf{Y}) = (\mathbf{I} - \mathbf{S})\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{S})^t \\ &= \boldsymbol{\Sigma} + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma} = \boldsymbol{\Sigma} + \mathbf{B} = \boldsymbol{\Sigma}_{\hat{\varepsilon}}, \end{aligned} \tag{2.11}$$

donde la matriz cuadrada $\mathbf{B} = \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma}$, representa el sesgo.

Considerando la relación existente entre el covariograma y el semivariograma, la relación (2.11) se puede expresar en términos del variograma de los residuos:

$$Var\left(\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j)\right) = Var\left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right) + b_{ii} + b_{jj} - 2b_{ij}, \tag{2.12}$$

donde b_{ij} representa el (i, j)-ésimo elemento de la matriz $\mathbf{B}$. Es claro entonces

que estimar el variograma teórico $\gamma(\cdot)$ a partir de los residuos $\hat{\boldsymbol{\varepsilon}}$ produce un sesgo, representado a través de los elementos de la matriz $\mathbf{B}$.

Como se comentó en la Sección 1.5, varios métodos paramétricos han sido propuestos para corregir el sesgo del variograma estimado a partir de los residuos. En el contexto no paramétrico, Fernández-Casal y Francisco-Fernández (2014) proponen un proceso iterativo no paramétrico de corrección del variograma estimado a partir de residuos, similar al presentado por Beckers y Bogaert (1998), con la diferencia de que no se recurre a modelos paramétricos para estimar la función tendencia o del variograma, y se utiliza el estimador lineal local en lugar del estimador empírico del variograma .

En este caso, el estimador lineal local del variograma calculado a partir de los residuos (o simplemente *variograma estimado residual*), se obtiene al reescribir (1.16) como la solución para α del siguiente problema de minimización por mínimos cuadrados:

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i,j} \left\{ \frac{1}{2} \left(\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j)\right)^2 - \alpha - \boldsymbol{\beta}^t(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}) \right\}^2 K_{\mathbf{G}}(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}). \quad (2.13)$$

Cabe mencionar que el estimador del variograma $\hat{\gamma}(\cdot)$ obtenido mediante la expresión (2.13), también puede ser expresado de forma análoga al estimador (2.1), es decir que la estimación lineal local del variograma corresponde al suavizado lineal de los datos $(\mathbf{x}_i - \mathbf{x}_j, (\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2)$. Asimismo, la matriz ventana $\mathbf{G}$ se puede seleccionar minimizando el error cuadrático obtenido por validación cruzada de los semivariogramas estimados:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\left(\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j)\right)^2 - 2\hat{\gamma}_{-(i,j)}(\mathbf{x}_i - \mathbf{x}_j) \right)^2, \quad (2.14)$$

o, de forma alternativa, el correspondiente error cuadrático relativo:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{(\hat{\varepsilon}(\mathbf{x}_i)-\hat{\varepsilon}(\mathbf{x}_j))^2}{2\hat{\gamma}_{-(i,j)}(\mathbf{x}_i-\mathbf{x}_j)}-1\right)^2, \tag{2.15}$$

donde $\hat{\gamma}_{-(i,j)}(\cdot)$ es la estimación obtenida sin considerar $(\hat{\varepsilon}(\mathbf{x}_i)-\hat{\varepsilon}(\mathbf{x}_j))^2$.

A partir de este estimador lineal local del variograma, el procedimiento de corrección consiste en sustituir las diferencias cuadradas de los residuos del estimador lineal local (2.13) por unas corregidas teniendo en cuenta la ecuación (2.12). A partir de este estimador corregido del variograma se reestiman las matrices $\hat{\mathbf{\Sigma}}$ y $\hat{\mathbf{B}}$, y se realiza una nueva corrección sobre $\hat{\gamma}(\cdot)$. Este proceso se repite hasta obtener convergencia. Para aproximar la matriz de varianzas y covarianzas, los autores recomiendan usar modelos paramétricos (como los presentados en la Sección 1.4.2), o utilizar el enfoque no paramétrico basado en los modelos flexibles de Shapiro Botha (ver Sección 1.4.4), lo cual puede implicar un alto coste computacional en la práctica.

En el presente trabajo se propone modificar este procedimiento de corrección de manera que en cada iteración, la matriz $\hat{\mathbf{\Sigma}}$ sea calculada directamente a partir de las estimaciones pilotos del variograma, usando pseudo-covarianzas:

$$\tilde{C}(\mathbf{u}) = \tilde{C}(0) - \hat{\gamma}(\mathbf{u}),$$

donde $\tilde{C}(0) = \underset{\mathbf{u}}{\max}\,\hat{\gamma}(\mathbf{u})$, donde $\mathbf{u} = \mathbf{x}_i - \mathbf{x}_j$ para $\mathbf{x}_i, \mathbf{x}_j \in D$. Se llevaron a cabo algunas pruebas empíricas para comparar ambos procedimientos, obteniéndose resultados similares, aunque se evidenció una reducción significativa en el tiempo de cálculo con esta segunda aproximación. A modo de ejemplo, considerando el conjunto de datos espaciales (formado por 1053 observaciones) que se presenta en la Sección 3.5 y 4.4, y utilizando un equipo con procesador I7-4770, con 3.4 GHz y 16 Gb de RAM, el tiempo de computación se reducía de 29.73 segundos a 4.68

segundos, es decir existe una reducción de alrededor del 84 % en comparación con el procedimiento original.

Este procedimiento de corrección (el cual se encuentra implementado en la función `np.svariso.corr` del paquete `npsp` del software R, Fernández-Casal, 2014) se puede resumir en los siguientes pasos :

Algoritmo 2.1: Corrección NP del variograma estimado basado en residuos

1 Obtener la tendencia estimada $\hat{\boldsymbol{\mu}}$ de la forma (2.1) y calcular $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I} - \mathbf{S})\mathbf{Y}$;

2 Obtener el estimador lineal local del variograma residual (2.13), y luego calcular los valores pilotos de $\hat{\boldsymbol{\Sigma}}^{(0)}$ y $\hat{\mathbf{B}}^{(0)}$ a partir de pseudocovarianzas;

3 Para $k \geq 1$, obtener el estimador $\hat{\gamma}(\cdot)$ reemplazando en (2.13) la diferencia $(\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2$ por $(\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2 - \hat{b}_{ii}^{(k-1)} - \hat{b}_{jj}^{(k-1)} + 2\hat{b}_{ij}^{(k-1)}$);

4 Actualizar las matrices $\hat{\boldsymbol{\Sigma}}^{(k)}$ y $\hat{\mathbf{B}}^{(k)}$ a partir de $\hat{\gamma}(\cdot)$;

5 Repetir los pasos 3 y 4 hasta la convergencia numérica.

Es importante mencionar que al finalizar este algoritmo, se obtiene una estimación corregida lineal local del variograma (o *variograma estimado corregido*), al que denotaremos por $\tilde{\gamma}(\cdot)$. A este estimador se le puede ajustar un modelo paramétrico válido a elección del usuario (utilizando técnicas como m.c.p.), o incluso se puede recurrir a modelos flexibles de Shapiro-Botha (1.23).

Por otra parte, este método de corrección de sesgo del variograma estimado basado en residuos requiere previamente una estimación no paramétrica de la tendencia, la cual permanece fija durante todo el proceso. Sin embargo, el estimador lineal local de la tendencia depende de una matriz ventana, la cual se debe seleccionar considerando la dependencia espacial, generándose el mismo problema circular que se presentó en la Sección 1.5.

Para evitar este inconveniente, Fernández-Casal y Francisco-Fernández (2014) proponen un procedimiento similar, al propuesto por Neuman y Jacobson (1984)

en el contexto paramétrico, en el cual se obtiene la estimación del variograma utilizando el análisis estructural paramétrico tradicional (Sección 1.4), para luego aplicar el criterio $CGCV$ para obtener $\mathbf{H}^{(0)}$.

En el presente trabajo, se propone un enfoque similar totalmente no paramétrico, utilizando el Algoritmo 2.1 para la corrección de la estimación del variograma. Este método de estimación conjunta de la tendencia y el semivariograma consta de los siguientes pasos:

Algoritmo 2.2: Estimación conjunta NP de la tendencia y el variograma

1. Obtener una matriz ventana inicial $\mathbf{H}^{(0)}$ para estimar la tendencia. Esta ventana puede ser seleccionada utilizando los criterios para datos independientes CV o GCV;
2. Para $k \geq 1$, usando la ventana $\mathbf{H}^{(k-1)}$ estimar la tendencia $\mu(\cdot)$ mediante el estimador lineal local (2.1), y calcular los residuos $\hat{\boldsymbol{\varepsilon}}$;
3. Obtener el variograma estimado corregido $\tilde{\gamma}(\cdot)$, aplicando el procedimiento de corrección NP 2.1 a partir de los residuos anteriores. Posteriormente, ajustar un modelo válido flexible Shapiro-Botha de variograma;
4. Construir un estimador $\hat{\Sigma}^{(k)}$ para aplicar el criterio $CGCV$ (o algunos de los descritos en la siguiente sección) y así obtener una nueva ventana $\mathbf{H}^{(k)}$;
5. Repetir los pasos 2 al 4 hasta obtener convergencia.

Si bien los criterios propuestos en el paso 1 dan lugar a ventanas iniciales pequeñas para estimar la tendencia (usando por ejemplo CV), esto puede resultar ventajoso a la hora de estimar el variograma en el paso 2. Finalmente, aunque el proceso puede ser aplicado de forma iterativa, en la práctica esto se recomienda solamente en los casos cuando existan grandes diferencias entre las ventanas inicial y final de la estimación de la tendencia. En muchos casos, en especial en las aplicaciones a datos reales que se presentarán en este y en los próximos capítulos, dos iteraciones de este procedimiento fueron generalmente suficientes para obtener una ventana adecuada para estimar la tendencia.

2.4. Criterios alternativos para la selección de la ventana bajo dependencia

En esta sección se proponen nuevas aproximaciones para corregir las funciones objetivo de los criterios CV y MCV. Para esto, consideraremos el conjunto $N = \{N(1), \ldots, N(n)\}$ donde $N(i)$ es el conjunto de observaciones en el vecindario de $\mathbf{x}_i$. Con esta notación, el estimador de la tendencia que se obtiene al omitir las observaciones dentro de un vecindario $N(i)$ del punto de estimación $\mathbf{x}_i$, utilizado en el criterio MCV (2.8), se puede escribir como:

$$\hat{\mu}_{-N(i)}(\mathbf{x_i}) = \left(\mathbf{s}_{\mathbf{x_i}}^{-N(i)}\right)^t \mathbf{Y},$$

donde $\mathbf{s}_{\mathbf{x_i}}^{-N(i)}$ se corresponde con la i-ésima fila de la matriz $\mathbf{S}_{-N}$, análoga a $\mathbf{S}$, la cual se obtiene añadiendo ceros en las posiciones correspondientes a las observaciones omitidas.

Luego, la función objetivo del criterio por validación cruzada modificada MCV, se puede reescribir de la siguiente manera:

$$MCV(\mathbf{H}) = \frac{1}{n}\left(\mathbf{Y} - \mathbf{S}_{-N}\mathbf{Y}\right)^t \left(\mathbf{Y} - \mathbf{S}_{-N}\mathbf{Y}\right),$$

y por tanto:

$$\mathbb{E}\left(MCV(\mathbf{H})\right) = \frac{1}{n}\mathbb{E}[\left(\mathbf{Y} - \mathbf{S}_{-N}\mathbf{Y}\right)^t \left(\mathbf{Y} - \mathbf{S}_{-N}\mathbf{Y}\right)]. \tag{2.16}$$

Bajo esta notación, el método CV corresponde a un caso particular en el cual $N = N_1 = \{\{\mathbf{x}_1\}, \ldots, \{\mathbf{x}_n\}\}$.

Por otro lado, la media de los errores cuadráticos para el suavizado lineal $\mathbf{SY}$ (es decir, considerando todos los datos muestrales) se puede expresar como la

esperanza matemática de una forma cuadrática respecto al vector $(\mathbf{Y} - \mathbf{SY})$, y entonces:

$$\frac{1}{n}\mathbb{E}\left((\mathbf{Y}-\mathbf{SY})^t(\mathbf{Y}-\mathbf{SY})\right) = \frac{1}{n}(\mathbf{S}\boldsymbol{\mu}-\boldsymbol{\mu})^t(\mathbf{S}\boldsymbol{\mu}-\boldsymbol{\mu}) + \frac{1}{n}tr\left((\mathbf{S}-\mathbf{I})\boldsymbol{\Sigma}(\mathbf{S}-\mathbf{I})^t\right).$$

A partir de esta última expresión, y considerando el criterio MASE definido en (2.6) se obtiene:

$$\frac{1}{n}\mathbb{E}\left((\mathbf{Y}-\mathbf{SY})^t(\mathbf{Y}-\mathbf{SY})\right) = \text{MASE}(\mathbf{H}) + \sigma^2 - \frac{2}{n}tr\left(\mathbf{S}\boldsymbol{\Sigma}\right),$$

lo cual da lugar a la siguiente aproximación para (2.16):

$$\mathbb{E}\left(MCV(\mathbf{H})\right) \approx \text{MASE}(\mathbf{H}) + \sigma^2 - \frac{2}{n}tr\left(\mathbf{S}_{-N}\boldsymbol{\Sigma}\right), \tag{2.17}$$

Cabe mencionar que una aproximación similar a (2.17) fue obtenida por (Opsomer *et al.*, 2001, p. 139) considerando el criterio CV para el caso de datos correlacionados unidimensionales regulamente espaciados.

Por otra parte, como σ^2 es constante en (2.17), entonces la selección de la matriz ventana $\mathbf{H}$ se podría realizar incorporando el término $\frac{2}{n}tr\left(\mathbf{S}_{-N}\boldsymbol{\Sigma}\right)$ a las funciones objetivos de validación cruzada (2.7) y (2.8). Esto permite proponer el siguiente criterio *corregido de validación cruzada*:

$$CCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{\mu}_{-i}(\mathbf{x}_i)\right)^2 + \frac{2}{n}tr\left(\mathbf{S}_{-N_1}\boldsymbol{\Sigma}\right),$$

o, de forma más general, el criterio *corregido de validación cruzada modificada*:

$$CMCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{\mu}_{-N(i)}(\mathbf{x_i})\right)^2 + \frac{2}{n}tr\left(\mathbf{S}_{-N}\boldsymbol{\Sigma}\right),$$

para un conjunto de vecindarios N.

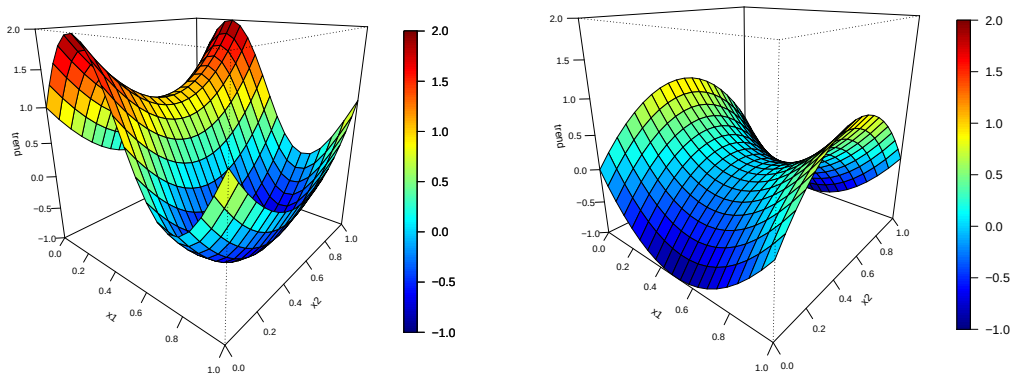
A diferencia de los criterios CV y MCV, sus versiones análogas corregidas se han construido incluyendo el término $\frac{2}{n}tr\left(\mathbf{S}_{-N}\mathbf{\Sigma}\right)$ para obtener una mejor aproximación del MASE($\mathbf{H}$). Consecuentemente, se espera que las matrices ventana obtenidas a partir de estos nuevos criterios proporcionen mejores aproximaciones a la ventana óptima $\mathbf{H}_{\text{MASE}}$ que sus versiones no corregidas.

De manera similar al método $CGCV$, estos criterios corregidos no pueden ser aplicados directamente en la práctica, pues dependen de la matriz de varianzas y covarianzas teórica que generalmente es desconocida. Sin embargo, se puede obtener una matriz $\hat{\mathbf{\Sigma}}$ utilizando los algoritmos de estimación no paramétrico descritos en la Sección 2.3, tal como se realizó en los estudios de simulación y en la aplicación a datos reales que se presentan en las siguientes secciones.

2.5. Estudios de simulación

A continuación se presentan algunos resultados de los estudios numéricos llevados a cabo para comparar el comportamiento de los diferentes criterios descritos en la sección anterior. Con este fin, se obtuvieron 1000 muestras en una rejilla regular en $D = [0,1]^2 \subset \mathbb{R}^2$. Se generaron muestras de tamaño $n = 10\times10, 15\times15$ y 20×20 de un proceso aleatorio gaussiano, bajo el modelo (1.24). Las funciones de tendencias consideradas fueron: $\mu_1(x_1,x_2) = \sin(2\pi x_1) + (2x_2-1)^2$ (no polinómica) y $\mu_2(x_1,x_2) = (2x_1-1)^2-(2x_2-1)^2$ (polinómica), tal como se muestran en las Figuras 2.1(a) y 2.1(b) respectivamente.

Los distintos efectos producidos por las coordenadas espaciales en estas funciones de tendencia nos permitió analizar el comportamiento de los selectores de ventana bajo diferentes escenarios. Por ejemplo, en el caso de datos simulados con la tendencia $\mu_2(\cdot)$ se puede asumir el uso de una ventana esférica $h\mathbf{I}$, pues el efecto de ambas coordenadas es el mismo. Sin embargo, para la tendencia $\mu_1(\cdot)$, el uso



(a) (b)

Figura 2.1: Tendencias teóricas (a) $\mu_1(x_1, x_2) = \sin(2\pi x_1) + (2x_2 - 1)^2$ (no polinómica) y (b) $\mu_2(x_1, x_2) = (2x_1 - 1)^2 - (2x_2 - 1)^2$ (polinómica).

de una ventana diagonal $\mathbf{H} = diag(h_{11}, h_{22})$ se puede considerar más razonable, pues no existe interacción entre las componentes espaciales.

Respecto a la estructura de dependencia, se consideró un modelo de variograma exponencial isotrópico $\gamma(u|\theta)$, definido en (1.17). Para analizar el efecto de los distintos grados de dependencia espacial sobre los criterios de selección de ventana, se asignaron los siguientes valores a los parámetros θ: $\sigma^2 = 1$, valores del nugget $c_0 = 0\,\%$, $20\,\%$, $40\,\%$ y $80\,\%$ de la varianza σ^2 y rango práctico $a = 0{,}3$, $0{,}6$, $0{,}9$.

Los primeros estudios de simulación se desarrollaron teniendo en cuenta las matrices de varianzas y covarianzas teóricas $\mathbf{\Sigma}$, para evitar la influencia de su estimación en los distintos resultados. Una vez generados los datos, se aplicaron los criterios de selección de ventana de la sección anterior, incluyendo la ventana óptima $\mathbf{H}_{\text{MASE}}$, para obtener estimaciones de la tendencia espacial en cada caso. Para comparar el comportamiento de cada criterio de selección de la ventana $\mathbf{H}$,

se calcularon los errores cuadrático definidos como:

$$SE_{\mathbf{H}}(\mathbf{x}) = (\hat{\mu}_{\mathbf{H}}(\mathbf{x}) - \mu(\mathbf{x}))^2.$$

Es importante mencionar que en los resultados numéricos que se presentan en esta sección no se han considerado las estimaciones correspondientes al borde de la rejilla de datos, para reducir la influencia del efecto frontera en la aproximación de los errores.

En la Tabla 2.1 se resumen los errores cuadráticos de las estimaciones para μ_1 y μ_2 con ventana diagonal $\mathbf{H}$ y ventana esférica $h\mathbf{I}$ respectivamente, para datos generados con $n = 20 \times 20$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$. Cabe indicar que los criterios indicados como MCV_k y $CMCV_k$ corresponden a los selectores MCV y $CMCV$ respectivamente, en los cuales se han omitido k observaciones adicionales en cada dirección (p.e. si k es igual a 1 o 2, significa que se han omitido 9 o 25 valores en torno al punto central, respectivamente).

Tabla 2.1: Estadísticos de los errores cuadráticos de las estimaciones de las tendencias μ_1 y μ_2 usando $\mathbf{\Sigma}$ teórica, donde $n = 20 \times 20$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$.

	$\mathbf{H}$ para μ_1			$h\mathbf{I}$ para μ_2		
Criterios	Media	Mediana	Desv. Est.	Media	Mediana	Desv. Est.
$MASE$	0.334	0.146	0.489	0.233	0.100	0.355
GCV	0.602	0.267	0.874	0.599	0.266	0.870
CV	0.562	0.250	0.814	0.562	0.251	0.813
MCV_1	0.466	0.206	0.682	0.451	0.199	0.664
MCV_2	0.402	0.177	0.589	0.363	0.156	0.542
$CGCV$	0.370	**0.164**	0.534	0.254	0.108	0.389
CCV	**0.368**	**0.165**	**0.525**	**0.249**	**0.106**	**0.381**
$CMCV_1$	0.375	0.166	0.542	0.262	0.111	0.401
$CMCV_2$	0.377	0.167	0.545	0.274	0.116	0.417

En esta tabla se observa que los valores de error más cercanos a los de la ventana óptima $\mathbf{H}_{\text{MASE}}$ corresponden a los criterios CCV y $CGCV$, e incluso los criterios corregidos $CMCV_k$ presentan mejores resultados que sus versiones sin

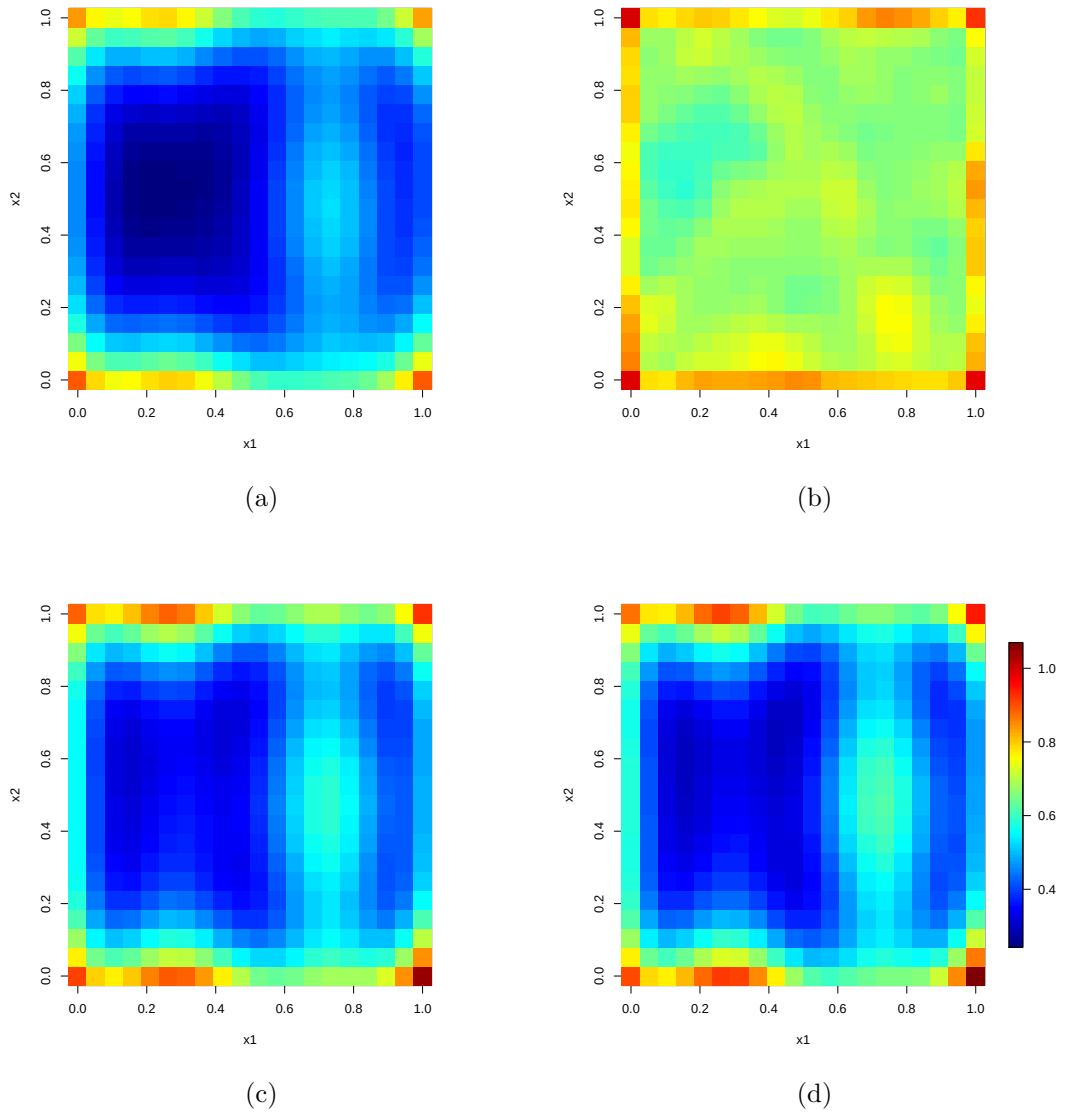


Figura 2.2: Superficies de errores cuadráticos medios de las estimaciones de la tendencia, obtenidas mediante los criterios (a) MASE, (b) CV, (c) CGCV, y (d) CCV, usando $\boldsymbol{\Sigma}$ teórica, donde $n = 20 \times 20$, $\mu = \mu_1$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$

corregir. Por otra parte, los criterios CV y GCV son los que presentan errores más elevados, lo cual era de esperar pues estos selectores no tienen en cuenta la dependencia espacial. Cabe mencionar que los errores para el caso de la ventana diagonal son mayores que los correspondientes a la ventana esférica, debido posiblemente a que en este último caso solo se requiere estimar un parámetro h y

que la función de tendencia tiene menos variabilidad.

Conclusiones similares se obtienen al analizar las Figuras 2.2(a), 2.2(b), 2.2(c) y 2.2(d) donde se muestran las superficies de errores cuadráticos de las estimaciones de la tendencia μ_1 obtenidas mediante los criterios $MASE$, CV, $CGCV$ y CCV respectivamente. Aquí se verifica que los criterios que toman en cuenta la dependencia proporcionan errores más parecidos a los que se obtienen con el criterio teórico $MASE$, a diferencia de lo que sucede con el criterio CV.

El efecto del tamaño muestral sobre la estimación de la tendencia obtenida por estos criterios se puede observar en la Tabla 2.2. Cuando n es pequeño, el selector MCV obtiene valores cercanos al óptimo $\mathbf{H}_{\text{MASE}}$. Pero conforme aumenta el tamaño muestral, los errores cuadráticos confirman que los criterios que toman en cuenta la estructura de la dependencia espacial proporcionan estimaciones más precisas de la tendencia, en especial los selectores CCV y $CGCV$. Este comportamiento pone en evidencia el efecto del tamaño de la muestra sobre la elección adecuada del vecindario a omitir en los criterios MCV_k. En general los criterios MCV_k tienden a seleccionar la ventana más pequeña posible, la cual está relacionada con el tamaño del vecindario. Para un valor k fijo, el vecindario es más grande para valores pequeños de n y por tanto también la correspondiente ventana mínima (obteniéndose buenos resultados en ese caso). Sin embargo, cuando n es grande, la ventana mínima es más pequeña y por eso se deben considerar valores de k más grandes para obtener buenos resultados.

Por otra parte, la Tabla 2.3 refleja el efecto de diferentes grados de dependencia espacial sobre los estadísticos de los errores cuadráticos, considerando distintos valores para el rango práctico a (un comportamiento parecido se observa al variar la dependencia espacial a través del efecto nugget). En esta tabla se observa nuevamente que los criterios corregidos muestran un comportamiento más satisfactorio, en comparación con los criterios tradicionales. Nuevamente, los mejores niveles

Tabla 2.2: Resumen de los errores cuadráticos de las estimaciones de la tendencia, por tamaño muestral, usando $\mathbf{\Sigma}$ teórica, donde $\mu = \mu_1$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$.

	n=10		n=15		n=20	
Criterios	Media	Desv. Est.	Media	Desv. Est.	Media	Desv. Est.
$MASE$	0.326	0.449	0.324	0.467	0.334	0.489
GCV	0.477	0.679	0.552	0.791	0.602	0.874
CV	0.403	0.563	0.494	0.701	0.562	0.814
MCV_1	**0.359**	**0.482**	0.393	0.559	0.466	0.682
MCV_2	0.418	0.535	**0.354**	**0.505**	0.402	0.589
$CGCV$	0.385	0.523	0.368	0.513	0.370	0.534
CCV	0.387	0.515	0.365	0.508	**0.368**	**0.525**
$CMCV_1$	0.371	0.494	0.370	0.519	0.375	0.542
$CMCV_2$	0.424	0.542	0.361	0.508	0.377	0.545

Tabla 2.3: Estadísticos de los errores cuadráticos de las estimaciones de la tendencia, por rango práctico, usando $\mathbf{\Sigma}$ teóricas, donde $\mu = \mu_1$, $n = 20 \times 20$, $\sigma^2 = 1$ y $c_0 = 20\,\%$.

	$a = 0{,}3$		$a = 0{,}6$		$a = 0{,}9$	
Criterios	Media	Desv. Est.	Media	Desv. Est.	Media	Desv. Est.
$MASE$	0.184	0.262	0.334	0.489	0.439	0.648
GCV	0.480	0.694	0.602	0.874	0.656	0.959
CV	0.410	0.585	0.562	0.814	0.631	0.924
MCV_1	0.280	0.406	0.466	0.682	0.559	0.824
MCV_2	0.220	0.316	0.402	0.589	0.506	0.748
$CGCV$	**0.211**	0.295	0.370	0.534	0.477	0.697
CCV	**0.212**	**0.291**	**0.368**	**0.525**	**0.474**	**0.688**
$CMCV_1$	0.213	0.299	0.375	0.542	0.481	0.705
$CMCV_2$	**0.212**	0.297	0.377	0.545	0.484	0.709

de error se obtienen con los criterios $CGCV$ y CCV, aunque éste último parece dar mejores resultados a medida que la dependencia espacial es mayor (cuando el rango es grande o el nugget es pequeño). Cabe indicar que como cabría esperar, a medida que el rango aumenta (y por tanto la dependencia espacial), la media de los errores cuadráticos de la ventana óptima $\mathbf{H}_{\text{MASE}}$ también se incrementa.

Asimismo, los resultados de estimación de la ventana esférica $h\mathbf{I}$ se pueden visualizar en la Figura 2.3, donde se muestran aproximaciones de las funciones de densidad de $log_{10}(h) - log_{10}(h_{\text{MASE}})$ para los distintos criterios de selección de la ventana. En esta figura se puede apreciar un comportamiento adecuado de los

criterios $CGCV$ (línea continua) y CCV (línea discontinua), dado que los valores representados se concentran principalmente alrededor del valor cero, mientras que el resto de selectores muestran un desplazamiento de la densidad hacia ventanas más pequeñas y mayor variabilidad.

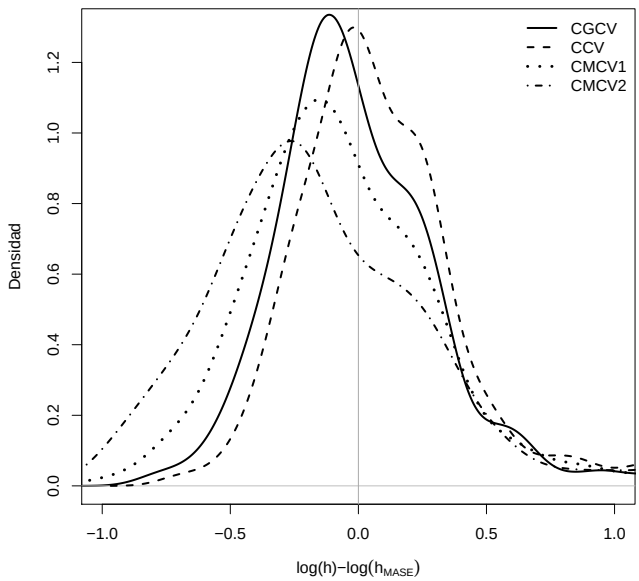


Figura 2.3: Estimación de la función de densidad de $log_{10}(h) - log_{10}(h_{\text{MASE}})$, usando los criterios $CGCV$, CCV, $CMCV_1$, $CMCV_2$ para estimar la ventana esférica $h\mathbf{I}$, con las matrices de varianzas y covarianzas teóricas, donde $\mu = \mu_2$, $n = 20 \times 20$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$.

Si bien hasta el momento, se observa que los criterios que toman en cuenta la estructura de correlación espacial proporcionan mejores aproximaciones a la ventana óptima frente a los otros selectores de ventana, en especial los selectores $CGCV$ y CCV, es importante mencionar que todos los estudios anteriores se realizaron considerando la matriz de varianzas y covarianzas teórica $\mathbf{\Sigma}$ conocida. Sin embargo, en el contexto aplicado esta matriz debe ser estimada, usualmente a partir de los residuos y por tanto se puede recurrir a los procedimientos de corrección de sesgo y de estimación conjunta no paramétrica presentados en la Sección 2.3.

Seguidamente se procedió a examinar el comportamiento de los criterios $CGCV$ y CCV en un proceso de estimación completo, es decir, que requiera la estimación de la matriz de varianzas y covarianza. Con este objeto, se realizó un estudio de simulación para comparar ambos criterios a la hora de estimar las tendencias teóricas μ_1 y μ_2 a partir de las ventanas diagonal y esférica respectivas. En cada simulación, se obtuvieron los residuos de la forma (2.10) donde $\hat{\boldsymbol{\mu}}$ se calculó con la ventana $\mathbf{H}_{\text{MASE}}$. Luego, se obtuvo el estimador lineal local del variograma $\hat{\gamma}(u_i)$ a partir de estos residuos, en los saltos $u_i = l * i$ con $i = 1, \ldots, q$, siendo l la distancia mínima entre las localizaciones en D, $q = \lfloor 0{,}55\sqrt{2}/l \rfloor$ donde $\lfloor r \rfloor$ denota la parte entera de r. La ventana g utilizada para estimar el variograma se seleccionó minimizando el error cuadrático definido en (2.14). La versión corregida $\tilde{\gamma}(u_i)$ se calculó utilizando el procedimiento de estimación de corrección NP de sesgo 2.1. Luego, ajustando modelos flexibles de Shapiro-Botha a ambas estimaciones del variograma residual y corregido, fue factible construir las matrices estimadas $\hat{\boldsymbol{\Sigma}}_{\hat{\varepsilon}}$ y $\hat{\boldsymbol{\Sigma}}$ respectivamente.

En la Figura 2.4 se muestran tanto el variograma teórico, como las medias obtenidas por simulación para $\hat{\gamma}(\cdot)$ y para $\tilde{\gamma}(\cdot)$. En este gráfico se puede visualizar el efecto del sesgo sobre el semivariograma lineal local residual, a la vez que se verifica que el procedimiento de corrección del sesgo permite obtener estimaciones corregidas bastante aproximadas al semivariograma teórico, en especial en saltos cercanos al origen, lo cual es de especial interés a la hora de realizar inferencias sobre el proceso espacial.

En la Tabla 2.4 se muestran los estadísticos de los errores cuadráticos obtenidos con los distintos criterios, verificándose en este caso que los criterios corregidos proporcionan mejores resultados frente a los criterios residuales (alrededor de un 8 % de reducción de la media del error cuadrático).

En la Figura 2.5 se presenta el boxplot de los errores cuadráticos de $\hat{\mu}_2$ para

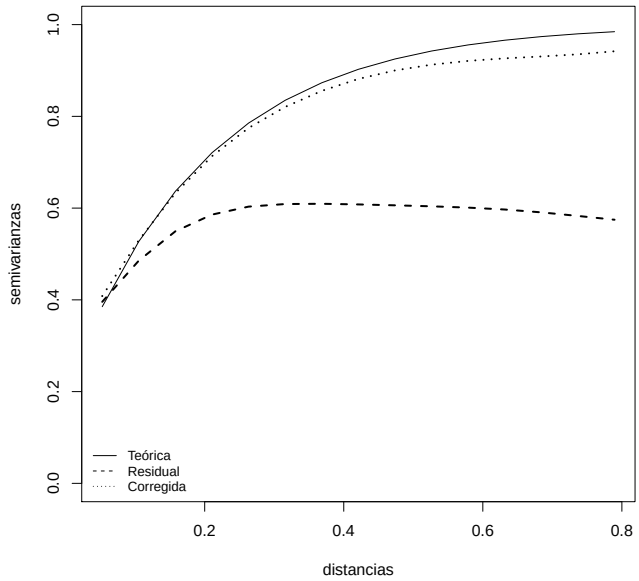


Figura 2.4: Semivariograma (a) teórico (línea continua), y medias del semivariograma (b) lineal local estimado a partir de los residuos (línea discontinua) y (c) corregido (línea de puntos), donde $\mu = \mu_1$, $n = 20 \times 20$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$.

Tabla 2.4: Estadísticos de los errores cuadráticos de las estimaciones de las tendencias μ_1 y μ_2, usando matrices de varianzas y covarianzas estimadas con procedimiento basado en residuos corregido $\hat{\mathbf{\Sigma}}$ y sin corrección (residuales) $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$, con $\sigma^2 = 1$, $n = 20 \times 20$, $a = 0{,}6$ y $c_0 = 20\,\%$.

Ventana	$\mathbf{H}$ para μ_1				$h\mathbf{I}$ para μ_2			
Método	Residuales		Corregidos		Residuales		Corregidos	
Criterios	*CGCV*	*CCV*	*CGCV*	*CCV*	*CGCV*	*CCV*	*CGCV*	*CCV*
Media	0.402	0.371	0.356	**0.352**	0.301	0.273	0.246	**0.238**
Mediana	0.176	0.161	**0.157**	**0.157**	0.131	0.118	0.105	**0.101**
Desv. Est.	0.592	0.546	0.517	**0.507**	0.453	0.413	0.374	**0.362**

el criterio $MASE$, y los criterios $CGCV$ y CCV residuales y corregidos. Aquí se observa que, si no se realiza la corrección de sesgo, al parecer el criterio CCV proporciona mejores estimaciones que el $CGCV$. Sin embargo, los niveles de errores mejoran en ambos casos, si las ventanas se obtienen con la matriz de varianzas y covarianzas corregida. Por otra parte, se observa que el gráfico de cajas correspondiente al criterio CCV corregido es muy similar al obtenido mediante el criterio teórico $MASE$, lo que sugiere que difícilmente otro selector de ventana

podría reproducir de mejor manera dicho comportamiento teórico.

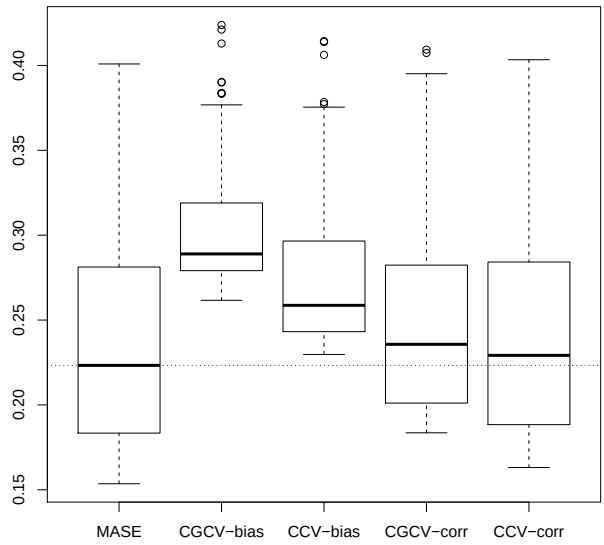


Figura 2.5: Boxplot de errores cuadráticos de las estimaciones de la tendencia, obtenidas mediante los criterios MASE, CGCV y CCV residuales y corregidos, para $\mu = \mu_2$, $n = 20 \times 20$, $a = 0{,}6$, $\sigma^2 = 1$ y $c_0 = 20\,\%$.

Los estudios numéricos realizados permiten resaltar el buen comportamiento de los criterios que tienen en cuenta la matriz de varianzas y covarianzas de los datos, tanto si se utiliza la matriz teórica o la estimada, en especial de los criterios $CGCV$ y CCV. En general, se observó que este último criterio proporciona estimaciones de la tendencia ligeramente mejores cuando el grado de correlación espacial es moderado o alto. Cuando se estima esta matriz, se puede concluir que el uso directo de los residuos sin corrección parece producir una subestimación de la variabilidad de pequeña escala, y en consecuencia, tanto el criterio $CGCV$ como el CCV no corregidos proporcionan ventanas más pequeñas. Eso pone en evidencia el efecto del sesgo, el cual parece ser corregido de forma adecuada mediante los procedimientos de corrección propuestos en la Sección 2.3. Cuando se utilizan las matrices de varianzas y covarianzas corregidas, las estimaciones mejoran en ambos casos.

2.6. Aplicación a datos reales

En esta sección, se emplearán los criterios de selección descritos anteriormente para obtener matrices ventana para la estimación lineal local de la tendencia en un contexto real. Primero se analizará el conjunto de datos utilizado en el capítulo anterior, y posteriormente se introducirá un nuevo conjunto de datos que mide el valor total mensual de las precipitaciones en EEUU.

Cabe mencionar que para realizar todas las aplicaciones a datos reales que se presentan a lo largo del presente trabajo, se desarrollaron códigos de programación con el software *R* recurriendo a varias funciones implementadas en el paquete *npsp* de Fernández-Casal (2014). Este paquete utiliza binning lineal para obtener las estimaciones de la tendencia en rejillas regulares, lo que permite acelerar los tiempos de computación, en especial cuando el conjunto de datos original tiene un gran número de observaciones irregularmente espaciadas. Cuando el número de datos es pequeño, como la rejilla binning coincide con la de estimación, el número de nodos binning puede ser superior al tamaño muestral. Incluso en este caso, el uso de binning suele producir una disminución significativa en los tiempos de computación, debido a la reducción de las evaluaciones de la función núcleo, a la facilidad para determinar los vecindarios para el ajuste local y a que en los cálculos solo se incluyen los nodos con peso binning no nulo. La rejilla binning facilita también establecer los vecindarios locales con los criterios MCV_k y $CMCV_k$. Por ejemplo, en los ejemplos siguientes, el subndice k indica el número adicional de nodos binning en cada dirección cuyas observaciones son omitidas en la estimación (de forma similar a como se procedió en el estudio de simulación).

Datos de Concentración de zinc en río Meuse

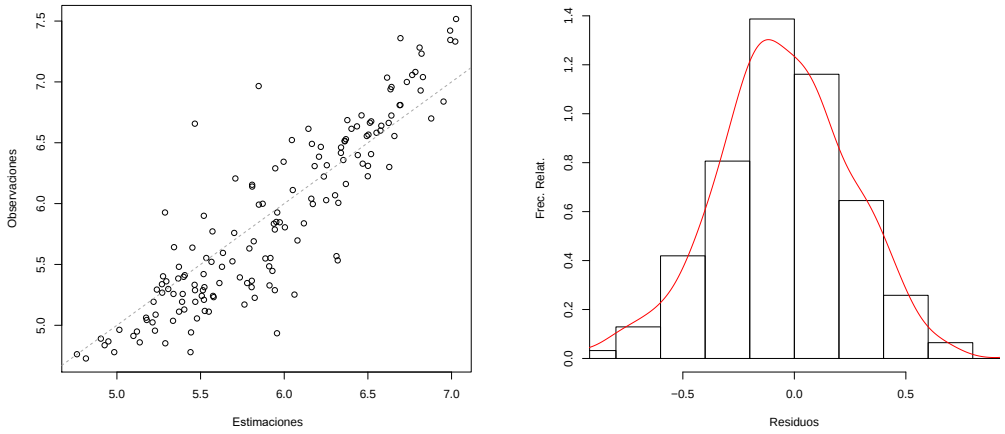
En este apartado analizaremos el conjunto de las 155 mediciones de concentra-

ción de zinc (en ppm, medidas en escala logarítmica), introducidos en el capítulo anterior, considerando en este caso una rejilla binning regular de 25×25 nodos. De acuerdo con el análisis exploratorio de estos datos, (ver Sección 1.1), se consideró adecuado aproximar una ventana diagonal $\mathbf{H} = diag(h_{11}, h_{22})$ Para obtener esta matriz ventana, se aplicaron en primer lugar los métodos tradicionales que no requieren la estimación de la matriz de varianzas y covarianzas (GCV, CV, MCV_k). Se tomaron valores de $k = 1, 2$ que implican eliminar aproximadamente el 3,6 % y el 10 % de los datos binning respectivamente.

Para los criterios que requieren el uso de $\hat{\mathbf{\Sigma}}$, se aplicó el procedimiento descrito en la Sección 2.3. De este modo, en primer lugar se estima la tendencia por suavizado lineal a partir de una ventana piloto obtenida por MCV_1. Para analizar si el comportamiento de los residuos obtenidos a partir de esta estimación piloto es adecuado, se realizaron varios análisis exploratorios, como los que se muestran en las Figuras 2.6(a) y 2.6(b). En estos gráficos se observa que la tendencia estimada se ajusta de forma razonable (el coeficiente de correlación estimado es de 0.806) e incluso se observa que los residuos tienen un comportamiento aproximadamente normal.

A partir de los correspondientes residuos, se obtiene el estimador lineal local del variograma, tomando saltos de la forma $u = 0{,}5 * l$ con $l = 1, \ldots, 30$ km., y utilizando una ventana g seleccionada mediante el criterio de error cuadrático de validación cruzada. Luego se empleó el Algoritmo 2.1 para obtener un estimador corregido del variograma. Posteriormente, se ajustó un modelo válido de Shapiro Botha (S-B) de variograma y se obtuvo la matriz de varianzas y covarianzas necesaria para aplicar los criterios corregidos de selección de ventana. (ver Figura 2.7). Las ventanas resultantes en los criterios considerados se presentan en la Tabla 2.5.

Es importante señalar que las ventanas más pequeñas corresponden a los



(a) (b)

Figura 2.6: (a) Gráfico de Estimaciones de la tendencia con ventana piloto MCV_1 vs. Datos observados, y (b) histograma de los residuos respectivos.

Tabla 2.5: Ventanas diagonales para la estimación de la tendencia de los datos de log(zinc), seleccionadas por cada uno de los criterios considerados.

h_{ii}	GCV	CV	MCV_1	MCV_2	$CGCV$	CCV	$CMCV_1$	$CMCV_2$
h_{11}	0.219	0.317	0.533	0.700	4.350	4.350	4.350	4.350
h_{22}	0.394	0.422	0.568	0.731	3.750	3.370	3.340	3.610

criterios que no tienen en cuenta la dependencia, GCV y CV. Por otra parte, todos los criterios corregidos propuestos en este trabajo dan lugar a ventanas mucho más grandes, y por tanto, proporcionan estimaciones mucho más suaves para la tendencia espacial (prácticamente lineales). Este comportamiento se puede observar al comparar las Figuras 2.8(a) y 2.8(b), en las cuales se presentan las tendencias aproximadas utilizando el estimador lineal local con la ventana piloto obtenida a partir del criterio MCV_1 y con el método $CGCV$, respectivamente.

Datos de Total de Precipitaciones mensuales en EEUU

En el presente apartado, se aplican los distintos criterios de selección de ventana para estimar la tendencia no paramétrica de un conjunto de datos que miden

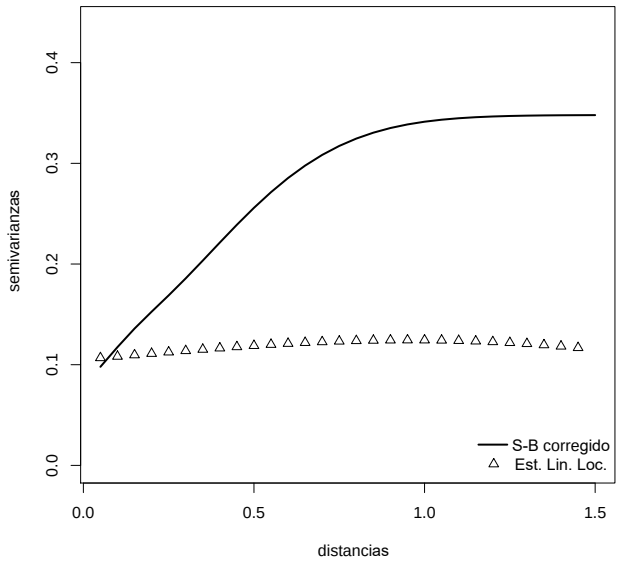
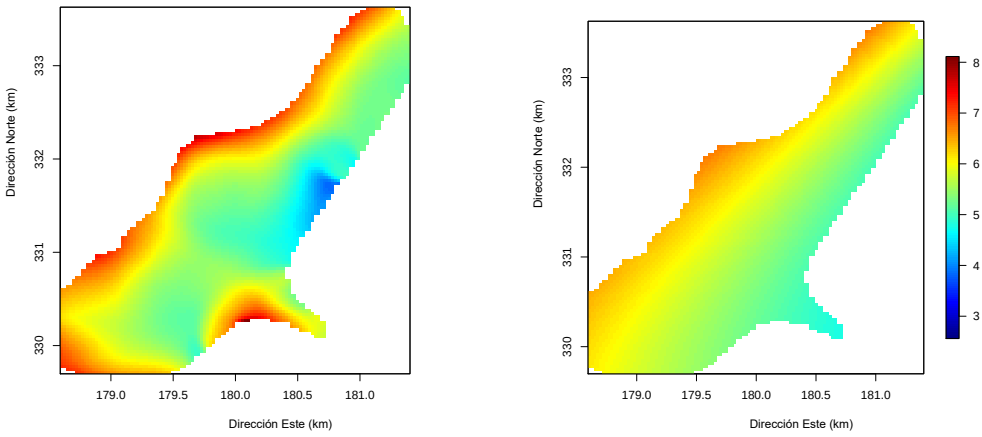


Figura 2.7: Estimación lineal local del Variograma sesgado de los residuos y variograma corregido y ajustado a un modelo S-B.



(a) (b)

Figura 2.8: Tendencia lineal local estimada a partir de una ventana diagonal seleccionada por: (a) el metodo MCV_1, y (b) el metodo $CGCV$.

el total de precipitaciones (en pulgadas de lluvia), registradas sobre 1053 localizaciones situadas en la parte continental de Estados Unidos de Norteamérica. Se puede acceder a los datos, que se muestran en la Figura 2.9, a través del sitio web

http://www.ncdc.noaa.gov.

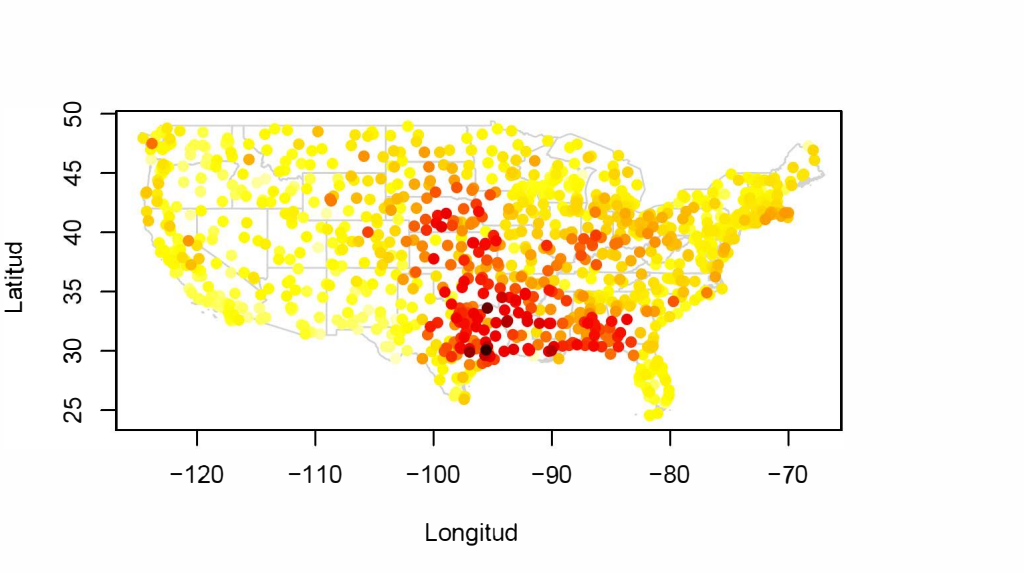
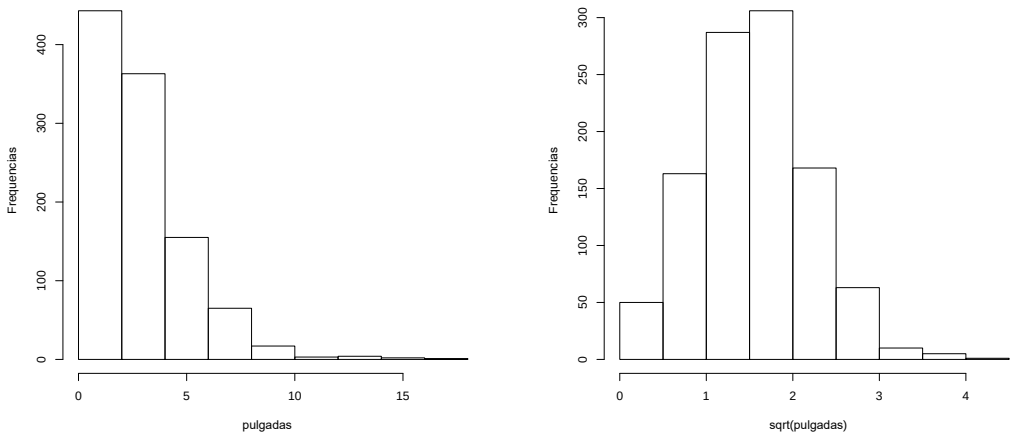


Figura 2.9: Distribución espacial de los datos observados sobre el total de precipitaciones (medidas en pulgadas) registradas en USA durante Marzo 2016



(a) (b)

Figura 2.10: Histograma de (a) los datos muestrales del total de precipitaciones en pulgadas de lluvia y (b) de los datos transformados en raíz cuadrada de pulgadas de lluvia, registradas en USA durante Marzo 2016

Luego de un análisis exploratorio, donde se confirmó la asimetría de los datos muestrales, estos fueron transformados aplicando su raíz cuadrada (ver Figuras

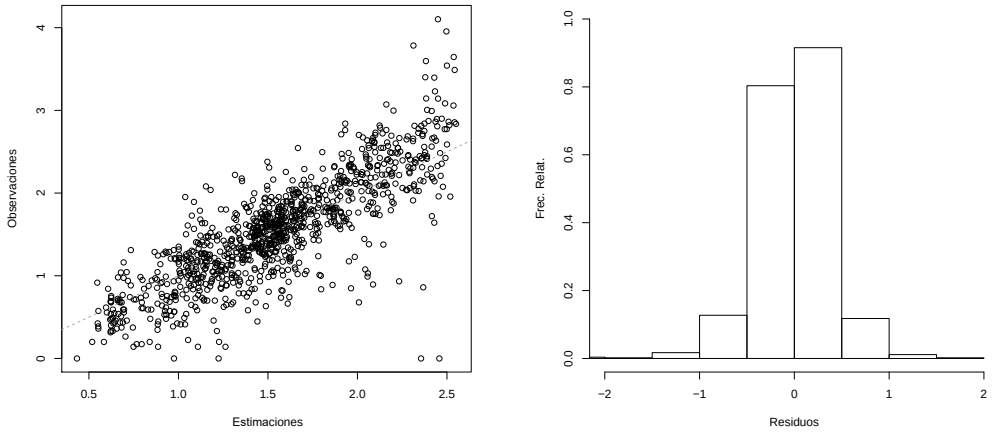
2.10(a) y 2.10(b)). De manera similar al ejemplo anterior, se consideró una matriz ventana diagonal para estimar la tendencia. Por otra parte, debido a que la matriz de varianzas y covarianzas teóricas es desconocida, se recurrió al Algoritmo 2.1 para obtener su estimación corregida. Igualmente, debido a que los datos se encuentran irregularmente espaciados, se recurrió a utilizar una rejilla binning de estimación de 30×30.

En primer lugar, la estimación piloto de la tendencia se obtuvo seleccionando una ventana piloto inicial $\mathbf{H}^{(0)}$ por MCV_1. Luego de obtener y analizar los residuos correspondientes (ver Figuras 2.11(a) y 2.11(b)), se observó que la hipótesis de isotropía era razonable, y por tanto, se estimó el variograma lineal local isotrópico $\hat{\gamma}(u_i)$ (2.13) utilizando una ventana $g = 2{,}3$ grados, seleccionada mediante la minimización del criterio de error cuadrático relativo (2.15). Se consideraron además los saltos equidistantes $u_i = i/6$ grados, con $i = 1, \ldots, 60$. A partir de esta estimación del variograma, se obtuvo su versión corregida mediante el procedimiento de corrección de sesgo del variograma, y posteriormente se le ajustó un modelo válido de Shapiro-Botha. El efecto de la corrección del sesgo en el variograma se puede observar en la Figura 2.12. A partir de la estimación corregida del variograma es factible aproximar la matriz de correlación $\mathbf{R}$, para posteriormente aplicar los distintos criterios de selección de ventana que toman en cuenta la estructura de dependencia de los datos.

Tabla 2.6: Ventanas diagonales para la estimación de la tendencia para los datos del total de precipitaciones en EEUU, durante Marzo 2016 (en sqrt(pulgadas)), seleccionadas por cada uno de los criterios considerados.

h_{ii}	GCV	CV	MCV_1	MCV_2	$CGCV$	CCV	$CMCV_1$	$CMCV_2$
h_{11}	2.936	4.893	6.851	8.808	10.087	10.586	9.060	8.882
h_{22}	1.263	2.105	2.946	7.557	18.396	19.244	11.769	9.100

En la tabla 2.6 se muestran las distintas ventanas estimadas con los criterios



(a) (b)

Figura 2.11: (a) Gráfico de Predicciones de tendencia estimada con ventana piloto MCV_1 vs. Datos observados, y (b) histograma de los residuos respectivos.

analizados en este capítulo. Como cabría esperar, se observa un comportamiento parecido al presentado en los estudios de simulación, donde las ventanas seleccionadas por los criterios que no toman en cuenta la dependencia espacial son mucho más pequeñas que las obtenidas por los criterios corregidos. Además se verifica que los criterios $CGCV$ y CCV proporcionan las ventanas más grandes (aunque sin llegar a generar a sobresuavizar los datos, a diferencia del ejemplo anterior) y son bastantes similares entre sí. Finalmente, en las Figuras 2.13(a) y 2.13(b) se representas las tendencias estimadas con las ventanas MCV_1 y $CGCV$ para los datos considerados, donde al parecer se evidencia el efecto de infrasuavizado que se produce con la ventana que no tiene en cuenta la dependencia espacial.

A modo de conclusión, se puede establecer que en el caso de datos correlados, los criterios de selección de ventana $CGCV$ y CCV son los más adecuados para estimar la tendencia del proceso espacial. En los estudios de simulación se evidencia que los métodos que no tienen en cuenta la dependencia tienden a proporcionar ventanas mucho más pequeñas. Además, estos resultados numéricos

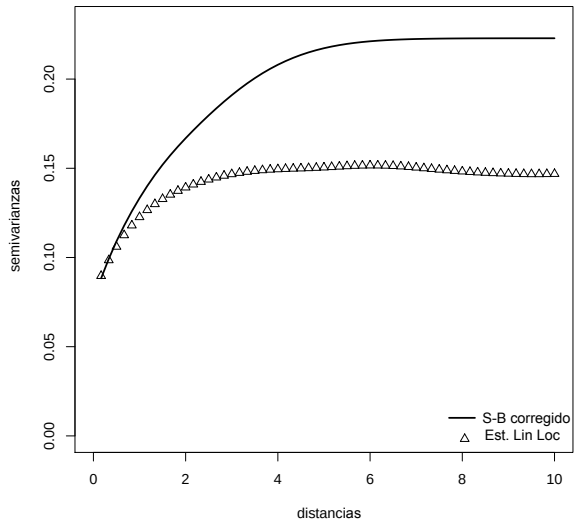
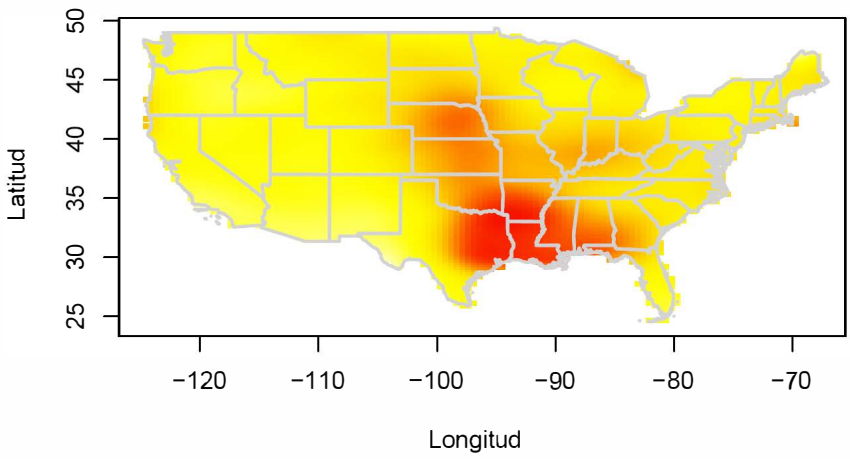


Figura 2.12: Estimación lineal local del Variograma sesgado de los residuos y variograma corregido y ajustado a un modelo S-B, para los datos de total de precipitación en EUU durante Marzo 2016.

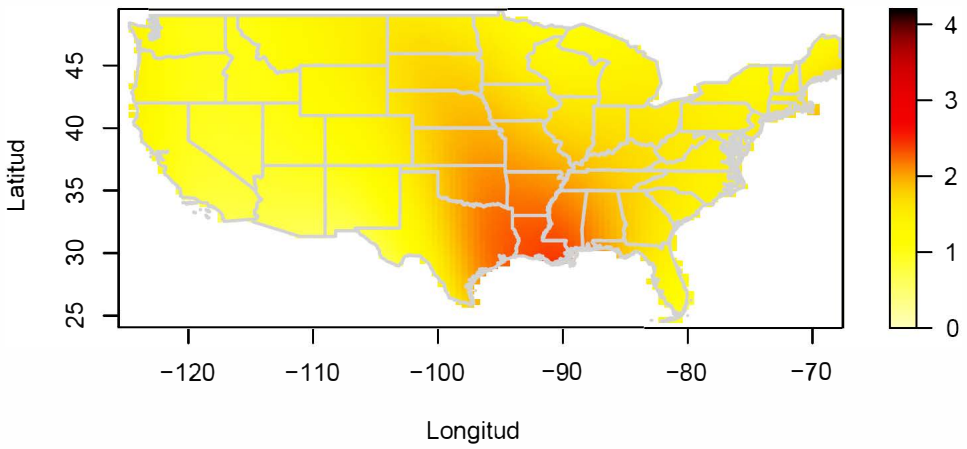
pusieron de manifiesto el efecto del sesgo debido al uso directo de residuos en los selectores de ventana. Por tal motivo, es recomendable utilizar el procedimiento de corrección de sesgo para el estimador del variograma con objeto de evitar seleccionar ventanas demasiado pequeñas.

Por otro lado, se verifica que gracias a los métodos no paramétricos expuestos en este capítulo se puede obtener una estimación más precisa y flexible del variograma $\gamma(\cdot)$, el cual puede ser utilizado para realizar predicciones espaciales. Sin embargo, no es factible realizar inferencia directamente a partir de este estimador. En el capítulo siguiente del presente documento se propone un método bootstrap que permite aproximar la variabilidad de este estimador, permitiendo por ejemplo, la construcción de bandas de confianza y mapas de probabilidad o riesgo no paramétricos.

Es importante mencionar que la mayoría de las rutinas de programación utilizadas en estos estudios se encuentran disponibles en el paquete `npsp` de `R`



(a)



(b)

Figura 2.13: Estimación lineal local de la tendencia con ventanas seleccionadas mediante criterios (a) MCV_1, y (b) $CGCV$, para los datos del total de precipitaciones (en raíz cuadrada de pulgadas de lluvia) registradas en EEUU durante Marzo 2016

(Fernández-Casal, 2014), como es el caso de la función `loc.pol` para obtener el estimador lineal local de la tendencia y la función `hcv.data` para seleccionar las ventanas óptimas con todos los criterios analizados en el presente capítulo.

Capítulo 3

Métodos bootstrap para procesos geoestadísticos

En este capítulo se propone un método bootstrap que reproduce de forma adecuada la variabilidad de un proceso espacial con tendencia no constante. Este método permite realizar inferencias sobre las características de dicho proceso, aunque en este trabajo nos centramos en el estudio del variograma. En la Sección 3.1 se presentan los métodos bootstrap disponibles en la literatura estadística, así como la aproximación bootstrap utilizada para estimar el sesgo y la precisión de un estimador. En la Sección 3.2 se introduce el método bootstrap no paramétrico (NPB), siguiendo la misma idea del procedimiento de corrección de sesgo presentado en el capítulo anterior. El comportamiento de este método se analiza mediante estudios numéricos para realizar inferencia puntual sobre el variograma, cuyos resultados se presentan en la Sección 3.3. Asimismo, mediante el bootstrap no paramétrico propuesto es factible construir un método que permite estimar la función de riesgo (denominación que procede del contexto medioambiental), definida como la probabilidad (incondicional) de que la variable de estudio exceda un valor límite o umbral en una ubicación determinada. Si se aproxima el

valor de la función de riesgo para las distintas localizaciones de una región, se podría construir el mapa de riesgo de dicha región. En la Sección 3.4 se introduce el algoritmo para la construcción de dichos mapas de riesgo. Finalmente, en la Sección 3.5 se muestra una aplicación a datos reales de los métodos propuestos en el presente capítulo. Las principales contribuciones y algunos de los resultados presentados en este capítulo se encuentran en Castillo-Páez *et al.* (2017b) y Fernández-Casal *et al.* (2017a).

3.1. Métodos de remuestreo para datos dependientes

Como se mencionó en el capítulo anterior, para un proceso espacial dado por el modelo (1.24), la estructura de dependencia se puede caracterizar a partir del variograma (o covariograma). Esto implica que la precisión del variograma estimado influye directamente sobre la inferencia en los componentes del modelo, especialmente en la varianza kriging (ver p.e. Sección 1.3.4), sobre la estimación de la tendencia (Sección 2.2), o incluso en la simulación de datos espaciales. Sin embargo, aunque el procedimiento de corrección NP de sesgo (Algoritmo 2.1) permite obtener estimaciones del variograma que reproducen de manera adecuada la variabilidad del proceso de error, no es factible realizar inferencias directamente a partir de él.

En el presente capítulo se propone un método bootstrap no paramétrico especialmente diseñado para procesos espaciales bajo tendencia no constante. Este método utiliza el procedimiento de corrección del sesgo propuesto en el capítulo anterior, para obtener estimaciones flexibles de las matrices de varianzas y covarianzas tanto del proceso de error $\varepsilon(\cdot)$ como de los residuos $\hat{\boldsymbol{\varepsilon}}$, y utiliza ambas

matrices para obtener las réplicas bootstrap del proceso espacial. A partir de este método es posible realizar inferencia sobre características del proceso espacial, utilizando la aproximación que se presenta en el siguiente apartado.

3.1.1. Aproximación bootstrap de la precisión y el sesgo de un estimador

Los métodos bootstrap tradicionales, diseñados originalmente para el caso de datos independientes, permiten aproximar propiedades de la distribución de un estimador, tales como su sesgo o varianza, midiendo estas características a través de réplicas obtenidas a partir de la distribución empírica de los datos observados (ver p.e. Efron y Tibshirani, 1994, Cap. 10). Con este mismo objeto, se han propuesto diversas técnicas de remuestreo para datos correlacionados, en especial para el contexto de series temporales. Algunas de estas técnicas están basadas en procedimientos paramétricos (p.e. suponiendo modelos autorregresivos), mientras que otros métodos suponen situaciones de dependencia más generales, como es el caso de los métodos de bootstrap por bloques. Una revisión de estas técnicas se pueden encontrar en Cao (1999).

Estos métodos bootstrap también se pueden utilizar para aproximar los momentos de la distribución de un estimador en el caso espacial. A modo de ejemplo, y por simplicidad, supondremos inicialmente que el vector $\mathbf{Y}$ representa n observaciones del proceso espacial (1.24), en cual la tendencia $\mu(\mathbf{x})$ es constante. Asimismo, consideraremos que el estimador de interés $\hat{\gamma}(u)$ proporciona una aproximación del variograma isotrópico teórico $\gamma(u)$ para un salto determinado u. En este caso, los métodos bootstrap estiman la distribución de $(\hat{\gamma}(\cdot) - \gamma(\cdot))$, y por tanto aproximan el sesgo y la varianza de $\hat{\gamma}(\cdot)$, mediante el siguiente procedimiento general:

1. Seleccionar un estadístico $\hat{\gamma}(u)$. En nuestro caso, se puede recurrir a estimadores piloto como el semivariograma empírico (1.13) o el estimador lineal local (1.16).

2. Utilizar un método bootstrap apropiado para generar B réplicas de los datos originales $\{Y_b^*(\mathbf{x}_1), \ldots, Y_b^*(\mathbf{x}_n)\}, b = 1, \ldots B$.

3. A partir de estas réplicas bootstrap, obtener B valores del semivariograma en el salto u, los cuales se denotan por $\{\hat{\gamma}_1^*(u), \ldots, \hat{\gamma}_B^*(u)\}$.

4. La versión bootstrap de $Var(\hat{\gamma}(u))$, se aproxima empíricamente mediante:

$$\widehat{Var}^*(\hat{\gamma}^*(u)) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\gamma}_b^*(u) - \bar{\hat{\gamma}}^*(u)\right)^2, \tag{3.1}$$

   donde $\bar{\hat{\gamma}}^*(u) = \sum_{b=1}^{B} \hat{\gamma}_b^*(u)/B$.

5. De forma similar, el análogo bootstrap del sesgo $Bias(\hat{\gamma}(u))$ puede ser numéricamente aproximado por:

$$\widehat{Bias}^*(\hat{\gamma}^*(u)) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\gamma}_b^*(u) - \hat{\gamma}(u)\right). \tag{3.2}$$

El procedimiento anterior permite la aproximación bootstrap del sesgo y la varianza del variograma estimado de un proceso estacionario de forma conjunta. Sin embargo, si se admite la presencia de una tendencia no constante, la estimación del semivariograma se realiza a partir de residuos, con el consabido problema del efecto del sesgo. En ese caso, la aproximación anterior necesita recurrir al uso de un estimador corregido $\tilde{\gamma}(u)$ del variograma, como se propone en la Sección 3.2.

Otro aspecto importante es la selección del método bootstrap en el paso 2.

Este método debe asegurar que las remuestras bootstrap realmente reproducen el comportamiento de los datos originales, tomando en cuenta la dependencia espacial. Al respecto, en Lahiri (2003) se extendieron los métodos bootstrap por bloques de datos de series de tiempo al caso multidimensional. Otros enfoques (García-Soidán *et al.*, 2014) sugieren generar réplicas del proceso a partir de un estimador no paramétrico del la distribución espacial conjunta. El método semiparamétrico, por otra parte, intenta obtener remuestras del proceso espacial bajo media no constante, extendiendo las ideas presentadas por Solow (1985) en el caso univariante. En las siguientes subsecciones revisaremos algunos detalles de los métodos más utilizados con datos bajo dependencia espacial.

3.1.2. Método Bootstrap por bloques

El método bootstrap por bloques BB fue introducido originalmente por Liu y Singh (1992) para el caso de datos unidimensionales regularmente espaciados (series de tiempo). Este método obtiene réplicas bootstrap de una serie temporal por medio de la concatenación de bloques de tamaño b de datos originales consecutivos. Una extensión de este procedimiento al caso multidimensional fue desarrollada en Lahiri (2003). El caso bidimensional, suponiendo bloques cuadrados por simplicidad, se expone a continuación.

Supongamos que el proceso estacionario $Y(\cdot)$ se encuentra definido sobre una rejilla regular $D = R \cap \mathbb{Z}^2 \subset \mathbb{R}^2$, donde R es una región bidimensional continua con área positiva centrada en el origen. El primer paso es dividir esta región R, en bloques cuadradados no solapados de la forma $b \times \mathcal{U}$, siendo $\mathcal{U} = (0, 1]^2$, y $b \in \mathbb{N}$. Esta partición genera un conjunto de índices $\mathcal{K} = \{\mathbf{k} \in \mathbb{Z}^2 : b(\mathbf{k} + \mathcal{U}) \subset R\}$ de todos los bloques completos separados de tamaño $b \times b$ contenidos en R, de modo que $n = |\mathcal{K}|b^2$. Luego, la muestra original $\mathbf{Y}$ puede ser reexpresada al concatenar

las submuestras $\{Y(\mathbf{x}_i) \in \mathbf{Y} : \mathbf{x}_i \in R(\mathbf{k})\}$, donde $R(\mathbf{k}) = R \cap [b(\mathbf{k} + \mathcal{U})]$, para $\mathbf{k} \in \mathcal{K}$.

Por otra parte, se define $\mathcal{I} = \{\mathbf{i} \in \mathbb{Z}^2 : \mathbf{i} + b\,\mathcal{U} \subset R\}$ como el conjunto de índices de todos los posibles bloques cuadrados con área b^2 dentro de R, con punto inicial en $\mathbf{i}$. Entonces, el conjunto $\mathcal{B} = \{\mathbf{i} + b\,\mathcal{U} : \mathbf{i} \in \mathcal{I}\}$ incluye todos los posibles bloques solapados dentro de la región R y con el mismo tamaño b. A partir de este punto, el método BB procede de forma similar al caso unidimensional, de manera que para cada índice $\mathbf{k} \in \mathcal{K}$, se obtiene un bloque bootstrap $R^*(\mathbf{k})$ asignándole un bloque del conjunto $\mathcal{B}$ mediante muestreo aleatorio sobre el conjunto de índices $\mathcal{I}$. Con estas asignaciones de bloques, la versión boostrap $\mathbf{Y}^*$ se obtiene mediante la unión de las submuestras correspondientes a cada bloque $R^*(\mathbf{k})$. Repitiendo este procedimiento B veces, se obtendrían B réplicas boostrap, a partir de las cuales se podrían aproximar las características de interés, como en la sección anterior.

Para asegurar la consistencia asintótica del método BB, Lahiri (2003) consideró las siguientes restricciones: El dominio espacial R debe satisfacer que $R = R_m = \lambda_m R_0 \subset\in \mathbb{R}^2$, donde $\{\lambda_m\}_{m\geq 1}$ es una secuencia de factores de escala que divergen a $+\infty$ y $R_0 \in \mathbb{R}^2$ es un subconjunto fijo que contiene una esfera bidimensional con área positiva y centrada en el origen; por ejemplo, $R_0 = (-1/2, 1/2]^2$. También se define $n = n_m$ tal que $n = A(R_0)\lambda_m^2$, donde $A(R_0)$ es el área de R_0. Además, se toma $b = b_m$ siendo $\{b_m\}_{m\geq 1}$ una sucesión de valores positivos verificando que $b_m^{-1} + b_m\lambda_m^{-1} = o(1)$ cuando m tiende a $+\infty$. Bajo estas condiciones, si se considera estadísticos que son funciones lineales o suaves de la media muestral, es posible obtener estimaciones consistentes de la varianza de dicho estadístico. Sin embargo, cuando se trata de la estimación del variograma, esta linealidad no se satisface (Clark y Allingham, 2011).

Asimismo, el tamaño óptimo del bloque b fue establecido por Hall *et al.* (1995) en el caso unidimensional, demostrando que éste es del orden $n^{1/3}$ para la esti-

mación bootstrap del sesgo y la varianza. Se obtuvieron resultados similares en Nordman *et al.* (2007) para el caso multidimensional. No obstante, debe considerarse que no siempre es posible obtener una partición exacta del dominio espacial para un determinado tamaño de bloque b, especialmente si D es una región irregular.

Otra desventaja del método BB es que no es estacionario. Esto ya fue observado en el caso de datos temporales (ver p.e. Goncalves y Politis, 2011). Para el caso unidimensional, se propone un método bootstrap estacionario en Politis y Romano (1994). Este procedimiento se basa en la selección aleatoria del tamaño del bloque, a partir de una distribución geométrica con parámetro p que representa el tamaño promedio de los bloques. Sin embargo, la extensión del bootstrap estacionario al caso multidimensional presenta varias dificultades y no se encontraron referencias de ninguna generalización a este contexto.

El método boostrap por bloques puede ser adaptado a procesos con tendencia, estimando previamente la tendencia mediante un modelo paramétrico, y luego utilizando los correspondientes residuos en lugar de la muestra original para obtener las réplicas bootstrap. Sin embargo, el efecto del sesgo puede afectar a la precisión de los resultados obtenidos de esta manera.

3.1.3. Método bootstrap semiparamétrico

Este método sigue las ideas propuestas por Solow (1985) para el caso unidimensional. Fue diseñado inicialmente para procesos estacionarios de segundo orden, y se basa en los métodos de simulación espacial (ver p.e. Cressie, 1993, Section 3.6.1). Este procedimiento realiza muestreo independiente de datos incorrelados, eliminando la estructura de dependencia a partir de un estimador paramétrico de la matriz de varianzas y covarianzas. Posteriormente, este método

semiparamétrico SPB fue aplicado a datos espaciales por Olea y Pardo-Igúzquiza (2011) y por Iranpanah *et al.* (2011), en procesos estacionarios y procesos con tendencia no constante, respectivamente.

Supongamos por simplicidad, que el proceso $Y(\cdot)$ es estacionario. En este caso, el método SPB consiste en los siguiente pasos:

Algoritmo 3.1: Método Bootstrap Semiparamétrico SPB

1. Obtener la matriz de varianzas y covarianzas estimada $\hat{\mathbf{\Sigma}}$ a partir de un modelo paramétrico válido ajustado a los datos $\mathbf{Y}$ (p.e. utilizando el análisis estructural descrito en la Sección 1.4);
2. Calcular la correspondiente matriz triangular $\mathbf{L}$ tal que $\hat{\mathbf{\Sigma}} = \mathbf{L}\mathbf{L}^t$, mediante la descomposición de Cholesky;
3. Obtener las variables incorreladas $\mathbf{e} = (e_1, e_2, \ldots, e_n)^t$, donde $\mathbf{e} = \mathbf{L}^{-1}\mathbf{y}$;
4. Centrar las variables anteriores, y generar los muestras bootstrap independientes de tamaño n a partir de ellas, denotadas por $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)^t$;
5. Finalmente, obtener las muestras bootstrap de los datos dependientes, dados por $\mathbf{Y}^* = \mathbf{L}\mathbf{e}^*$;

Clark y Allingham (2011); Pardo-Igúzquiza y Olea (2012), entre otros autores, han aplicado el método SPB para aproximar mediante bootstrap el sesgo y varianza del estimador empírico del variograma (1.13), demostrando su buen comportamiento en datos espaciales estacionarios.

La principal ventaja de esta técnica es que no depende de la configuración espacial de la región de observación D, permitiendo su implementación en regiones más generales, a diferencia del método BB. Además, Iranpanah *et al.* (2011) verificaron que para el caso de procesos estacionario, el método semiparamétrico reproduce mejor la variabilidad de estimadores basados en la media de procesos estacionarios, en comparación con el método por bloques. La precisión de este último método depende del tamaño del boque considerado, y por lo general

subestima la variabilidad de cada estimador analizado, especialmente cuando la dependencia espacial es fuerte.

Sin embargo, en el caso de la presencia de una tendencia, el método semiparamétrico no considera efecto del sesgo debido al uso de residuos y la precisión de los resultados dependen de la apropiada selección de los modelos paramétricos.

3.2. Método bootstrap no paramétrico

En esta sección se propone un procedimiento bootstrap no paramétrico NPB, para el caso de procesos espaciales con tendencia no constante, es decir, cuando el proceso espacial $Y(\cdot)$ sigue el modelo (1.24). El método NPB utiliza el estimador lineal local de la tendencia (2.1) y el correspondiente estimador del variograma calculado a partir de los residuos (2.13), así como sus estimaciones piloto corregidas mediante el procedimiento de corrección NP del variograma 2.1. Asimismo, utiliza los modelos flexibles de Shapiro-Botha, para aproximar las matrices de varianzas y covarianzas de los residuos $\mathbf{\Sigma}_{\hat{\varepsilon}}$ y de los datos originales $\mathbf{\Sigma}$, las cuales luego son utilizadas para generar las réplicas bootstrap. El procedimiento NPB se puede resumir en el Algoritmo 3.2.

Es importante recalcar que el método NBP no depende de la configuración espacial de la región de observación D, por lo que puede ser implementado en regiones más generales, a diferencia del método BB. Sin embargo, la principal ventaja de este procedimiento es que proporciona réplicas bootstrap para datos con tendencia no constante, tratando de reproducir correctamente la variabilidad del proceso espacial. Para este fin, el método propuesto utiliza estimaciones de la matriz de varianzas y covarianzas residual y corregida a la hora de generar dichas réplicas. Además, el método NPB no está afectado por los problemas de mala especificación inherentes a la selección de modelos de las técnicas paramétricas

Algoritmo 3.2: Método Bootstrap No Paramétrico NPB

1 Seleccionar una matriz ventana $\mathbf{H}$ para estimar la tendencia (por ejemplo, mediante el criterio $CGCV$) y obtener $\hat{\mu}_{\mathbf{H}}(\cdot)$ en base a (2.1);

2 Calcular los residuos $\hat{\varepsilon}_i = Y(\mathbf{x}_i) - \hat{\mu}_{\mathbf{H}}(\mathbf{x}_i)$, para $i = 1, \ldots, n$. Luego, aplicar el procedimiento de corrección de sesgo 2.1 para obtener el estimador lineal local del variograma residual $\hat{\gamma}(\cdot)$ y su versión corregida $\tilde{\gamma}(\cdot)$;

3 Ajustar modelos de Shapiro Botha a los estimadores $\hat{\gamma}(\cdot)$ y $\tilde{\gamma}(\cdot)$, y a partir de estos variogramas válidos construir las matrices de varianzas y covarianzas estimadas $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$ y $\hat{\mathbf{\Sigma}}$ respectivas;

4 Descomponer las matrices anteriores mediante la factorización de Cholesky, de manera que $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}} = \mathbf{L}_{\hat{\varepsilon}}\mathbf{L}_{\hat{\varepsilon}}^t$ y $\hat{\mathbf{\Sigma}} = \mathbf{L}\mathbf{L}^t$;

5 Calcular los residuos "independentes" $\mathbf{e} = \mathbf{L}_{\hat{\varepsilon}}^{-1}\hat{\varepsilon}$, donde $\mathbf{e} = (e_1, \ldots, e_n)^t$;

6 Centrar las variables anteriores, y a partir de estas, obtener muestras bootstrap independientes de tamaño n, denotados por $\mathbf{e}^* = (e_1^*, \ldots, e_n^*)^t$;

7 Calcular los errores bootstrap $\varepsilon^* = \mathbf{L}\mathbf{e}^*$, donde $\varepsilon^* = (\varepsilon_1^*, \ldots, \varepsilon_n^*)^t$, y finalmente, generar las remuestras bootstrap del proceso espacial, haciendo $Y^*(\mathbf{x}_i) = \hat{\mu}_{\mathbf{H}}(\mathbf{x}_i) + \varepsilon_i^*,\ i = 1, 2, \ldots, n$.

o semiparamétricas, ni en la caracterización de la tendencia ni de las matrices de varianzas y covarianzas. Incluso, el método no paramétrico se puede adaptar fácilmente al caso estacionario, en los cuales no se requiere estimar $\mu(\cdot)$ ni $\mathbf{\Sigma}_{\hat{\varepsilon}}$.

Si bien el NPB se puede utilizar para realizar inferencias sobre distintas características del proceso espacial, en el presente estudio nos centraremos en determinar si este método reproduce de manera adecuada la varianza de distintos estimadores del variograma, mediante las aproximaciones bootstrap explicadas en la Sección 3.1.1. Para esto, se puede estimar la tendencia en cada réplica bootstrap $\{Y^*(\mathbf{x}_1), ..., Y^*(\mathbf{x}_n)\}$, y a partir de los residuos bootstrap se puede evaluar el correspondiente estimador $\hat{\gamma}^*(\cdot)$. Procediendo de esta manera de forma iterativa, es posible aproximar el sesgo y varianza bootstrap mediante las expresiones (3.1) y (3.2), considerando el modelo de Shapiro-Botha ajustado a $\tilde{\gamma}(u)$ en lugar de

$\hat{\gamma}(u)$ en la aproximación del sesgo.

El método NPB también es válido para realizar otro tipo de inferencias, por ejemplo, para obtener intervalos de confianza puntual, bandas de confianza, contrastes de hipótesis, etc. Un ejemplo ilustrativo acerca de la aplicabilidad del NPB para el primer caso, se presenta en la Sección 3.5.

Finalmente, el método NPB puede ser programado e implementado de forma automática, sin que se requiera la interacción del usuario para ajustar las estimaciones, lo que constituye otra ventaja de orden práctico frente a los otros métodos disponibles.

3.3. Estudios de simulación

3.3.1. Resultados en procesos estacionarios

En esta sección se describen los resultados obtenidos al aplicar los distintos métodos bootstrap (BB, SPB y NPB) para aproximar la variabilidad del estimador empírico (1.13) y lineal local (1.16) del variograma en el caso de procesos estacionarios. Para esto, se consideró un proceso gaussiano definido sobre la región $\mathcal{D} = [0,1]^2$, con tendencia nula, y función teórica del variograma $\gamma(\cdot)$ dada por el modelo isotrópico de Matérn (1.19). Inicialmente se fijó el parámetro de suavizado $v = 0{,}5$, con lo cual el modelo de Matérn coincide con el modelo exponencial (1.17). Asimismo, se consideraron los siguientes parámetros: $\sigma^2 = 1$ ($c_1 = \sigma^2 - c_0$), $a = 0{,}3$, $0{,}6$ y $0{,}9$, y tomando efectos nugget equivalentes al $0\,\%$, $20\,\%$ y $50\,\%$ de la varianza total.

Para cada combinación de los parámetros antes mencionados, se generaron $N = 1000$ muestras de tamaños $n = 12 \times 12$ y 24×24 sobre una rejilla regular bidimensional contenida en la región $\mathcal{D}$. Cabe indicar que estos tamaños muestrales

fueron seleccionados de manera que permitan una partición exacta de la región D para el método BB. Además, procediendo de este modo, se pueden comparar los resultados obtenidos en el presente estudio con aquellos presentados en el trabajo de Iranpanah *et al.* (2011).

Utilizando cada uno de los métodos bootstrap, se generaron $B = 1000$ réplicas y se calculó para cada muestra bootstrap los estimadores del variograma considerados. A partir de estos estadísticos bootstrap y mediante el procedimiento descrito en la Sección 3.1.1, se obtuvieron las respectivas aproximaciones del sesgo y de la varianza. En nuestro caso nos centraremos en analizar los resultados obtenidos por los distintos métodos bootstrap a la hora de aproximar la variabilidad de $\hat{\gamma}(u_i)$. Con este fin se recurrió a la relación (3.1) para estimar $Var^*(\hat{\gamma}(u_i))$ en los saltos considerados $u_i = l * i$, con $i = 1, \ldots, q$, siendo l la distancia mínima entre las localizaciones en D, $q = \lfloor 0{,}55\sqrt{2}/l \rfloor$ donde $\lfloor r \rfloor$ denota la parte entera de r.

Para comparar la precisión obtenida mediante las distintas técnicas bootstrap, se utilizaron las siguientes medidas de error, aproximadas numéricamente por simulación:

$$AE(u) = \mathbb{E}\left[|Var^*(\hat{\gamma}^*(u)) - Var(\hat{\gamma}(u))|\right];$$
$$SE(u) = \mathbb{E}\left[(Var^*(\hat{\gamma}^*(u)) - Var(\hat{\gamma}(u)))^2\right];$$
$$RSE(u) = \mathbb{E}\left[(Var^*(\hat{\gamma}^*(u))/Var(\hat{\gamma}(u)) - 1)^2\right]$$

donde AE, SE y RSE corresponden las medias del error absoluto, error cuadrático y error cuadrático relativo, respectivamente.

Respecto al método BB, fue necesario determinar en primer lugar el tamaño de bloque óptimo para ambos estimadores considerados. Para esto, se tomaron distintos tamaños de bloque que permiten particiones exactas, es decir, $b = \sqrt{n}/2$, $\sqrt{n}/3$, $\sqrt{n}/4$, $\sqrt{n}/6$. Luego, el tamaño óptimo b_{opt} se seleccionó como aquel valor

b que minimice el promedio de los valores RSE:

$$ARSE = \frac{1}{q}\sum_{i=1}^{q} RSE(u_i).$$

El ajuste paramétrico del método SPB se realizó utilizando el paquete `geoR` de `R`, tomando como semivariograma teórico el modelo exponencial (1.17), evitando de esta manera el efecto de una mala especificación del modelo paramétrico teórico (fijando $v = 0{,}5$). Por otra parte, la ventana óptima g_{opt} del estimador lineal local (1.16) se obtuvo minimizando el criterio de error RSE. En la Tabla 3.1 se presentan los resultados para el caso de $n = 24 \times 24$, considerando el estimador lineal local del variograma y el método BB.

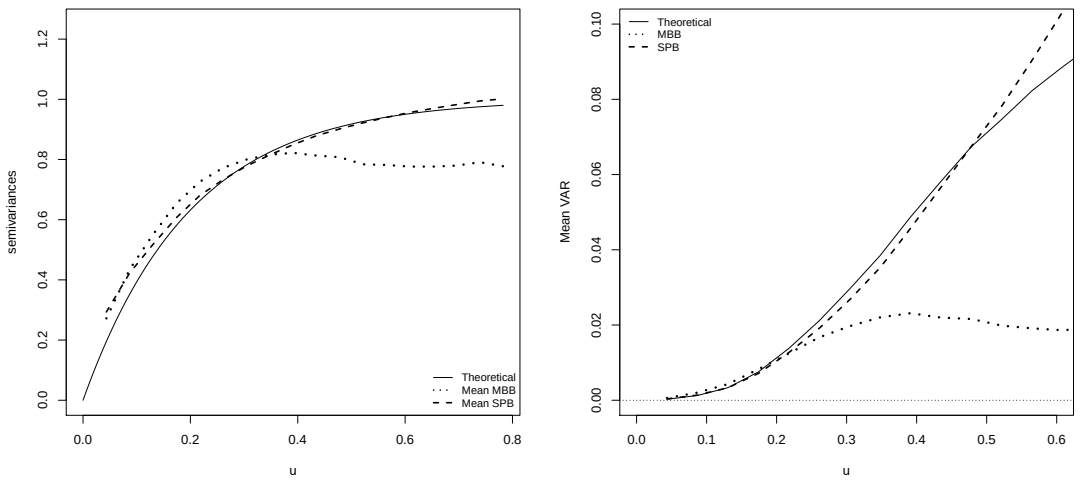
Tabla 3.1: Valores de g_{opt} para el estimador lineal local, $ARSE$ y b_{opt} para el método BB, para $n = 24 \times 24$.

			$ARSE$				
c_0	a	g_{opt}	$b = 4$	$b = 6$	$b = 8$	$b = 12$	b_{opt}
0 %	0.3	0.093	0.600	0.310	0.230	0.260	8
	0.6	0.138	4.770	1.950	0.980	0.430	12
	0.9	0.184	15.700	6.030	2.820	0.950	12
20 %	0.3	0.140	0.330	0.210	0.190	0.260	8
	0.6	0.250	1.390	0.650	0.410	0.330	12
	0.9	0.300	2.870	1.190	0.660	0.390	12
50 %	0.3	0.184	0.157	0.126	0.138	0.252	6
	0.6	0.342	0.420	0.290	0.250	0.320	8
	0.9	0.387	0.560	0.370	0.300	0.340	8

En estos resultados se puede apreciar el efecto de la estructura de dependencia sobre la elección del tamaño óptimo del bloque. Por ejemplo, si el rango a aumenta de 0,3 a 0,9 (y por tanto la dependencia espacial es más fuerte), el valor b_{opt} también aumenta, pasando de $\sqrt{n}/3$ a $\sqrt{n}/2$. Si se varía el efecto nugget se observan patrones similares (a menores valores de c_0 se requieren tamaños óptimos mayores). En resumen, el método BB requiere valores de b_{opt} más grandes

conforme la correlación espacial se vuelve más fuerte. Estas conclusiones se mantienen para los otros parámetros de simulación considerados, y se verifica además que conforme n aumenta, también lo hacen los tamaños óptimos de bloque. Otro aspecto relevante de esta tabla se refiere a la ventana óptima g_{opt}, donde se evidencia que a medida que el rango aumenta, las ventanas tienden a ser más grandes. Estos resultados son similares a los obtenidos cuando se considera el estimador empírico.

Seguidamente se analizaron las medias de los dos estimadores del variograma y se compararon con el modelo teórico de variograma considerado. Posteriormente, se analizó el comportamiento de los métodos bootstrap a la hora de reproducir la variabilidad de cada estimador del variograma. Con este fin, se calcularon las medias de las aproximaciones numéricas de $Var^*(\hat{\gamma}^*(u))$, y se las comparó con las correspondientes $Var(\hat{\gamma}(u))$ aproximadas por simulación.



(a) (b)

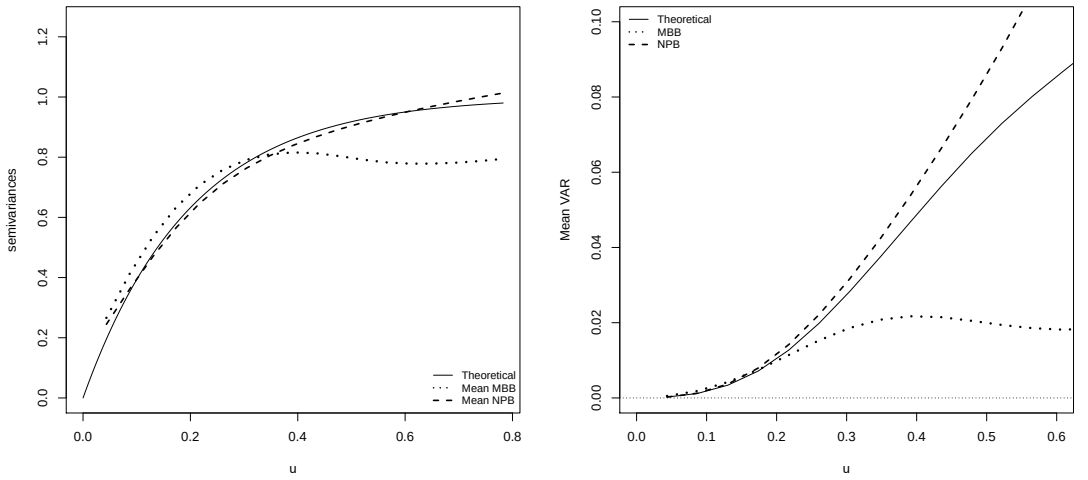
Figura 3.1: (a) Comparación del variograma teórico (línea continua) con la aproximación bootstrap de la media del estimador empírico del variograma, y (b) aproximaciones obtenidas por simulación para $Var(\hat{\gamma}(u))$ (línea continua) y para $Var^*(\hat{\gamma}^*(u))$ con los métodos BB y SPB (líneas de puntos y línea discontinua, respectivamente), con $n = 24 \times 24$, $\sigma^2 = 1$, $c_0 = 0\,\%$ y $a = 0{,}6$.

En la Figura 3.1(a) se muestran los resultados obtenidos para la aproximación

bootstrap de la media del estimador empírico del variograma con los métodos BB y SPB. Cabe mencionar que el comportamiento del método NPB fue muy similar al presentado por el método SPB, y por tal razón sus resultados no fueron incluidos en este gráfico. Esto permite además comparar el BB con el método SPB considerando su enfoque de estimación tradicional (es decir, basado en el estimador empírico). En este gráfico se observa que el método SPB proporciona mejores aproximaciones de la media bootstrap del estimador respecto al modelo teórico, mientras que el método por bloques presenta aproximaciones bootstrap que parecen sobreestimar el variograma teórico en saltos pequeños, mientras que a medida que el salto es más grande, su media bootstrap subestima a $\gamma(\cdot)$. Por otra parte, la Figura 3.1(b) presenta las aproximaciones por simulación de la varianza del estimador y de su respectiva media bootstrap. Se puede observar que el método BB nuevamente tiende a subestimar la variabilidad cuando la distancia entre las observaciones se incrementa. Por su parte, el método SPB presenta mejores aproximaciones de la varianza del estimador empírico.

Por otra parte, al analizar las mismas aproximaciones bootstrap utilizando el estimador lineal local, los métodos BB y NPB presentaron un comportamiento parecido a los obtenidos con el estimador empírico, respecto a las medias de $\hat{\gamma}^*(u)$, tal como se puede verificar en las Figuras 3.2(a) y 3.2(b). Asimismo, los resultados del método SPB son similares a los obtenidos por NPB para el caso de procesos estacionarios, y sus resultados no se incluyeron para permitir la comparación entre el BB y el enfoque tradicional de estimación del método no paramétrico.

Los gráficos anteriores parecen sugerir que los métodos SPB y NPB proporcionan mejores aproximaciones del sesgo y varianza teóricos, en comparación con el método BB, independientemente del estimador del variograma considerado. Estos efectos sobre el método BB se deben posiblemente a la concatenación de los bloques, en los cuales los datos en las réplicas bootstrap se colocan a distancias



(a) (b)

Figura 3.2: (a) Comparación del variograma teórico (línea continua) con la aproximación bootstrap de la media del estimador lineal local del variograma, y (b) aproximaciones obtenidas por simulación para $Var(\hat{\gamma}(u))$ (línea continua) y para $Var^*(\hat{\gamma}^*(u))$ con los métodos BB y NPB (líneas de puntos y línea discontinua, respectivamente), con $n = 24 \times 24$, $\sigma^2 = 1$, $c_0 = 0\,\%$ y $a = 0{,}6$.

más cercanas o más lejanas de lo que realmente se encuentran en la muestra original, lo cual parece producir una ligera sobrestimación del variograma en saltos pequeños y una más evidente subestimación en saltos más grandes. Esto último también se presenta en el caso de la aproximación de la varianza en ambos estimadores.

Para comparar numéricamente el comportamiento de los métodos bootstrap, se calcularon las medidas de error definidas anteriormente, cuyos valores se resumen en las Tablas 3.2 y 3.3, para el caso del estimador empírico y lineal local respectivamente. Estas tablas conducen a conclusiones similares a las obtenidas en los gráficos anteriores, dando cuenta de la superioridad de los métodos SPB y NPB respecto al BB. Como cabría esperar, los valores obtenidos para los errores AE y SE se incrementan conforme la dependencia espacial es más fuerte (medida en términos de las variaciones del efecto nugget o del rango).

Resultados similares se obtuvieron con las diferentes configuraciones de simu-

Tabla 3.2: Promedios de AE, SE y RSE ($\times 10^{-2}$) de $\widehat{Var}^*(\hat{\gamma}^*(u))$ obtenidos con los métodos BB y SPB, considerando el estimador empírico del variograma, con $n = 24 \times 24$ y $\sigma^2 = 1$.

		AAE		ASE		ARSE	
c_0	a	BB	SPB	BB	SPB	BB	SPB
0 %	0.3	1.188	0.392	0.020	0.003	23.493	3.585
	0.6	3.697	0.948	0.248	0.027	43.547	3.701
	0.9	4.868	0.601	0.489	0.012	93.919	2.377
20 %	0.3	0.780	0.250	0.009	0.002	18.730	2.970
	0.6	2.400	0.900	0.103	0.023	31.800	5.300
	0.9	3.130	1.070	0.201	0.027	37.100	4.800
50 %	0.3	0.329	0.110	0.002	0.000	12.750	2.206
	0.6	0.921	0.456	0.015	0.005	25.159	6.862
	0.9	1.208	0.683	0.030	0.011	29.654	8.860

Tabla 3.3: Promedios de AE, SE y RSE ($\times 10^{-2}$) de $\widehat{Var}^*(\hat{\gamma}^*(u))$ obtenidos con los métodos BB y NPB, considerando el estimador lineal local del variograma, con $n = 24 \times 24$ y $\sigma^2 = 1$.

		AAE		ASE		ARSE	
c_0	a	BB	NPB	BB	NPB	BB	NPB
0 %	0.3	1.160	0.600	0.019	0.010	23.050	8.890
	0.6	3.620	2.200	0.239	0.121	43.200	12.300
	0.9	4.800	3.200	0.472	0.242	94.500	14.600
20 %	0.3	0.750	0.420	0.008	0.005	18.600	9.400
	0.6	2.310	1.430	0.096	0.053	33.000	12.800
	0.9	3.100	2.000	0.191	0.104	38.800	14.400
50 %	0.3	0.313	0.191	0.001	0.001	12.600	7.470
	0.6	0.870	0.570	0.014	0.009	25.300	11.600
	0.9	1.170	0.810	0.028	0.017	30.400	13.400

lación con ambos estimadores, en los cuales se observó en general que el método SPB y NPB presentan mejores aproximaciones bootstrap respecto al BB. Sin embargo se debe considerar que en estos estudios numéricos, la matriz Σ fue estimada paramétricamente considerando un modelo exponencial, es decir, evitando problemas de mala especificación.

Para analizar este efecto, se llevo a cabo un estudio final de simulación, mo-

dificando únicamente el paramétro $v = 0{,}25$ del modelo de variograma teórico (1.19) a la hora de simular los datos espaciales. Este parámetro afecta a la forma del variograma en los saltos cercanos al origen, lo cual tiene una gran relevancia en términos de inferencia y predicción espacial (ver p.e. Sección 1.3.4). El resto de parámetros se mantuvieron sin cambio alguno, mientras que para el ajuste paramétrico del método SPB se consideró el modelo exponencial (es decir, tomando $v = 0{,}50$). Se realizaron nuevamente los análisis anteriores, y se calcularon las medidas de error con el fin de comparar la variabilidad obtenida para los estimadores respectivos con los distintos métodos bootstrap, obteniéndose resultados similares a los que se muestran, a modo de ejemplo, en la Tabla 3.4.

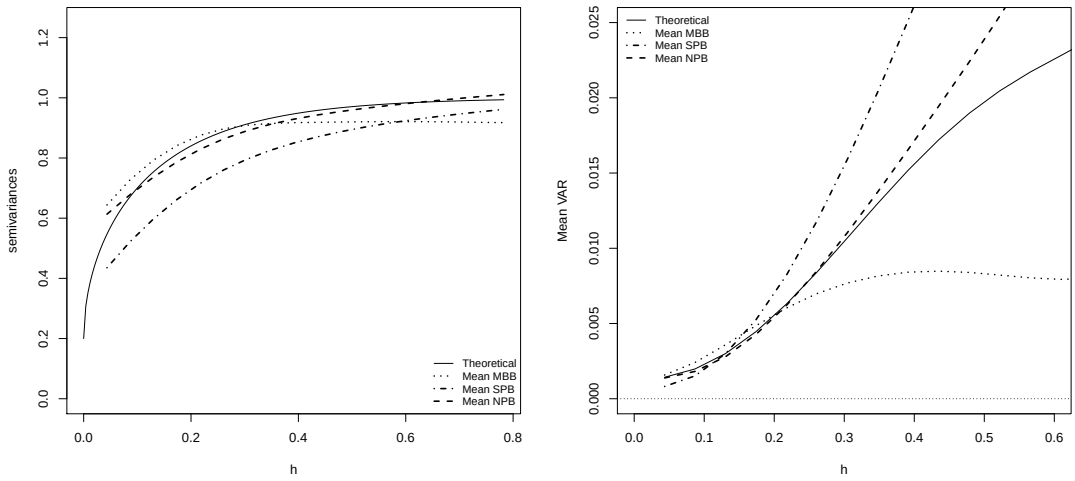
Tabla 3.4: Promedios de AE, SE y RSE ($\times 10^{-2}$) de $\widehat{Var}^*(\hat{\gamma}^*(u))$ con los distintos métodos bootstrap, tomando $v = 0{,}25$ en el modelo de variograma teórico y fijando $v = 0{,}50$ en el ajuste paramétrico, con $n = 24 \times 24$, $\sigma^2 = 1$, $c_0 = 20\,\%$ y $a = 0{,}6$.

Estimador	Método	AAE	ASE	ARSE
Empírico	BB	0.860	0.012	23.000
	SPB	1.700	0.055	89.700
	NPB	0.540	0.007	11.000
Lineal Local	BB	0.822	0.011	22.624
	SPB	1.661	0.054	92.609
	NPB	0.546	0.008	11.590

En esta última tabla se puede observar que la aproximación de la varianza del estimador con el el método SPB conduce a errores mucho más grandes, sobrepasando incluso los obtenidos por el método BB. Por otra parte, se verifica que el método no paramétrico propuesto presenta las mejores aproximaciones bootstrap de la variabilidad del estimador del variograma. Conclusiones similares se obtienen al analizar las Figuras 3.3(a) y 3.3(b), en las cuales se observa que las estimaciones bootstrap obtenidas por método SPB presentan mayores errores cuando el modelo de variograma está mal especificado. En este último caso, las medias bootstrap subestiman al variograma teórico, mientras que las varianzas

bootstrap sobreestiman considerablemente la varianza del estimador lineal local del variograma.

Este último estudio pone en evidencia la vulnerabilidad del método SPB respecto a la mala especificación de los modelos paramétricos seleccionados para estimar la matriz de varianzas y covarianzas, en procesos estacionarios. Este problema puede tener mayor relevancia en el caso de tendencia espacial no constante, pues en este caso, el método SPB requiere ajustar un modelo paramétrico adicional, además del efecto del sesgo debido a los residuos. Esta situación fue analizada mediante los estudios numéricos que se presentan en la siguiente sección.



(a) (b)

Figura 3.3: (a) Comparación del variograma teórico (línea continua) con la aproximación bootstrap de la media del estimador lineal local del variograma, y (b) aproximaciones obtenidas por simulación para $Var(\hat{\gamma}(u))$ (línea continua) y para $Var^*(\hat{\gamma}^*(u))$ con los métodos BB (línea de puntos), SPB (línea discontinua con puntos) y NPB (línea discontinua), coniderando mala especificación del modelo de variograma en el SPB, con $n = 24 \times 24$, $\sigma^2 = 1$, $c_0 = 0\,\%$ y $a = 0{,}6$.

3.3.2. Resultados en procesos con tendencia no constante

En este apartado se consideró un proceso espacial bajo tendencia no constante, es decir, se supone que $Y(\mathbf{x})$ sigue el modelo no estacionario (1.24), con

una función tendencia $\mu(\cdot)$ determinística no constante. Bajo las mismas configuraciones de simulación utilizadas en el caso estacionario se desarrollaron varios estudios numéricos para comparar los métodos bootstrap a la hora de aproximar la variabilidad del estimador empírico y lineal local del variograma, esta vez calculado a partir de residuos. Por tanto, los datos se generaron en una rejilla regular $n \times n$ definida sobre la región D, con los mismos parámetros de simulación para N, n y B. Además, se consideraron dos tipos de tendencias: $\mu_1(x_1, x_2) = c_1 * [sin(2\pi x_1) + 4(x_2 - 0{,}5)^2]$ (modelo no polinómica), y $\mu_2(x_1, x_2) = c_2 * [x_1 - x_2 + x_2^2]$ (modelo polinómico). Las constantes c_1 y c_2 fueron seleccionadas siguiendo la propuesta de Fan y Gijbels (1996, pp. 111), donde la razón $\tau =$ ruido/señal, representa una medida de la dificultad para caracterizar la tendencia, tal que:

$$\tau = \sigma^2 / Var(\mu(X_1, X_2))$$

Para establecer estas constantes, se consideró que las localizaciones espaciales siguen una distribución uniforme en $[0, 1]^2$ y que la estructura de correlación corresponde al modelo de variograma exponencial (1.17) con $\sigma^2 = 1$. Entonces, c_1 y c_2 se obtuvieron de forma que la $Var(\mu(X_1, X_2))$ fuese igual a 2 o 3, de manera que la razón τ tomase valores de $1/2$ o $1/3$, respectivamente.

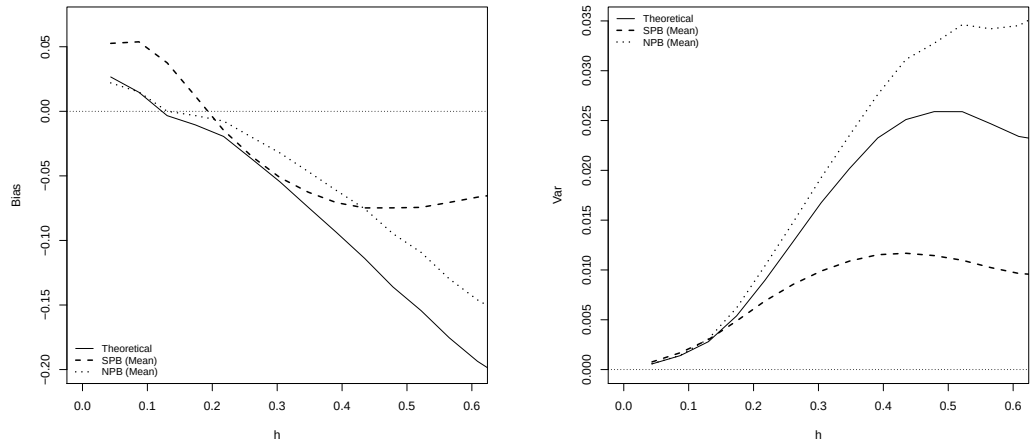
En este caso, y teniendo en cuenta la subestimación producida por el método BB en el caso estacionario, solamente se estudiaron los resultados para los métodos SPB y NPB. En el caso del método semiparamétrico, la tendencia espacial se caracterizó mediante el modelo paramétrico $\mu_\beta(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_2^2$, mientras que la matriz $\mathbf{\Sigma}$ fue estimada mediante ajuste paramétrico a partir del estimador empírico de los correspondientes residuos, tomando como semivariograma teórico el modelo exponencial (para evitar problemas de mala especificación del modelo de variograma). Cabe mencionar que la elección del modelo

paramétrico para la tendencia responde al hecho de que, en efecto, la función $\mu_2(x_1, x_2)$ pertenece a la familia paramétrica considerada (por tanto, el modelo estaría correctamente especificado), y a su vez representa a la mejor tendencia polinómica que se puede ajustar a la función $\mu_1(x_1, x_2)$ (ver p.e. Fernández-Casal y Francisco-Fernández (2014)).

Respecto al método NPB, para reducir el efecto de la selección de la ventana en la estimación no paramétrica de la tendencia, la ventana $\mathbf{H}$ del estimador lineal local de la tendencia se eligió mediante el criterio $MASE$ definido en (2.6). Luego, para la aproximación bootstrap en el método NPB se reemplazó la estimación del variograma $\hat{\gamma}(u)$ en la aproximación boostrap del sesgo (3.2) por el correspondiente modelo de Shapiro-Botha ajustado al estimador corregido $\tilde{\gamma}(u)$, y de esa manera evitar el efecto del sesgo debido al uso de residuos.

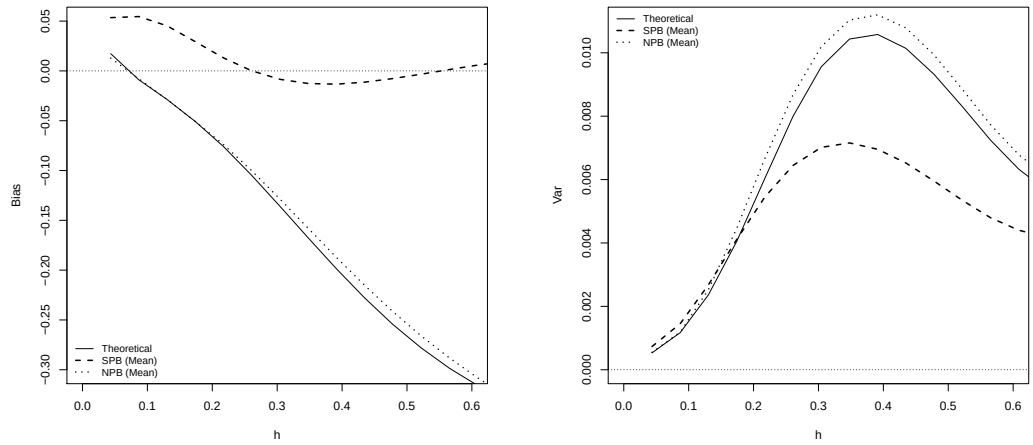
Una vez obtenidas las réplicas bootstrap con cada método, se calcularon los estimadores del variograma a partir de los residuos bootstrap. Para el caso del estimador empírico, la tendencia se estimó paramétricamente a partir del modelo $\mu_\beta(x_1, x_2)$ y para el caso del estimador lineal local del variograma, estos residuos se obtuvieron utilizando el estimador no paramétrico de la tendencia 2.1. Luego, se procedió a construir las aproximaciones bootstrap del sesgo y la varianza de cada estimador y se procedió a calcular las medidas de error descritas anteriormente.

De manera general, se verifica que las aproximaciones bootstrap para el sesgo y la varianza del estimador del variograma obtenido mediante el método NPB presenta mejores resultados en comparación con los valores presentados por la técnica SPB. Por ejemplo, las Figuras 3.4(a) y 3.4(b) muestran las aproximaciones por simulación de los sesgos y varianzas teóricos respectivamente, obtenidas para el estimador empírico (líneas continuas). Además, en cada caso se presentan las aproximaciones bootstrap del sesgo y varianza de dicho estimador utilizando el método SPB (línea discontinua) y el NPB (línea de puntos). Estas figuras po-



(a) (b)

Figura 3.4: Aproximaciones por simulación para (a) $Bias(\hat{\gamma}(u))$ y (b) $Var(\hat{\gamma}(u))$ (línea continua) considerando el estimador empírico, con sus aproximaciones bootstrap $Bias^*(\hat{\gamma}^*(u))$ y $Var^*(\hat{\gamma}^*(u))$ respectivas, obtenidas mediante el método SPB (línea discontinua) y con el método NPB (línea de puntos), para $\tau = 1/2$, $\mu_2(\cdot)$, $n = 24 \times 24$, $\sigma^2 = 1$, $c_0 = 20\,\%$ y $a = 0{,}6$.



(a) (b)

Figura 3.5: Aproximaciones por simulación para (a) $Bias(\hat{\gamma}(u))$ y (b) $Var(\hat{\gamma}(u))$ (línea continua) considerando el estimador lineal local, con sus aproximaciones bootstrap $Bias^*(\hat{\gamma}^*(u))$ y $Var^*(\hat{\gamma}^*(u))$ respectivas, obtenidas mediante el método SPB (línea discontinua) y con el método NPB (línea de puntos), para $\tau = 1/2$, $\mu_2(\cdot)$, $n = 24 \times 24$, $\sigma^2 = 1$, $c_0 = 20\,\%$ y $a = 0{,}6$.

nen en evidencia un mejor comportamiento del método no paramétrico respecto al SPB para el caso del estimador empírico.

En las Figuras 3.5(a) y 3.5(b) se presenta las mismas aproximaciones anteriores, obtenidas a partir del estimador lineal local del variograma. Estos gráficos evidencian la superioridad del método NPB a la hora de reproducir el sesgo como la varianza de ambos estimadores del variograma. Por su parte, el método SPB parece subestimar la variabilidad de dichos estimadores, y este efecto parece aumentar a medida que los saltos se alejan del origen.

Por otra parte, en la Tabla 3.5 se verifica que las medidas de error obtenidas por el método NPB son mejores que las correspondientes al método SPB en todos los casos. Además, cuando la razón τ decrece, la dificultad para estimar la tendencia aumenta. También se verifica que en esta situación los errores del método SPB no varían, mientras que los errores del método no paramétrico se reducen, evidenciando la robustez de los resultados obtenidos mediante el método NPB.

Tabla 3.5: Estadísticos de AE y SE ($\times 10^{-2}$) obtenidos para $\widehat{Var}^{*}(\hat{\gamma}^{*}(u))$, con $\tau = 1/2, 1/3$, $\mu_2(\cdot)$, $n = 24 \times 24$, $c_0 = 20\,\%$ y $a = 0{,}6$.

			$\tau = 1/2$		$\tau = 1/3$	
Estimador	Error	Método	Media	Desv. Est.	Media	Desv. Est
Empírico	AE	SPB	0.840	0.570	0.840	0.570
		NPB	0.600	0.500	0.590	0.490
	SE	SPB	0.010	0.009	0.010	0.009
		NPB	0.006	0.007	0.006	0.007
	RSE	SPB	21.30	12.50	21.30	12.50
		NPB	13.60	16.80	13.10	16.10
Lineal Local	AE	SPB	0.172	0.127	0.156	0.114
		NPB	0.046	0.020	0.032	0.013
	SE	SPB	0.000	0.000	0.000	0.000
		NPB	0.000	0.000	0.000	0.000
	RSE	SPB	7.280	4.670	7.040	4.580
		NPB	0.540	0.280	0.350	0.230

Cabe mencionar que en todos los resultados anteriores se consideró la tenden-

cia teórica lineal (es decir, cuando se utiliza el modelo paramétrico correcto para estimar la tendencia en el método SPB). Por tanto, este comportamiento de los errores en el método semiparamétrico se deben principalmente al sesgo debido al uso de los residuos sin corrección a la hora de estimar la variabilidad de pequeña escala. Es de esperar que los errores del método SPB se incrementen si se estima paramétricamente la tendencia teórica no polinómica $\mu_1(x, y)$ mediante $\mu_\beta(x, y)$. Este efecto se puede comprobar en la Tabla 3.6, donde también se verifica que el método NPB no se encuentra afectado por este problema de mala especificación, pues este no asume ningún modelo paramétrico concreto.

Tabla 3.6: Promedios de AE, SE y RSE ($\times 10^{-2}$) de $\widehat{Var}^*(\hat{\gamma}^*(u))$ para el estimador empírico y lineal local, considerando μ_1 y μ_2, con $\tau = 1/3$, $c_0 = 20\,\%$ y $a = 0{,}6$.

		AAE		ASE		ARSE	
Estimador	Tendencia	SPB	NPB	SPB	NPB	SPB	NPB
Empírico	$\mu_1(x, y)$	3.130	1.140	0.176	0.022	29.10	1.600
	$\mu_2(x, y)$	0.840	0.590	0.010	0.006	21.30	13.10
Lineal Local	$\mu_1(x, y)$	0.899	0.018	0.009	0.000	29.93	1.400
	$\mu_2(x, y)$	0.156	0.032	0.000	0.000	7.040	0.350

Los resultados de los estudios numéricos realizados en los distintos escenarios confirman la superioridad del método NPB, en especial en el caso de procesos espaciales no estacionarios en media, pues esta técnica recurre a un procedimiento de corrección de sesgo para aproximar la distribución de los estimadores del variograma. Además cuenta con la ventaja añadida de que este método no depende de la región de observación (al contrario del BB), ni requiere suponer ningún modelo paramétrico como en el caso del SPB, proporcionando estimaciones más robustas frente a los otros métodos bootstrap considerados.

A partir de este método bootstrap no paramétrico es factible realizar inferencia sobre características de interés de un proceso espacial con tendencia. Este método puede resultar adecuado para construir intervalos de confianza (como se presenta

en la Sección 3.5) o en contrastes de hipótesis sobre el variograma. El método NPB también puede ser utilizado, por ejemplo, en la construcción de mapas de riesgo geoestadístico. Una muestra de esta última aplicación se presenta a continuación.

3.4. Mapas de riesgos basados en NPB

Un objetivo de estudio muy importante dentro de muchos problemas medioambientales corresponde a la estimación de la incertidumbre de ciertas variables de interés. El control de los niveles de contaminación en el suelo o el aire, o la prevención de desastres naturales, son solo algunos ejemplos en los cuales se recurren a técnicas estadísticas que permitan estimar las probabilidades de que una variable exceda un valor límite (o umbral) en ciertas ubicaciones geógraficas. Estas herramientas que permiten construir mapas de probabilidades estimadas o mapas de riesgo, proporcionan información importante para los organismos de control, y suelen ser de gran ayuda para la toma de decisiones y en el diseño de políticas de prevención para evitar los efectos adversos de la contaminación, particularmente en aquellas áreas donde los niveles estimados de riesgo son elevados.

El enfoque tradicional en geoestadística consiste en la estimación de la probabilidad condicional de que una variable sobrepase un umbral determinado. Formalmente esto significa que, dado un vector de observaciones $\mathbf{Y}$ de un proceso espacial $\left\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\right\}$, se desea estimar la probabilidad condicional de que la variable Y exceda a un valor crítico c en una ubicación específica $\mathbf{x}_0$, $P\left(Y(\mathbf{x}_0) \geq c \,|\mathbf{Y}\right)$.

Aunque existen varios métodos geoestadísticos que permiten construir este tipo de mapas de probabilidad, como el kriging disyuntivo (DK) (ver p.e. Oliver *et al.*, 1996), o los modelos geoestadísticos basados en cadenas de Markov

(p.e. Li *et al.*, 2010), ha sido el kriging indicador (IK) (p.e. Goovaerts *et al.*, 1997) la técnica más utilizada y estudiada en este contexto. La idea general del IK consiste en aproximar la probabilidad condicional mediante la predicción lineal kriging a partir de funciones indicadoras $I_{\{Y(\mathbf{x}_0)\geq c\}}$, teniendo en cuenta que $\mathbb{E}\left(I_{\{Y(\mathbf{x}_0)\geq c\}}\,|\mathbf{Y}\right) = P\left(Y(\mathbf{x}_0)\geq c\,|\mathbf{Y}\right)$. (ver, p.e. Journel, 1983). Sin embargo, el IK presenta serias desventajas, tales como la pérdida de información debido a la discretización de los datos, o problemas de relación de orden (tales que si $c_1 < c_2$, no se puede asegurar que $I_{\{Y(\mathbf{x}_0)\geq c_1\}} \leq I_{\{Y(\mathbf{x}_0)\geq c_2\}}$, ver p.e. Chilès y Delfiner, 2012, p.384). Además el método IK puede dar lugar a estimaciones negativas de las probabilidades o también mayores que 1 (Chilès y Delfiner, 2012, Section 6.3.3).

Para evitar estas dificultades, Tolosana-Delgado *et al.* (2008) propusieron el método llamado "simplicial indicator kriging", el cual emplea una aproximación simplex para datos composicionales para estimar la función de distribución condicional acumulada. Este enfoque ha sido extendido al contexto bayesiano por Guardiola-Albert y Pardo-Igúzquiza (2011). Otra alternativa viable al método IK es el kriging disyuntivo (Oliver *et al.*, 1996). El método DK es una técnica de estimación no lineal que se suele aplicar bajo el supuesto de un modelo gaussiano isofactorial para el proceso geoestadístico (para más detalles al respecto, ver p.e. Wackernagel, 2003, Cap. 34, o Chilès y Delfiner, 2012, Sección 6.3.2). Sin embargo, Lark y Ferguson (2004) compararon ambos métodos y demostraron que no existe una evidencia empírica que determine una preferencia entre el IK o el DK.

Un aspecto crítico de todos los métodos anteriores, es que en la práctica asumen modelos paramétricos y por tanto, están expuestos a problemas de mala especificación. Por tal razón, se ha recurrido a técnicas no paramétricas, como por ejemplo, el método propuesto por García-Soidán *et al.* (2012), el cual sugiere utilizar estimadores tipo núcleo para el variograma indicador. En el contexto espa-

cio temporal, Draghicescu e Ignaccolo (2009) propuso estimar el riesgo mediante suavizado tipo núcleo en el dominio temporal conjuntamente con interpolación espacial. Cameletti *et al.* (2013) extendió esta última técnica, considerando variables adicionales exógenas en la interpolación kriging, y utilizando un algoritmo de bootstrap por bloques para obtener regiones de confianza para las probabilidades estimadas.

Por otra parte, este enfoque tradicional consistente en aproximar el riesgo mediante la estimación de probabilidades condicionales, (al igual que los métodos expuestos anteriormente), puede ser apropiado cuando interesa una realización específica del proceso, por ejemplo, para valoraciones de recursos mineros, o en predicciones del clima a corto plazo (comúnmente llamado predicción del tiempo). Sin embargo, en otros casos interesa estudiar la distribución del proceso espacial bajo ciertas condiciones (especificadas a través de una tendencia, o por variables exógenas), como es el caso de cierto tipo de estudios climáticos (p.e. en estudios sobre el calentamiento global). En estos casos, puede ser de mayor interés la estimación de la probabilidad incondicional, o también llamado *riesgo a largo plazo* (ver p.e Krzysztofowicz y Sigrest, 1997; Franks y Kuczera, 2002, para comentarios adicionales sobre las probabilidades condicionales e incondicionales).

En nuestro caso, nos centraremos en este último problema, es decir, estamos interesados en estimar la probabilidad incondicional de que la variable Y exceda un umbral c en una ubicación $\mathbf{x}_0$:

$$r_c\left(\mathbf{x}_0\right) = P\left(Y(\mathbf{x}_0) \geq c\right). \tag{3.3}$$

Es importante indicar que la probabilidad condicional puede ser muy distinta de la incondicional, debido a que la distribución de $Y(\mathbf{x}_0)$ condicionada a $\mathbf{Y}$ puede tener mucha menor variabilidad que la distribución marginal de $Y(\mathbf{x}_0)$, espe-

cialmente en localizaciones cercanas a los datos observados. Por esta razón, los métodos geoestadísticos descritos anteriormente no son apropiados para estimar la probabilidad incondicional. Asimismo, aunque estos métodos fueron originalmente diseñados para procesos estacionarios, también pueden ser adaptados para el caso de tendencia espacial no constante, en cuyo caso se suele considerar un modelo paramétrico para la tendencia y luego aplicar las técnicas respectivas sobre los residuos (con los consabidos problemas de sesgo mencionados en secciones anteriores).

3.4.1. Algoritmo NPB para mapas de riesgo

En esta sección se propone un procedimiento no paramétrico que permite estimar el riesgo espacial (3.3), para el caso de procesos espaciales con tendencia no constante de la forma (1.24) y sin necesidad de asumir ninguna forma paramétrica para la función tendencia o variograma. El procedimiento propuesto se detalla en el Algoritmo 3.3.

Este nuevo algoritmo es una modificación del método semiparamétrico propuesto en Francisco-Fernández *et al.* (2011) que se aplica en estudios sismológicos. El enfoque propuesto extiende el método anterior en dos direcciones. Primero, trata de reproducir de forma adecuada la variabilidad de los datos, para lo cual se recurre al método NPB expuesto anteriormente. En segundo lugar, la estimación de las probabilidades se realiza a partir de las predicciones espaciales obtenidas mediante kriging simple de los residuos, en lugar de utilizar las estimaciones de la tendencia como en el artículo original. Cabe mencionar que, al contrario de lo que ocurriría en la estimación de probabilidades condicionales, las réplicas bootstrap generadas al final del paso 1 en general no coinciden con los valores observados.

Alternativamente, los valores de la variable respuesta se pueden generar di-

Algoritmo 3.3: Algoritmo NPB para mapas de riesgo

1 Utilizar el método NPB para obtener la remuestra bootstrap del proceso espacial, en las posiciones observadas, tal que $Y^*(\mathbf{x}_i) = \hat{\mu}_{\mathbf{H}}(\mathbf{x}_i) + \varepsilon_i^*$, $i = 1, 2, \ldots, n$;

2 Para cada localización no muestreada $\mathbf{x}_0$, obtener $\hat{\mu}^*_{\mathbf{H}}(\mathbf{x}_0)$ mediante el estimador lineal local de la tendencia (2.1) aplicado sobre la réplica bootstrap $\{Y^*(\mathbf{x}_1), \ldots, Y^*(\mathbf{x}_n)\}$, utilizando la misma matriz ventana $\mathbf{H}$ del paso anterior;

3 Obtener $\hat{\varepsilon}^*(\mathbf{x}_0)$, mediante la predicción kriging simple (ver Sección 1.3.1) obtenida a partir de los residuos correspondientes;

4 Calcular las predicciones kriging $\hat{Y}^*(\mathbf{x}_0) = \hat{\mu}^*_{\mathbf{H}}(\mathbf{x}_0) + \hat{\varepsilon}^*(\mathbf{x}_0)$;

5 Repetir los pasos 1 al 4 un gran número de veces B. De tal manera, para cada $\mathbf{x}_0$, se obtienen B réplicas bootstrap $\hat{Y}^{*(1)}(\mathbf{x}_0), \ldots, \hat{Y}^{*(B)}(\mathbf{x}_0)$;

6 Finalmente, se obtiene un mapa con las estimaciones del riesgo incondicional dado por (3.3), calculando las frecuencias relativas obtenidas a partir de las réplicas bootstrap, midiendo la cantidad de veces que la observación bootstrap supera dicho umbral en cada localización $\mathbf{x}_0$, es decir, calculando:

$$\hat{r}_c\left(\mathbf{x}_0\right) = \frac{1}{B}\sum_{j=1}^{B} I_{\left\{\hat{Y}^{*(j)}(\mathbf{x}_0) \geq c\right\}}, \tag{3.4}$$

rectamente en las ubicaciones de predicción en lugar de las ubicaciones originales (y por tanto se eliminaría el paso 2 del algoritmo). De esta forma se esperaría que la variabilidad del proceso se reprodujese de mejor manera. Sin embargo, se obtuvieron resultados muy similares con ambas aproximaciones. Una ventaja del algoritmo propuesto, es que permite realizar inferencias sobre las características del proceso espacial (por ejemplo, sobre la tendencia o el variograma) de manera simultánea. Además, este método se puede utilizar para la construcción de intervalos de predicción o contrastes de hipótesis, o incluso se puede aplicar a datos independientes.

3.4.2. Resultados de simulación

El método propuesto fue analizado mediante estudios de simulación, considerando distintos escenarios. En cada caso, se tomaron $N = 1,000$ muestras de tamaño $n = 10 \times 10$, 17×17 y 20×20 generadas sobre una rejilla regular definida sobre la región $D = [0, 1]^2$. Los datos simulados responden al modelo (1.24), donde la función tendencia tiene la forma $\mu(x_1, x_2) = 2{,}5 + \sin(2\pi x_1) + 4(x_2 - 0{,}5)^2$, y los errores ε_i siguen una distribución normal de media cero y variograma isotrópico exponencial (1.17). Los parámetros considerados para este modelo de dependencia fueron $a = 0{,}25$, $0{,}50$ y $0{,}75$, $\sigma^2 = 0{,}16$, $0{,}32$, y $0{,}64$, con efectos nugget de $0\,\%$, $25\,\%$ y $50\,\%$ de σ^2. Estos valores fueron seleccionados para poder realizar comparaciones con el método presentado en Francisco-Fernández *et al.* (2011).

Considerando las funciones teóricas de tendencia y variograma, y dado un valor de umbral c, se calcularon las probabilidades teóricas $r_c(\mathbf{x}_0)$, dadas por (3.3) para una rejilla regular de predicción de tamaño 50×50. Luego, para cada simulación se aplicó el método bootstrap propuesto considerando $B = 1000$, y de esta forma se obtuvieron las estimaciones de las probabilidades $\hat{r}_c(\mathbf{x}_0)$ mediante la ecuación (3.4). Este proceso se repitió para distintos valores de umbral $c = 2{,}0$, $2{,}5$, $3{,}0$ y $3{,}5$.

La estimación de la tendencia se realizó de forma similar a los estudios de simulación presentados en la Sección 3.3.2, es decir, utilizando el estimador lineal local dado por (2.1) con matriz ventana $\mathbf{H}_{\text{MASE}}$ que minimiza el criterio $MASE$ (2.6). De igual manera, la estimación del semivariograma de los residuos y su versión corregida se realizó a partir del estimador lineal local correspondiente (2.13), considerando los saltos de manera similar al estudio presentado en la Sección 3.3. La ventana g se obtuvo mediante el criterio de minimización del error cuadrático relativo dado por (2.15). Finalmente, se ajustó un modelo isotrópico de Shapiro-

Botha a los variogramas estimados anteriores para obtener las correspondientes matrices de varianzas y covarianzas necesarias de cara a aplicar el método NPB.

Para estudiar el efecto del método de estimación de la matriz de varianzas y covarianzas en el comportamiento del algoritmo bootstrap, se consideraron tres aproximaciones distintas. En primer lugar, se aplicó el método propuesto utilizando la matriz teórica $\mathbf{\Sigma}$ para aproximar las probabilidades de exceder el umbral determinado (al que denominaremos como método téorico en los resultados mostrados a continuación). Para el segundo procedimiento, las estimaciones del riesgo se obtuvieron a partir de la matriz estimada de varianzas y covarianzas de los residuos sin corregir $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$ (denotado como método residual). Finalmente, el procedimiento completo propuesto en la Sección 3.4.1, o método corregido, se aplicó para obtener las correspondientes probabilidades estimadas (3.4).

Para comparar el comportamiento de estas tres distintas aproximaciones, se calcularon los errores cuadráticos:

$$\mathrm{SE}\left(\mathbf{x}\right) = \left(r_c\left(\mathbf{x}\right) - \hat{r}_c\left(\mathbf{x}\right)\right)^2,$$

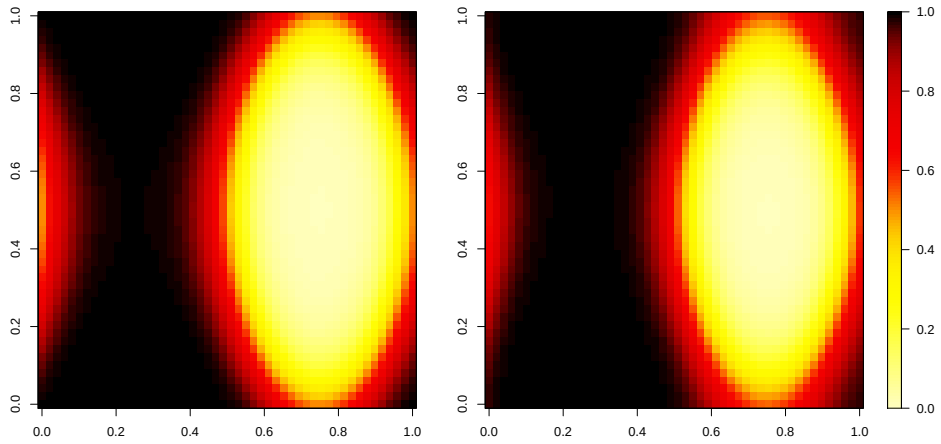
los cuales fueron calculados en las posiciones de la rejilla de predicción.

Como se observó un comportamiento parecido en los diferentes escenarios de simulación considerados, se muestran solamente algunos resultados representativos. Por ejemplo, la media, mediana y desviación estándar de los errores cuadráticos ($\times 10^{-2}$), para $c = 2{,}5$, $\sigma^2 = 0{,}16$, $r = 0{,}5$ y $c_0 = 0{,}04$, se presentan en la Tabla 3.7. Como cabría esperar, los mejores resultados se obtienen para el método teórico (cuando se utiliza la matriz de varianzas y covarianzas $\mathbf{\Sigma}$). Sin embargo, el método corregido proporciona resultados muy próximos a los valores teóricos, sobre todo cuando se comparan las medias de los errores cuadráticos. De manera general se observa que el procedimiento de corrección reduce aproximadamente

en un 40 % las medias de los errores obtenidas por el procedimiento basado en el uso directo de los residuos (método residual). Asimismo, el promedio de errores cuadráticos obtenidos por el método propuesto son aproximadamente un 55 % más bajos que aquellos presentados en el trabajo de Francisco-Fernández *et al.* (2011).

Tabla 3.7: Estadísticos de errores cuadráticos ($\times 10^{-2}$) para los métodos teórico, residual y corregido, para el umbral $c = 2{,}5$, $\sigma^2 = 0{,}16$, $a = 0{,}5$ y $c_0 = 0{,}04$, considerando los tamaños muestrales $n = 10 \times 10$, 17×17 y 20×20.

n	Método	Media	Mediana	Desv. Est.
10×10	Teórico	2.30	0.09	5.40
	Residual	5.00	0.16	11.00
	Corregido	2.60	0.09	6.40
17×17	Teórico	2.05	0.08	4.97
	Residual	4.10	0.16	9.40
	Corregido	2.40	0.09	5.80
20×20	Teórico	1.90	0.08	4.65
	Residual	3.80	0.16	8.60
	Corregido	2.20	0.08	5.40



(a) (b)

Figura 3.6: (a) Probabilidades teóricas de exceder el umbral de $c = 2{,}5$, y (b) medias obtenidas por simulación de las correspondientes probabilidades estimadas usando el método corregido, para $n = 20 \times 20$, $\sigma^2 = 0{,}16$, $a = 0{,}5$ y $c_0 = 0{,}04$.

Este buen comportamiento del procedimiento propuesto puede visualizarse

también en las Figuras 3.6(a) y 3.6(b), en las cuales las probabilidades teóricas $r_c(\mathbf{x}_0)$, para un umbral de $c = 2{,}5$, se comparan con los correspondientes promedios (calculados a partir de las N réplicas) de las probabilidades estimadas $\hat{r}_c(\mathbf{x}_0)$ obtenidas mediante el método corregido.

Por otra parte, las Figuras 3.7(a), 3.7(b) y 3.7(c) muestran los errores cuadráticos medios aproximados por simulación, obtenidos mediante el método corregido, considerando un valor de umbral de $c = 2{,}5$, $\sigma^2 = 0{,}16$, $r = 0{,}5$, $c_0 = 0{,}04$, y los distintos tamaños muestrales. A partir de estos gráficos (y de los resultados de la Tabla 3.7) se puede inferir la consistencia del algoritmo propuesto.

El efecto de la dependencia espacial se puede analizar a partir de los resultados que se muestran en la Tabla 3.8 (con valores de $c = 2{,}5$, $\sigma^2 = 0{,}16$, $n = 20 \times 20$, y distintos valores para el nugget y el rango práctico). En todos los casos, el método corregido presenta mejores resultados respecto a las medias de los errores cuadráticos. Además, como era de esperar, la precisión de la estimación mejora conforme disminuye el grado de dependencia espacial (es decir, cuando el efecto nugget se incrementa o el rango decrece, se obtienen errores medios más bajos; ver p.e. Cressie, 1993, Sección 1.3., para comentarios generales acerca de la estimación bajo dependencia).

Tabla 3.8: Medias de errores cuadráticos ($\times 10^{-2}$) de las probabilidades estimadas mediante el método residual y corregido, para umbral $c = 2{,}5$, $\sigma^2 = 0{,}16$, $n = 20 \times 20$, y distintos valores de efecto nuggets y rangos prácticos.

Nugget	$c_0 = 0\,\%$		$c_0 = 25\,\%$		$c_0 = 50\,\%$	
Método	Residual	Corregido	Residual	Corregido	Residual	Corregido
$a = 0{,}25$	2.90	1.60	3.00	1.60	3.00	1.80
$a = 0{,}50$	3.80	2.34	3.80	2.20	3.70	2.25
$a = 0{,}75$	4.50	2.90	4.40	2.70	4.10	2.70

Por último, se realizaron estudios de simulación considerando muestreo espa-

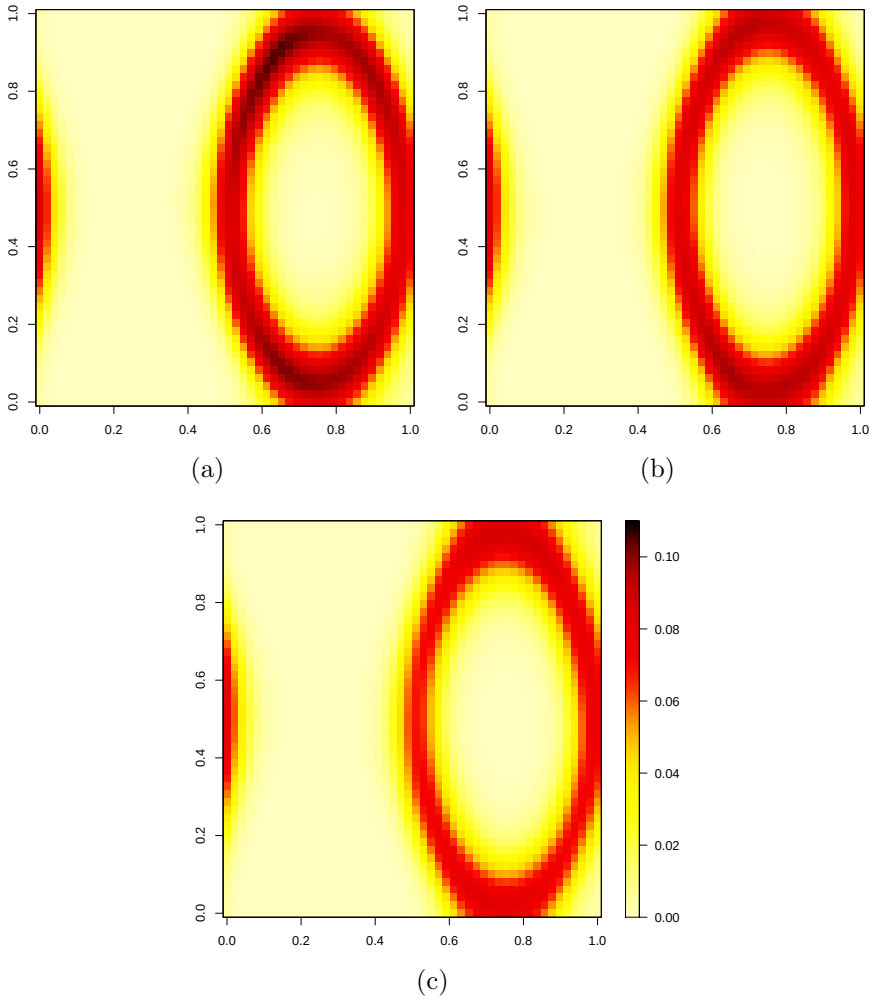


Figura 3.7: Superficies de errores cuadráticos medios, obtenidos mediante el método corregido para $c = 2{,}5$, $\sigma^2 = 0{,}16$, $r = 0{,}5$, $c_0 = 0{,}04$, y distintos tamaños muestrales: (a) $n = 10 \times 10$, (b) $n = 17 \times 17$ y (c) $n = 20 \times 20$.

cial irregular. Los resultados obtenidos en estos casos siguen la misma línea de los obtenidos bajo muestreo regular, y por tanto las conclusiones se mantienen válidas en ambas situaciones. A modo de ejemplo, la Tabla 3.9 presenta resultados similares a los que se obtuvieron en la Tabla 3.7, considerando ahora que las localizaciones muestrales se generaron a partir de una distribución uniforme sobre el cuadrado unitario $[0, 1]^2$. Aunque los resultados bajo diseño espacial irregular son parecidos a los obtenidos anteriormente, es importante indicar que en el primer caso el coste computacional es mucho mayor. Esto se debe a que en diseño fijo,

Tabla 3.9: Estadísticos de errores cuadráticos ($\times 10^{-2}$) para los métodos teórico, residual y corregido, bajo muestreo irregular (con distribución uniforme sobre la región $[0,1]^2$), considerando el umbral de $c = 2{,}5$, $\sigma^2 = 0{,}16$, $a = 0{,}5$ y $c_0 = 0{,}04$, y tamaños muestrales de $n = 10 \times 10$, 17×17 y 20×20.

n	Método	Media	Mediana	Desv. Est.
10×10	Teórico	2.50	0.13	6.00
	Residual	4.50	0.17	10.00
	Corregido	2.70	0.17	6.50
17×17	Teórico	2.20	0.09	5.20
	Residual	4.52	0.21	10.70
	Corregido	2.30	0.10	5.60
20×20	Teórico	2.10	0.08	5.10
	Residual	6.20	0.34	14.00
	Corregido	2.28	0.09	5.63

los parámetros ventana para la estimación de la tendencia y del variograma, así como la matriz de suavizado $\mathbf{S}$ solo se calculan una vez, mientras que en el caso de diseño aleatorio, estas matrices debe ser recalculadas en cada iteración.

3.5. Aplicación a datos reales

Para ilustrar la aplicación en la práctica del método NPB propuesto en la Sección 3.2 y del algoritmo bootstrap para la construcción de mapas de riesgos no paramétricos (expuesto en la Sección 3.4.1), consideraremos el conjunto de datos sobre el total de precipitaciones en la parte continental de Estados Unidos de Norteamérica, que se analizó anteriormente en la Sección 2.6.

Para obtener las matrices de varianzas y covarianzas necesarias para el NPB, se realizaron dos iteraciones del Algoritmo 2.2 de estimación NP conjunta. En la primera etapa, la estimación piloto de la tendencia se obtuvo seleccionando una ventana piloto inicial $\mathbf{H}^{(0)}$ por CV, en el paso 1. Luego, utilizando las mismas especificaciones utilizadas en el estudio realizado en el capítulo anterior, se

estimó el variograma lineal local $\hat{\gamma}(u_i)$ (2.13) asumiendo isotropía considerando los 60 saltos equidistantes u_i. Luego, mediante el procedimiento de corrección de sesgo del variograma, se obtuvo la versión corregida del variograma estimado y posteriormente se le ajustó un modelo válido de Shapiro-Botha para construir la estimación de la matriz de correlación $\mathbf{R}$ (paso 3). A partir de esta matriz, en el paso 4 se aplicó el criterio $CGCV$ (2.9), obteniéndose la nueva ventana $\mathbf{H} = diag(10{,}09, 18{,}40)$, con la cual se construyó una nueva estimación de la tendencia

En la segunda iteración, se utilizaron los nuevos residuos para reestimar el variograma residual $\hat{\gamma}(\cdot)$, observándose que ambas estimaciones proporcionaban resultados similares, razón por la cual se decidió no continuar con las iteraciones. La estimación lineal local de la tendencia obtenida con la ventana $\mathbf{H}$ final se presenta en la Figura 3.8.

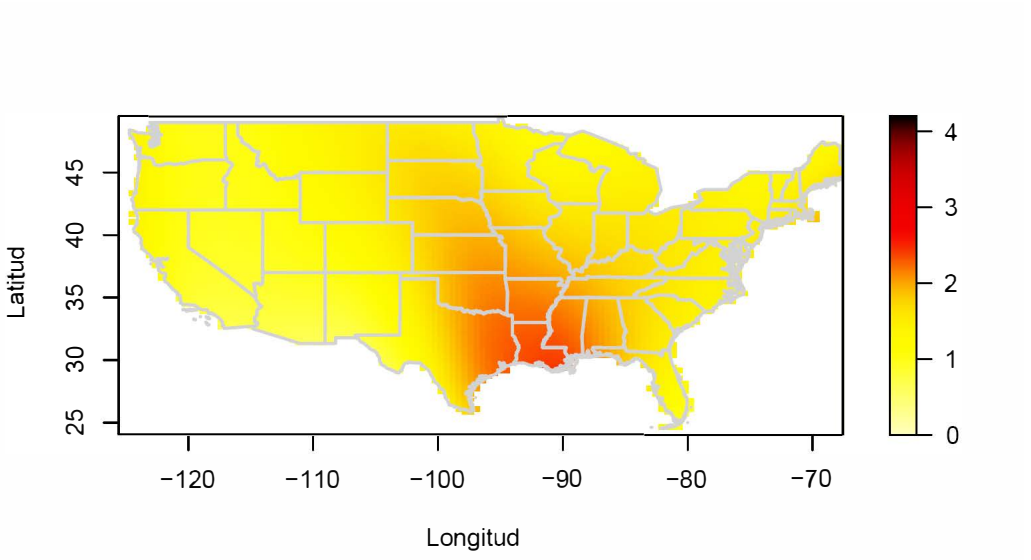
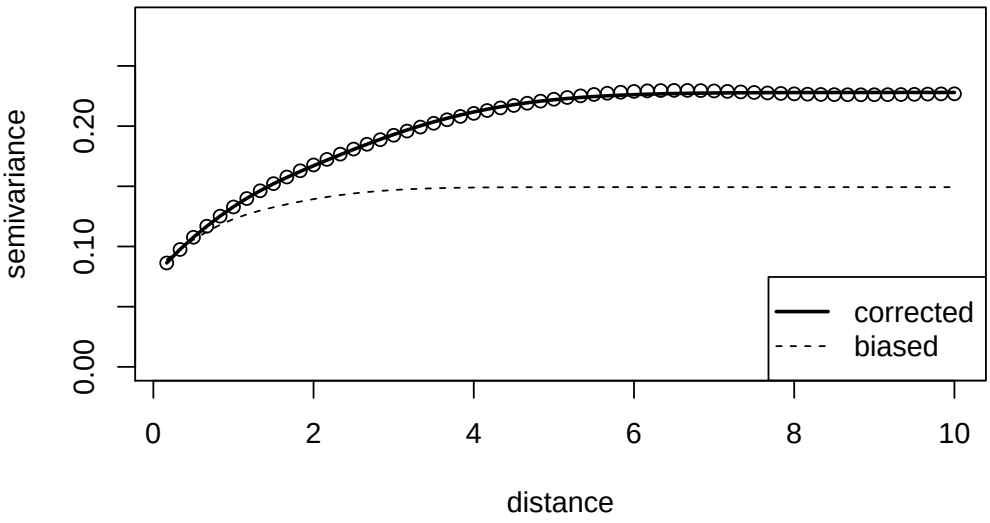


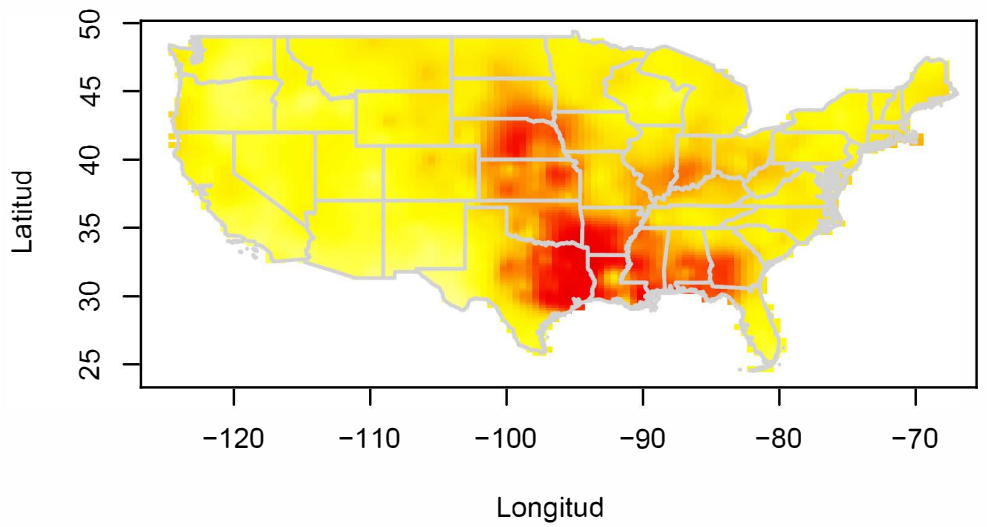
Figura 3.8: Tendencia lineal local final estimada con ventana $\mathbf{H}$ que minimiza el criterio $CGCV$, para los datos del total de precipitaciones (en raíz cuadrada de pulgadas de lluvia) registradas en EEUU durante Marzo 2016

Luego, a partir del variograma residual $\hat{\gamma}(\cdot)$ obtenido en la segunda iteración,

se obtuvo la correspondiente versión corregida $\tilde{\gamma}(\cdot)$ y se ajustaron los modelos de Shapiro-Botha respectivos, los cuales se representan en la Figura 3.9(a).



(a)



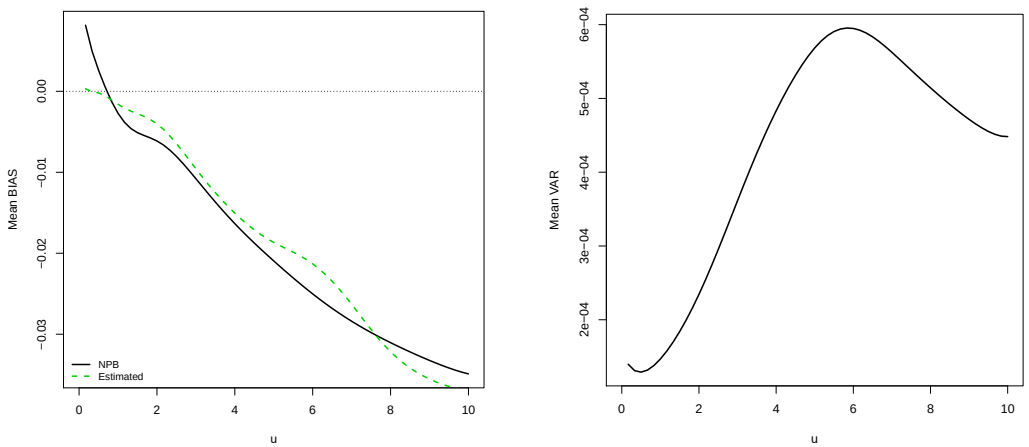
(b)

Figura 3.9: (a) Modelos de Shapiro-Botha ajustados a las estimaciones residuales (línea discontinua) y corregidas (línea continua) del variograma, y (b) predicciones kriging, correspondientes al total de precipitaciones en Marzo 2016 (medida en raíz cuadrada de pulgadas de lluvia)

Este gráfico pone en evidencia el efecto del sesgo debido al uso de los residuos no corregidos, pues su correspondiente variograma subestima la variabilidad es-

pacial, lo cual puede influir significativamente en las inferencias que se realicen sobre el proceso espacial. Además, utilizando las estimaciones finales de la tendencia y el variograma corregido fue factible obtener las predicciones kriging, las cuales se muestran en la Figura 3.9(b).

A partir de estos modelos ajustados de variograma se pudieron estimar las matrices $\boldsymbol{\Sigma}$ y $\boldsymbol{\Sigma}_{\hat{\varepsilon}}$ necesarias para aplicar el método bootstrap no paramétrico. Repitiendo este procedimiento $B = 1000$ veces, se obtuvieron las aproximaciones bootstrap del sesgo y la varianza mediante las expresiones (3.2) y (3.1) respectivamente. Los sesgos bootstrap se compararon con el sesgo estimado, calculado mediante la diferencia entre el estimador corregido y el estimador residual del variograma. Los resultados obtenidos mediante el método NPB para las aproximaciones bootstrap del sesgo y la varianza se presentan en las Figuras 3.10(a) y 3.10(b) respectivamente.



(a) (b)

Figura 3.10: (a) Sesgo estimado, aproximado por ($\hat{\gamma}(u)$ - $\tilde{\gamma}(u)$) y $\widehat{Bias}^*(\hat{\gamma}^*(u))$ (líneas discontinua y continua respectivamente), y (b) $\widehat{Var}^*(\hat{\gamma}^*(u))$, obtenidas mediante el método NPB.

La Figura 3.10(a) pone en evidencia la apropiada aproximación del sesgo

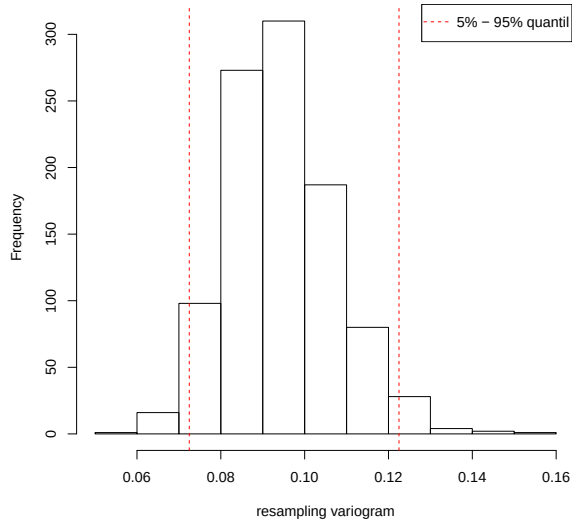
bootstrap respecto al sesgo estimado, el cual toma valores pequeños en saltos cercanos al origen, y se incrementa a medida que el salto es más grande. Por otra parte, la aproximación bootstrap de la varianza del estimador, mostrado en la Figura 3.10(b), tiene un comportamiento similar a los resultados obtenidos por simulación en la Sección 3.3.2.

Como se mencionó en la sección 3.2, a partir del NPB es factible construir intervalos de confianza puntuales para $\gamma(u_i)$, de la forma siguiente (ver p.e. Davison y Hinkley, 1997, Sección 5.2):

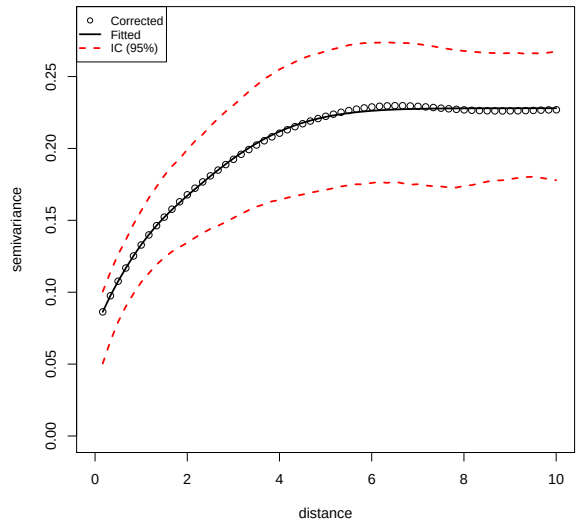
$$\left[\tilde{\gamma}(u_i) + \hat{\gamma}(u_i) - \hat{\gamma}^*_{[(B+1)(1-\alpha/2)]}(u_i), \tilde{\gamma}(u_i) + \hat{\gamma}(u_i) - \hat{\gamma}^*_{[(B+1)(\alpha/2)]}(u_i)\right], i = 1, \ldots, B.$$

donde $\hat{\gamma}^*_{[(B+1)(\alpha)]}(u_i)$ es la aproximación del cuantil de orden α obtenido a partir de las $B = 1000$ réplicas bootstrap. A modo de ejemplo, la Figura 3.11(a) muestra el histograma resultante y los cuantiles al 5 % y 95 % para el salto u_1. En la Figura 3.11(b) representa de forma conjunta, los intervalos de confianza al 95 % (líneas discontinuas), el variograma estimado corregido (línea de puntos) y la versión ajustada del modelo de Shapiro-Botha (línea continua). Estos resultados podrían ser de utilidad para seleccionar modelos paramétricos de variograma o diagnosticar independencia de los datos espaciales, entre otras posibles aplicaciones.

Finalmente, se construyó un mapa de riesgo geoestadístico para la región de observación considerada, utilizando el procedimiento descrito en la Sección 3.4.1. En este caso, se estimó la probabilidad de que la cantidad total de precipitación superase el valor umbral c en dicha región, para distintos valores de c. Las Figuras 3.12(a) y 3.12(b) muestran los mapas de riesgo $\hat{r}_c(\mathbf{x})$, para $c = 1{,}0$ y $2{,}0$ (1.0 y 4.0 pulgadas de lluvia, respectivamente). Estos mapas indican que existen áreas con escaso riesgo de precipitación alta, correspondientes a los estados de Arizona, Utah y Nuevo México. Estas regiones se caracterizan por ser área secas con clima



(a)

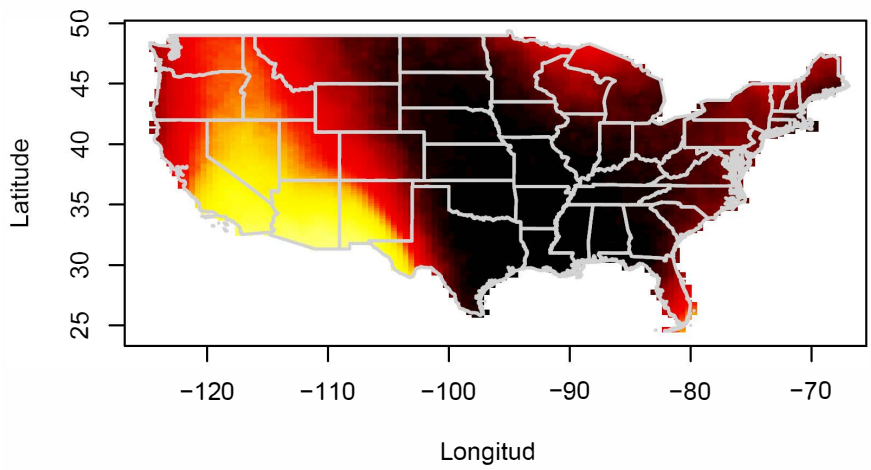


(b)

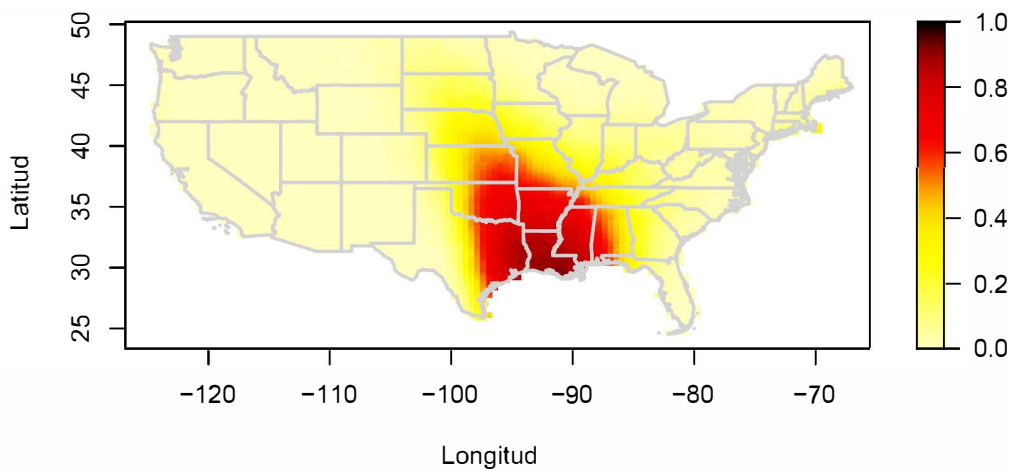
Figura 3.11: (a) Histograma de las réplicas bootstrap para el variograma estimado en el salto u_1, y valores de los cuantiles 5 % y 95 % (líneas discontinuas). (b) Estimador lineal local del variograma corregido (línea de puntos), variograma válido de Shapiro-Botha (línea continua) y límites del intervalo de confianza al 95 % (líneas discontinuas).

semiárido. Asimismo, estos mapas indican zonas con alta probabilidad de lluvias, en especial en ciertos estados sureños. Cabe mencionar, que en el período consi-

derado, se produjeron lluvias fuertes e inundaciones en estas áreas, sobre todo en las riberas del río Sabine, localizado entre los estados de Texas y Louisiana. Estos mapas de probabilidad incondicional estimada constituyen un ejemplo simple del tipo de resultados que se pueden obtener gracias al uso del método NPB. Además, estos métodos podrían también ser de utilidad para analizar conjuntos de datos más amplios, y considerando modelos más complejos (como en el caso espacio-temporal), lo cual constituye una interesante línea de trabajo futuro.



(a)



(b)

Figura 3.12: Mapas con las probabilidades estimadas incondicionales, $\hat{r}_c(\mathbf{x})$, de la ocurrencia de una precipitación total mayor o igual a (a) $c = 1{,}0$ y (b) $c = 2{,}0$ (medida en raíz cuadrada de pulgadas de lluvia) para el área de estudio.

Capítulo 4

Estimación no paramétrica en procesos geoestadísticos heterocedásticos

En los capítulos anteriores se ha supuesto que el proceso espacial es no estacionario en media, es decir, que la tendencia cambia dependiendo la posición espacial. También se ha supuesto que la variación de pequeña escala (dependencia espacial) en el modelo (1.24) es estacionaria. Por tanto, la caracterización del proceso espacial se reduce a la estimación del variograma de los errores. Desde esta perspectiva, los modelos anteriores se podrían definir como *homocedásticos*.

Sin embargo, existen procesos espaciales no estacionarios en los que la estructura de variabilidad de pequeña escala varía localmente. Dependiendo de las suposiciones que se hagan sobre los modelos de variabilidad espacial no estacionaria, se recurren a distintos procedimientos, como por ejemplo, los modelos de deformación espacial (Sampson y Guttorp, 1992; Anderes y Stein, 2008), o aquellos basados en la descomposición espectral (Fuentes, 2001, 2002). Sin embargo, debido a la complejidad de este tipo de modelos, el proceso de estimación de los

parámetros suele ser complicado y en muchos casos se requiere contar con un volumen considerable de datos para realizar dicha estimación.

En el presente capítulo, se trabajará con procesos espaciales heterocedásticos, desde la perspectiva de la regresión tipo núcleo bajo errores correlacionados heterocedásticos, basándonos en especial en los métodos no paramétricos para estimar la función varianza, como se puede observar en los trabajos de Ruppert *et al.* (1997) Opsomer *et al.* (1999), Vilar-Fernández y Francisco-Fernández (2006), Pérez-González *et al.* (2010) entre otros.

En la Sección 4.1 se presenta el modelo espacial heterocedástico, así como ciertas propiedades y diferencias respecto al modelo homocedástico considerado hasta ahora. A continuación, se exponen los métodos no paramétricos comúnmente utilizados para estimar la función varianza, y un método general de estimación basado en residuos para procesos espaciales heterocedásticos. Sin embargo, los métodos anteriores suelen ignorar los efectos de la correlación espacial o el sesgo debido al uso de residuos. Por tal razón se introduce un nuevo método iterativo de estimación y corrección de sesgo en la Sección 4.2. Este procedimiento permite obtener de forma conjunta estimaciones no paramétricas de las funciones tendencia y varianza, así como un estimador piloto del variograma del proceso de error. Estas aproximaciones se obtienen mediante suavizado lineal empleando el estimador lineal local. Finalmente, en la Secciones 4.3 y 4.4 se estudia el comportamiento de los algoritmos propuestos con datos simulados y reales, respectivamente.

Las principales contribuciones y algunos de los resultados del presente capítulo se encuentran en Fernández-Casal *et al.* (2017b).

4.1. Procesos geoestadísticos heterocedásticos

Consideremos que el proceso espacial $\{Y(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d\}$ puede ser modelado de la siguiente manera:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon(\mathbf{x}), \tag{4.1}$$

donde $\mu(\cdot)$ es la tendencia determinística del proceso y $\sigma^2(\cdot)$ es la función varianza, determinística y estrictamente positiva. Además, el término de error $\varepsilon(\cdot)$ corresponde ahora a un proceso estacionario de segundo orden, tal que $\mathbb{E}(\varepsilon(\mathbf{x})) = 0$ y $Var(\varepsilon(\mathbf{x})) = 1$, para todo $\mathbf{x} \in D$.

Bajo este supuesto, el correlograma del proceso de error $\varepsilon(\cdot)$, viene dado por $\rho(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$, y su variograma correspondiente es:

$$2\gamma_\varepsilon(\mathbf{u}) = Var(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u})) = 1 - \rho(\mathbf{u}) \tag{4.2}$$

Por otra parte, el covariograma del proceso espacial heterocedástico $Y(\cdot)$ bajo el modelo (4.1) viene dado por:

$$Cov\left[Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})\right] = \sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{u})\rho(\mathbf{u}). \tag{4.3}$$

Asimismo, el variograma del proceso heterocedástico se escribiría como:

$$2\gamma(\mathbf{x}, \mathbf{x}+\mathbf{u}) = (\sigma(\mathbf{x}) - \sigma(\mathbf{x}+\mathbf{u}))^2 + 2\sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{u})\gamma_\varepsilon(\mathbf{u}), \tag{4.4}$$

donde se verifica que este variograma no es estacionario, pues a diferencia de lo que ocurre con el variograma del error (4.2), el variograma del proceso (4.4) no depende únicamente del salto $\mathbf{u}$, sino también de la función varianza evaluada en

las posiciones espaciales $\mathbf{x}$ y $\mathbf{x}+\mathbf{u}$. Por tanto, la caracterización de la variabilidad de pequeña escala en un proceso heterocedástico implica la estimación tanto de $\sigma(\cdot)$ como de $\gamma_\varepsilon(\mathbf{u})$. De esta última expresión también se verifica que si $\sigma(\cdot)=1$ (es decir, si el proceso es homocedástico), el variograma del proceso espacial coincide con el variograma de los errores.

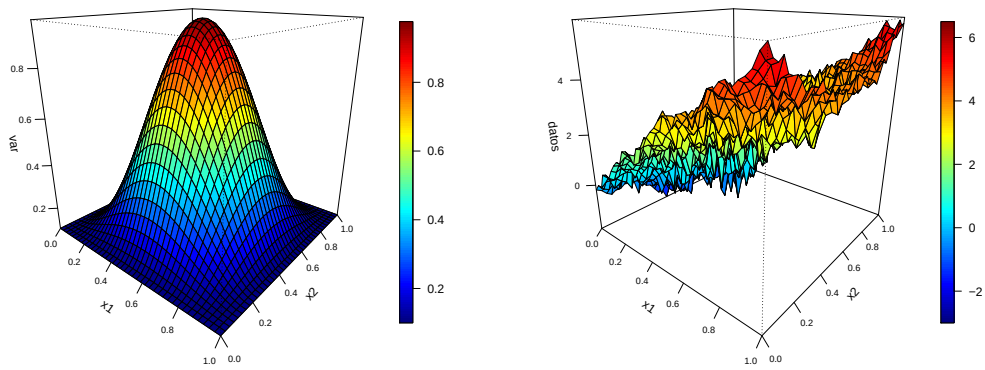
Para un vector $\mathbf{Y}$ de n observaciones del proceso heterocedástico, y a partir de (4.3), se obtiene (ver Ruppert *et al.*, 1997, Lema 2, pp.272):

$$\boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}^t \odot \boldsymbol{R} \tag{4.5}$$

donde $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas de $\mathbf{Y}$, $\boldsymbol{R}$ es la matriz de correlación de los errores estacionarios $\varepsilon(\mathbf{x}_i)$, con $i=1,\ldots,n$, $\boldsymbol{\sigma}=(\sigma(\mathbf{x}_1),\ldots,\sigma(\mathbf{x}_n))^t$, y $\odot$ representa el producto Hadamard de matrices.

A modo de ejemplo, se representan $n=40\times 40$ datos espaciales simulados bajo este modelo en una rejilla regular bidimensional definida en $D=[0,1]^2\subset\mathbb{R}^2$, utilizando las funciones tendencia y variograma del ejemplo mostrado en el caso homocedástico (ver Fig. 1.8(a) y Fig. 1.4(a) respectivamente), añadiendo ahora la componente $\sigma^2(x_1,x_2)=\left(\frac{15}{16}\right)^2(1-(2x_1-1)^2)^2(1-(2x_2-1)^2)^2+0{,}1$ representada en la Figura 4.1(a). En los resultados que se presentan en la Figura 4.1(b), se puede observar claramente el efecto que tiene la función varianza sobre el comportamiento de estos datos simulados al compararlos con el proceso homocedástico (ver 1.8(b)), sobre todo en las regiones en las cuales la varianza alcanza valores más grandes.

El modelo (4.1) nos permite analizar el proceso espacial $Y(\cdot)$ bajo el enfoque de la regresión heterocedástica, en el cual el modelo de regresión se representa a través de una función tendencia más una función de varianza no constante y un término de error o ruido. Luego, es factible extender varias de las técnicas



(a) (b)

Figura 4.1: (a) Función Varianza $\sigma^2(x_1, x_2) = \left(\frac{15}{16}\right)^2 (1 - (2x_1 - 1)^2)^2(1 - (2x_2 - 1)^2)^2 + 0{,}1$ y (b) Datos espaciales simulados sobre $D = [0, 1]^2$, bajo el modelo $Y(\mathbf{x}) = \mu(\mathbf{x}) + \sigma(\mathbf{x})\varepsilon(\mathbf{x})$, tomando $\mu(x_1, x_2) = 5{,}8(x_1 - x_2 + x_2^2)$, y semivariograma isotrópico exponencial $\gamma_\varepsilon(u)$ con $\sigma^2 = 1$, $a = 0{,}6$ y $c_0 = 0{,}2$.

de regresión heterocedástica para caracterizar las componentes de dicho modelo, tomando en consideración que en este caso los errores presentan una correlación espacial.

4.1.1. Estimación no paramétrica de la función varianza

Desde el contexto de la regresión para datos incorrelados, se han propuesto varios métodos de estimación no paramétrica de la función varianza, aunque generalmente se recurre al método basado en las diferencias de los momentos del proceso espacial (Härdle y Tsybakov, 1997), o al método basado en la aproximación de la media de los residuos cuadrados (Fan y Yao, 1998).

Supongamos que el vector $\mathbf{Y}$ corresponde a n observaciones de un proceso aleatorio bajo el modelo heterocedástico (4.1), donde el término $\varepsilon(\mathbf{x})$ es un proceso de ruido independiente con media cero y varianza unidad. El método basado en diferencia, trata de aproximar la función $\sigma(\cdot)$ a partir de la expresión:

$$\sigma^2(\mathbf{x}_i) = \mathbb{E}[Y^2(\mathbf{x}_i)] - \mathbb{E}[Y(\mathbf{x}_i)]^2 = g(\mathbf{x}_i) - \mu^2(\mathbf{x}_i),\ i = 1, \ldots, n$$

Este procedimiento generalmente consta de los siguientes pasos:

1. Obtener el estimador polinómico local de grado p_1, con núcleo K_1 y ventana h_1 para estimar $\mu(\mathbf{x}_i)$.

2. A partir de los valores $Y^2(\mathbf{x}_i)$, obtener $\hat{g}(\mathbf{x}_i)$ por regresión polinómica local con la misma ventana del paso 1.

3. Calcular $\hat{\sigma}^2(\mathbf{x}_i) = \hat{g}(\mathbf{x}_i) - \hat{\mu}^2(\mathbf{x}_i)$

El método basado en residuos cuadrados, en cambio, parte de la siguiente definición:

$$\sigma^2(\mathbf{x}_i) = \mathbb{E}[(Y(\mathbf{x}_i) - \mu(\mathbf{x}_i))^2], i = 1 \ldots, n$$

donde el término $(Y(\mathbf{x}_i) - \mu(\mathbf{x}_i))$ representan los errores heterocedásticos (independientes e iguales a $\sigma(\mathbf{x}_i)\varepsilon(\mathbf{x}_i)$). Los pasos a seguir son:

1. Obtener $\hat{\mu}(\mathbf{x}_i)$ mediante regresión polinómica local de grado p_1, con núcleo K_1 y ventana h_1.

2. Calcular los residuos $r_i = Y(\mathbf{x}_i) - \hat{\mu}(\mathbf{x}_i)$ y obtener r_i^2.

3. Estimar $\sigma^2(\mathbf{x}_i)$ a partir de r_i^2 mediante regresión polinómica local de grado p_2, con núcleo K_2 y ventana h_2.

Las propiedades estadísticas de ambos estimadores han sido estudiadas en datos correlacionados unidimensionales (ver p.e. Vilar-Fernández y Francisco-Fernández, 2006). Estos autores observaron de forma general, el mejor comportamiento para el caso del estimador basado en residuos cuadrados. Esto ha sido confirmado, incluso en estudios de estimación con datos perdidos (Pérez-González

et al., 2010). Sin embargo, para el caso de datos dependientes, ambos procedimientos introducen sesgos en el estimador de la varianza (Ruppert *et al.*, 1997).

4.1.2. Estimación basada en residuos de un proceso espacial heterocedástico.

La estimación del proceso espacial bajo el modelo (4.1) implica la necesidad de aproximar la variabilidad de gran escala, representada a través de la función tendencia, y de la variabilidad de pequeña escala, que en este caso está compuesta por la función varianza $\sigma(\cdot)$ y el variograma (o correlograma) de los errores $\varepsilon(\cdot)$.

Con este fin se han propuesto distintos enfoques, ya sea adaptando métodos estadísticos generales al contexto espacial o extendiendo técnicas geoestadísticas al caso heterocedástico, como es el caso de la estimación basada en residuos. En el primer caso se incluyen, por ejemplo, el uso de modelos de regresión lineales o parcialmente lineales aplicados en datos espaciales (ver p.e. Robinson y Thawornkaiwong, 2012). En el segundo grupo, Opsomer *et al.* (1999) utilizan una combinación de estimación no paramétrica de la varianza (mediante residuos cuadrados) conjuntamente con una tendencia aproximada paramétricamente y un modelo válido de variograma para realizar predicciones kriging con datos espaciales heterocedásticos. Sin embargo, en este último caso no se tiene en cuenta el sesgo debido al uso de residuos en la estimación de la función varianza ni en la del variograma.

De manera similar al caso de procesos con tendencia no constante, se puede utilizar un método de estimación basado en residuos para aproximar las componentes del modelo (4.1). Para esto, a partir de una realización del proceso espacial heterocedástico $\mathbf{Y}$, se efectuarían los siguientes pasos:

1. Estimar la tendencia $\mu(\cdot)$ del proceso y calcular los residuos $\mathbf{r} = (r_1, \ldots, r_n)$

tal que $\mathbf{r} = \mathbf{Y} - \hat{\boldsymbol{\mu}}$.

2. Obtener el estimador de la varianza $\hat{\boldsymbol{\sigma}} = (\hat{\sigma}(\mathbf{x}_1), \ldots, \hat{\sigma}(\mathbf{x}_n))$ mediante el método de residuos cuadrados, y posteriormente construir los residuos estandarizados estimados $\hat{e}(\mathbf{x}_i) = r_i/\hat{\sigma}(\mathbf{x}_i)$.

3. Aproximar el variograma de los errores $\gamma_\varepsilon(\mathbf{u})$ a partir de un estimador $2\hat{\gamma}_{\hat{e}}(\mathbf{u})$ previamente reescalado.

En este punto, es importante aclarar la notación utilizada en el procedimiento anterior, que será utilizada en las secciones siguientes. Mientras los *errores heterocedásticos* se denotan por la diferencia $(\mathbf{Y} - \boldsymbol{\mu})$, los residuos "heterocedásticos" (o simplemente *residuos*) $\mathbf{r}$ se obtienen a partir de la tendencia estimada $\hat{\boldsymbol{\mu}}$ (ver paso 1). Llamaremos *residuos estandarizados* al cociente $\tilde{\varepsilon}(\mathbf{x}_i) = r_i/\sigma(\mathbf{x}_i)$, mientras que si utiliza además la estimación de la función varianza, se denotarán como *residuos estandarizados estimados* $\hat{e}(\mathbf{x}_i) = r_i/\hat{\sigma}(\mathbf{x}_i)$ (ver paso 2). Esta notación permite establecer claramente la diferencia que existe entre los residuos $\mathbf{r}$ y las estimaciones del término de error $\hat{\varepsilon}$ para el caso del modelo heterocedástico, en comparación con lo que sucede en el modelo bajo tendencia no constante (1.24). Bajo homocedasticidad, ambos términos coinciden (ver p.e. (2.10)), y por esa razón, el variograma del proceso se estima directamente a partir de los residuos (ya que en ese caso los variogramas de los errores y del proceso espacial $Y(\cdot)$ también coinciden). Sin embargo, en el modelo heterocedástico (4.1), $\gamma_\varepsilon(\mathbf{u})$ se aproxima a partir de los residuos estandarizados estimados $\hat{e}(\cdot)$ (ver paso 3). Además, el variograma del proceso y el variograma de los errores son diferentes bajo heterocedasticidad (ver (4.4)).

Cabe indicar que en el contexto paramétrico, en Opsomer *et al.* (1999) también se utiliza un procedimiento para reescalar el variograma estimado a partir de los residuos estandarizados estimados (paso 3). Esto se debe principalmente a que

la varianza de estos residuos no es unitaria, a diferencia de lo que sucede con los errores teóricos $\varepsilon(\cdot)$.

Por otra parte, si bien se pueden utilizar modelos paramétricos en cada uno de los pasos anteriores, no es menos cierto que existen varios aspectos que deben tenerse en cuenta a la hora de su aplicación. De manera similar al caso de procesos homocedásticos, hay que considerar el efecto de la dependencia espacial sobre la estimación de la tendencia, pero adicionalmente este efecto también va a incidir en la estimación de la varianza. Además, el sesgo debido al uso de los residuos $\mathbf{r}$ afectarán a las estimaciones de la variabilidad de pequeña escala, y hay que tener en cuenta el efecto adicional de la estimación de la función varianza sobre el variograma estimado del proceso de error, presentándose nuevamente varios problemas circulares de estimación. A todo lo anterior, se deberían añadir los posibles problemas de mala especificación de modelos paramétricos, que ahora se incrementarían al tener que estimar paramétricamente la nueva componente, que viene dada por la función $\sigma(\cdot)$.

4.2. Estimación conjunta no paramétrica en procesos heterocedásticos.

A continuación se propone un método no paramétrico de estimación conjunta de los componentes del modelo espacial heterocedástico (4.1). Se trata de un proceso iterativo que corrige el sesgo debido al uso de los residuos $\mathbf{r}$, adaptando el método propuesto por Fernández-Casal y Francisco-Fernández (2014) al caso heterocedástico (ver Sección 2.3), para corregir de forma conjunta el variograma y la función varianza en cada iteración. Para obtener las aproximaciones no paramétricas de las componentes $\hat{\mu}(\mathbf{x})$, $\hat{\sigma}(\mathbf{x})$ y $\hat{\gamma}_\varepsilon(\mathbf{u})$, se empleó el suavizado li-

neal mediante el estimador lineal local dado en (2.1), debido a sus propiedades asintóticas tales como la reducción de los efectos frontera, entre otros (Fan y Gijbels, 1996; García-Soidán *et al.*, 2003). A continuación se detalla el método propuesto considerando en primer lugar el caso sin tendencia, que luego se extiende al contexto donde se admite la presencia de una tendencia determinística no constante.

4.2.1. Estimación en procesos heterocedásticos sin tendencia

Supongamos que $\mu(\mathbf{x}) = 0$, para todo $\mathbf{x} \in D$. Por simplicidad, se asumirá que el correlograma de los errores es isotrópico. El algoritmo iterativo propuesto para la estimación no paramétrica conjunta de la función varianza y el variograma de los errores, consta de los siguientes pasos:

Algoritmo 4.1: Estimación NP en procesos heterocedásticos sin tendencia

1. Obtener $\hat{\sigma}^2(\mathbf{x})$ por suavizado lineal de $(\mathbf{x}_i, Y^2(x_i)))$, con una función tipo núcleo K_2 y una matriz ventana $\mathbf{H}_2$;
2. Obtener los valores estimados estandarizados de la variable respuesta, $\hat{e}(\mathbf{x}_i) = Y(\mathbf{x}_i)/\hat{\sigma}(\mathbf{x}_i)$ y luego aproximar el correspondiente variograma $2\hat{\gamma}_{\hat{e}}$ mediante suavizado lineal de $(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{e}(\mathbf{x}_i) - \hat{e}(\mathbf{x}_j))^2)$, con una función tipo núcleo K_3 y una ventana g;
3. Obtener una estimación de la varianza de los valores estandarizados estimados $\hat{\sigma}^2_{\hat{e}}$, y calcular la estimación del variograma del error, reescalando los valores del variograma $\hat{\gamma}_{\hat{e}}(\cdot)$ anterior, tal que $\hat{\gamma}_{\varepsilon}(\mathbf{u}) = \hat{\gamma}_{\hat{e}}(\mathbf{u})/\hat{\sigma}^2_{\hat{e}}$.

Para la selección de las matrices ventana, se pueden considerar distintos criterios, incluyendo alternativas locales y globales. Aunque en el contexto heterocedástico puede ser preferible utilizar ventanas locales, por simplicidad se propone utilizar el criterio *CGCV* (2.9) para seleccionar $\mathbf{H}_2$. Para esto, si se considera

que el proceso espacial es gaussiano[1], y teniendo en cuenta las expresiones (4.3) y (4.5), entonces la matriz de covarianzas del proceso $Y^2(\cdot)$ se puede expresar como:

$$Var(\boldsymbol{Y}^2) = 2\boldsymbol{\Sigma} \odot \boldsymbol{\Sigma} \tag{4.6}$$

donde $\boldsymbol{Y}^2 = (Y^2(\mathbf{x}_1), \ldots, Y^2(\mathbf{x}_n))^t$. En cuanto al parámetro ventana g requerido para la estimación del variograma, este puede ser seleccionado igual que en el caso homocedástico, minimizando el error cuadrático relativo de validación cruzada del variograma estimado (2.15).

4.2.2. Estimación en procesos heterocedásticos con tendencia

Para el caso de una tendencia determinística no nula, su estimación no paramétrica se obtiene a partir del suavizado lineal de los datos $(\mathbf{x}_i, Y(\mathbf{x}_i))$ utilizando el estimador lineal local definido en (2.1), de forma que $\hat{\boldsymbol{\mu}} = \mathbf{SY}$, donde $\mathbf{S}$ es la correspondiente matriz de suavizado. Luego, la varianza y el variograma se pueden estimar a partir de los residuos $\mathbf{r} = \mathbf{Y} - \mathbf{SY}$. Sin embargo, como se mencionó en la Sección 2.3, la varianza de estos residuos $\mathbf{r}$ presenta sesgos respecto a la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$. La relación (2.11) sigue siendo válida en este caso:

$$Var(\mathbf{r}) = \boldsymbol{\Sigma} + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^t - \boldsymbol{\Sigma}\mathbf{S}^t - \mathbf{S}\boldsymbol{\Sigma}. \tag{4.7}$$

[1]En ese caso, si definimos $Y_1 = Y(\mathbf{x}_1)$, $Y_2 = Y(\mathbf{x}_2)$, $\sigma_i = \sigma(\mathbf{x}_i)$ y $\sigma_{12} = Cov[\varepsilon(\mathbf{x}_1), \varepsilon(\mathbf{x}_2)]$, entonces estas variables se pueden reescribir como $Y_1 = \sigma_1 Z_1$ y $Y_2 = aZ_1 + bZ_2$, donde Z_1 y Z_2 son variables normales estándar independientes. Como $\mathbb{E}[Y_1 Y_2] = \sigma_1\sigma_2\sigma_{12} = a\sigma_1$ y $Var[Y_2] = a^2 + b^2$, luego se obtiene: $Y_1 = \sigma_1 Z_1$ y $Y_2 = \sigma_2\sigma_{12}Z_1 + (\sigma_2\sqrt{1-\sigma_{12}^2})Z_2$. Suponiendo normalidad, se tiene que: $\mathbb{E}[Y_1^2 Y_2^2] = 2\sigma_1^2\sigma_2^2\sigma_{12}^2 + \sigma_1^2\sigma_2^2$, y finalmente $Cov[Y_1^2 Y_2^2] = 2\sigma_1^2\sigma_2^2\sigma_{12}^2$.

Para corregir estos sesgos se puede utilizar la siguiente aproximación, obtenida a partir del suavizado lineal de la matriz de varianzas covarianzas heterocedástica (4.5):

$$Var(\mathbf{r}) \approx \boldsymbol{\sigma}\boldsymbol{\sigma}^{\boldsymbol{t}} \odot (\mathbf{R} + \mathbf{B}), \tag{4.8}$$

donde $\mathbf{B} = \mathbf{SRS}^t - \mathbf{RS}^t - \mathbf{SR}$ es la matriz de sesgo debido al uso de los residuos $\mathbf{r}$. Cabe indicar que en el caso homocedástico se tiene la igualdad en la expresión anterior y que esta aproximación es equivalente a la empleada por Ruppert *et al.* (1997) para el caso de datos independientes. A partir de (4.8), se obtiene que:

$$Var\left(r_i/\sqrt{1+b_{ii}}\right) \approx \sigma^2(\mathbf{x}_i).$$

Si se calculan los residuos estandarizados $\tilde{\varepsilon}(\mathbf{x}_i) = r_i/\sigma(\mathbf{x}_i)$, se verifica que:

$$Var\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right) \approx Var\left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right) + b_{ii} + b_{jj} - 2b_{ij},$$

donde b_{ij} representa el (i,j)-ésimo elemento de la matriz $\mathbf{B}$. Esta relación es similar a la obtenida en (2.12) para el caso de modelos homocedásticos.

Luego, suponiendo que la tendencia fue estimada mediante el estimador lineal local (2.1) con una matriz ventana $\mathbf{H}$ y aproximando los residuos estandarizados $\tilde{\varepsilon}(\mathbf{x}_i)$ por su versión estimada $\hat{e}(\mathbf{x}_i)$, se propone realizar la estimación conjunta del modelo (4.1) mediante en el Algoritmo 4.2.

Al finalizar este algoritmo, se obtienen estimaciones no paramétricas de las funciones de tendencia y varianza, así como una estimación piloto del semivariograma $\hat{\gamma}_\varepsilon(u)$. A este variograma piloto se le puede ajustar un modelo paramétrico según el criterio del usuario (por ejemplo, utilizando m.c.p.), o un modelo flexible de Shapiro-Botha (1.23). A partir de estos resultados, se pude obtener un

Algoritmo 4.2: Estimación NP en procesos heterocedásticos con tendencia

1 Calcular los residuos $\mathbf{r} = \mathbf{Y} - \mathbf{SY}$ y obtener una estimación piloto de la matriz de correlación $\mathbf{R}$ (por ejemplo, $\hat{\mathbf{R}} = \mathbf{I}$);

2 Obtener $\hat{\mathbf{B}} = \mathbf{S}\hat{\mathbf{R}}\mathbf{S}^t - \hat{\mathbf{R}}\mathbf{S}^t - \mathbf{S}\hat{\mathbf{R}}$;

3 Calcular la función varianza estimada $\hat{\sigma}^2(\mathbf{x})$ mediante suavizado lineal de $\left(\mathbf{x}_i, r_i^2/(1+\hat{b}_{ii})\right)$, con función tipo núcleo K_2 y una matriz ventana $\mathbf{H}_2$;

4 Calcular los residuos estandarizados estimados $\hat{e}(\mathbf{x}_i) = r_i/\hat{\sigma}(\mathbf{x}_i)$ y aproximar su variograma correspondiente a partir del suavizado lineal (corregido) de $\left(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{e}(\mathbf{x}_i) - \hat{e}(\mathbf{x}_j))^2 - \hat{b}_{ii} - \hat{b}_{jj} + 2\hat{b}_{ij}\right)$, con función tipo núcleo K_3 y ventana g, para obtener $2\hat{\gamma}_{\hat{e}}$;

5 Obtener una estimación de la varianzas de los residuos estandarizados estimados $\hat{\sigma}^2_{\hat{e}}$ a partir del variograma anterior $\hat{\gamma}_{\hat{e}}$ y reescalarlo para obtener el semivariograma estimado de los errores $\hat{\gamma}_{\varepsilon}(\mathbf{u}) = \hat{\gamma}_{\hat{e}}(\mathbf{u})/\hat{\sigma}^2_{\hat{e}}$;

6 Calcular una nueva estimación de la matriz de correlación $\hat{\mathbf{R}}$ a partir de $\hat{\gamma}_{\varepsilon}(u)$ y repetir los pasos 2 al 6 hasta obtener convergencia.

estimador del variograma no estacionario del proceso heterocedástico (4.4):

$$2\hat{\gamma}(\mathbf{x}, \mathbf{x}+\mathbf{u}) = (\hat{\sigma}(\mathbf{x}) - \hat{\sigma}(\mathbf{x}+\mathbf{u}))^2 + 2\hat{\sigma}(\mathbf{x})\hat{\sigma}(\mathbf{x}+\mathbf{u})\hat{\gamma}_{\varepsilon}(\mathbf{u}). \tag{4.9}$$

Cabe indicar que en Ruppert *et al.* (1997) se propuso una corrección similar para la estimación de la función varianza (paso 3) para el caso de datos independientes. Estos autores utilizaron un estimador $\hat{\sigma}^2 = S_2\mathbf{r}^2/(\mathbf{1} + S_2 diag(\mathbf{B}))$, donde S_2 es la correspondiente matriz de suavizado de los residuos cuadrados. Por otra parte, el procedimiento de corrección de sesgo utilizado en el paso 4, es análogo al propuesto por Fernández-Casal y Francisco-Fernández (2014) para la estimación del variograma bajo homocedasticidad.

Adicionalmente, se puede proponer un algoritmo alternativo, al considerar que

la relación (4.7) puede ser reescrita como:

$$Var(\mathbf{r}) = \mathbf{\Sigma} + \mathbf{B}_C,$$

donde $\mathbf{B}_C = \mathbf{S\Sigma S}^t - \mathbf{\Sigma S}^t - \mathbf{S\Sigma}$. En ese caso, la función varianza se puede estimar mediante el suavizado lineal de $\left(\mathbf{x}_i, r_i^2 - \hat{b}_{ii}^C\right)$ en el paso 3. Estudios numéricos preliminares mostraron que ambos procedimientos obtienen resultados similares, aunque este segundo algoritmo utiliza mayor tiempo de computación, pues en cada iteración se requiere calcular las dos matrices de sesgo $\mathbf{B}$ y $\mathbf{B}_C$. Se puede obtener una ligera mejora computacional si se considera la aproximación: $\hat{\mathbf{B}}_C \approx \hat{\boldsymbol{\sigma}}\hat{\boldsymbol{\sigma}}^t \odot \hat{\mathbf{B}}$.

En cuanto a la selección de las ventanas, para la estimación de la tendencia se propone utilizar el criterio $CGCV$ para seleccionar $\mathbf{H}$, con el fin de tener en cuenta el efecto de la dependencia espacial (a través de la matriz de correlación $\mathbf{R}$). El mismo criterio se puede utilizar para seleccionar $\mathbf{H}_2$ para el suavizado lineal de los residuos cuadrados (paso 3), utilizando en este caso la siguiente aproximación:

$$Var(\boldsymbol{r}^2) \approx 2(Var(\boldsymbol{r})) \odot (Var(\boldsymbol{r})). \tag{4.10}$$

Sin embargo, el uso de este criterio va a depender de la estimación de la matriz de varianzas y covarianzas heterocedástica $\mathbf{\Sigma}$ y de la matriz de sesgo $\mathbf{B}$. Para evitar este problema circular, se puede recurrir a un método iterativo para la selección de dicha ventana, similar al propuesto en la Sección 2.3 para el caso homocedástico. En la Sección 4.4 se propone un método de dos pasos para la selección de dichas ventanas.

4.3. Estudios de simulación

En esta sección, se analiza el comportamiento de los estimadores presentados en las secciones anteriores mediante varios estudios de simulación. Los datos se generaron inicialmente considerando una rejilla regular bidimensional en el soporte $D = [0,1]^2$, bajo el modelo (4.1). Para analizar el efecto de la forma de la tendencia sobre las estimaciones, se han considerado tres tipos de funciones: $\mu_1(x_1, x_2) = 0$ (tendencia nula), $\mu_2(x_1, x_2) = 5{,}8(x_1 - x_2 + x_2^2)$ (tendencia polinómica) y $\mu_3(x_1, x_2) = \sin(2\pi x_1) + 4(x_2 - 0{,}5)^2$ (tendencia no polinómica). Asimismo se seleccionaron tres modelos para la función de varianza: $\sigma_1^2(x_1, x_2) = 1$ (varianza constante), $\sigma_2^2(x_1, x_2) = 0{,}5(1 + x_1 - x_2)$ (varianza lineal) y $\sigma_3^2(x_1, x_2) = \left(\frac{15}{16}\right)^2 (1 - (2x_1 - 1)^2)^2(1 - (2x_2 - 1)^2)^2 + 0{,}1$ (varianza no lineal).

Se generaron los errores aleatorios con una distribución normal multivariante, de media cero, varianza unidad y variograma isotrópico exponencial (1.17), con los siguientes paramétros: $a = 0{,}3$, $0{,}6$ y $0{,}9$, $c_0 = 0$, $0{,}2$, $0{,}4$, y $0{,}8$. Para cada escenario, se simularon $N = 1,000$ muestras de tamaño n igual a 10×10, 15×15 y 20×20 (de forma similar a los valores utilizados en la sección 2.5).

Para evitar el efecto de la selección de ventana en los resultados, se calculó la matriz de varianzas y covarianzas teórica dada por (4.5), y a partir de esta se seleccionaron las ventanas minimizando el correspondiente criterio MASE. Para las tendencias no nulas ($\mu_2(\cdot)$ y $\mu_3(\cdot)$), la ventana $\mathbf{H}$ se seleccionó utilizando el criterio dado por (2.6), mientras que para la selección de la ventana $\mathbf{H}_2$ se utilizó el mismo criterio considerando la aproximación (4.10). Cuando se considera la tendencia nula $\mu_1(\cdot)$, se utilizó la expresión (4.6) en el criterio $MASE$ a la hora escoger la ventana $\mathbf{H}_2$ correspondiente. La ventana g para la estimación lineal local del variograma se seleccionó minimizando el error cuadrático relativo por validación cruzada.

Para comparar el comportamiento de los algoritmos propuestos, se calcularon los errores cuadráticos:

$$SE(\hat{\sigma}(\mathbf{x})) = (\hat{\sigma}(\mathbf{x}) - \sigma(\mathbf{x}))^2,$$

$$SE(\hat{\gamma}(u)) = (\hat{\gamma}(u) - \gamma(u))^2.$$

En primer lugar, consideramos el modelo de tendencia nula $\mu_1(\cdot)$, para de esa manera analizar el comportamiento de los estimadores propuestos de la varianza y variograma sin la presencia del sesgo debido a la estimación de dicha tendencia. Utilizando el algoritmo propuesto 4.1, se verifica que ambos estimadores proporcionan buenas aproximaciones a sus valores teóricos. Un resumen de los errores en este caso se presentan en la Tabla 4.1, para $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$ y $a = 0{,}6$. Estos valores pueden servir de referencia en comparación con las estimaciones obtenidas bajo la presencia de una tendencia no constante.

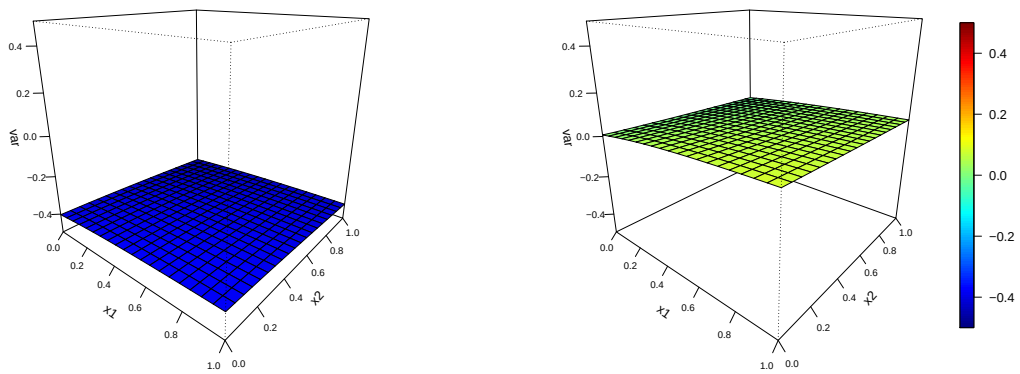
Tabla 4.1: Resúmenes de errores cuadrados de los estimadores de la varianza y el variograma, para datos sin tendencia, varianza lineal $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$ y $a = 0{,}6$.

Estimador	Varianza			Variograma		
n	10×10	15×15	20×20	10×10	15×15	20×20
Media	0.172	0.160	0.155	0.009	0.010	0.010
Mediana	0.055	0.050	0.049	0.003	0.003	0.003
Desv. Est.	0.458	0.440	0.439	0.017	0.017	0.017

Para el caso general, considerando las tendencias $\mu_2(\cdot)$ y $\mu_3(\cdot)$, el método propuesto (al que denominaremos "corregido" en los resultados) se comparó con la estimación obtenida mediante el uso directo de los residuos (denotada por "residual"), donde la varianza se aproxima utilizando el suavizado lineal de los residuos cuadrados (sin corregir).

Respecto a la estimación de la función varianza, en general el método residual produce una subestimación de la varianza teórica, mientras que el estimador

corregido se muestra aproximadamente insesgado. Por ejemplo, las Figuras 4.2(a) y 4.2(b) representa las superficies del sesgo promedio para ambos estimadores, considerando $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $a = 0{,}6$ y $c_0 = 0{,}2$.



(a) (b)

Figura 4.2: Promedio de los sesgos obtenidos para los estimadores residuales (a) y corregido (b) de la función varianza , con $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $a = 0{,}6$ y $c_0 = 0{,}2$.

Asimismo, considerando los distintos escenarios de simulación, las estimaciones corregidas de la función varianza proporcionaron errores inferiores que las estimaciones residuales. Esto se puede observar, por ejemplo, en el caso de $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$, $a = 0{,}6$, cuyos valores se representan en la Tabla 4.2 para distintos tamaños muestrales. Aqui se puede ver que en todos los casos, el método corregido presenta menores promedios de error cuadráticos, y que estos errores van disminuyendo conforme el tamaño muestral aumenta, lo que sugiere la consistencia del método propuesto.

En las Tablas 4.3 y 4.4 se presenta la misma información anterior, considerando distintos valores de efecto nugget y rango, para un tamaño fijo de $n = 20 \times 20$. Estos resultados nuevamente evidencian el buen comportamiento del estimador corregido, en especial cuando la dependencia espacial es alta (valores bajos del

nugget o altos del rango). Además, se puede observar que los estimadores residuales presentan errores mayores cuando los datos presentan una correlación espacial fuerte. Si la dependencia espacial es débil se puede observar que los resultados con ambos estimadores resultan mas similares entre sí.

Tabla 4.2: Resúmenes de los errores cuadráticos para los estimadores de la función varianza, para $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$, $a = 0{,}6$ y diferentes tamaños muestrales.

Estimador	Residual			Corregido		
n	10×10	15×15	20×20	10×10	15×15	20×20
Media	0.212	0.193	0.183	0.152	0.090	0.085
Mediana	0.175	0.164	0.159	0.042	0.029	0.026
Desv. Est.	0.165	0.135	0.119	0.486	0.236	0.389

Tabla 4.3: Resúmenes de los errores cuadráticos para los estimadores de la función varianza, para $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $a = 0{,}6$ y diferentes valores de nugget.

c_0	0		0.4		0.8	
Estimador	Residual	Corregido	Residual	Corregido	Residual	Corregido
Media	0.259	0.149	0.120	0.053	0.035	0.034
Mediana	0.227	0.042	0.102	0.017	0.022	0.010
Desv. Est.	0.165	0.624	0.084	0.216	0.040	0.073

Tabla 4.4: Resúmenes de los errores cuadráticos para los estimadores de la función varianza, para $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$ y distintos valores de rango.

a	0.3		0.6		0.9	
Estimador	Residual	Corregido	Residual	Corregido	Residual	Corregido
Media	0.073	0.107	0.183	0.085	0.280	0.079
Mediana	0.055	0.028	0.159	0.026	0.254	0.039
Desv. Est.	0.069	0.401	0.119	0.389	0.151	0.234

Por otra parte, el estimador corregido del variograma $\gamma_\varepsilon(\cdot)$ se comparó con el correspondiente variograma sin corregir (obtenido a partir de los residuos estandarizados estimados $\hat{e}(\mathbf{x})$, utilizando el estimador "residual" de la varianza). Los resultados obtenidos muestran en general que el estimador corregido proporciona buenas aproximaciones al modelo teórico, mientras que el estimador residual

tiende a subestimar la dependencia espacial. Por ejemplo, la Figura 4.3 muestra al semivariograma teórico y los promedios de los estimadores no paramétricos del variograma para $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $a = 0{,}6$ y $c_0 = 0{,}2$. En este gráfico se observan claramente grandes diferencias entre ambos métodos, especialmente en la estimación del efecto nugget.

Las conclusiones obtenidas es la estimación del variograma para las distintas configuraciones de simulación, se mantienen al comparar los estadísticos de error de la función varianza, aunque en este caso se puede observar un menor efecto del tamaño muestral sobre la precisión de las estimaciones (ver p.e. Tabla 4.5). Esto puede ser debido a que en ambos métodos, los variogramas estimados se reescalan (de manera que el umbral sea igual a 1).

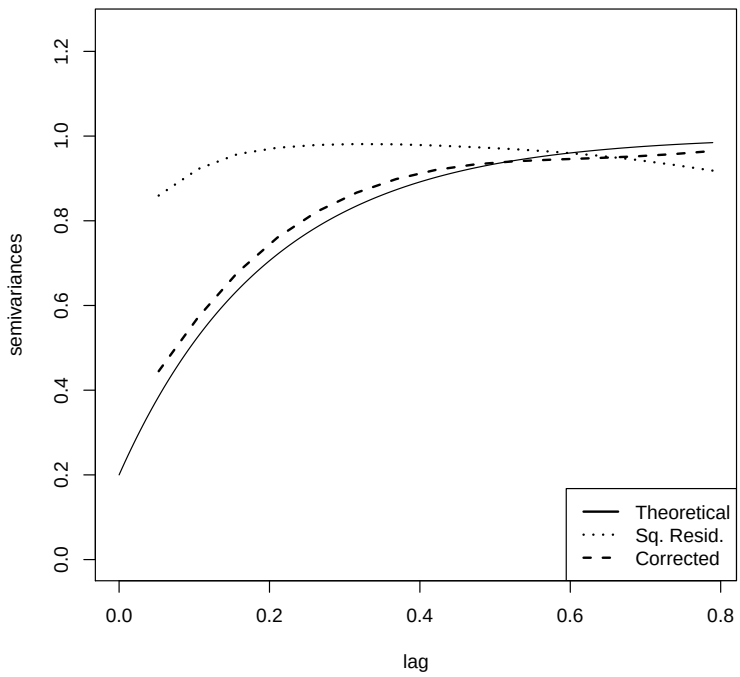


Figura 4.3: Semivariograma teórico de los errores (línea continua), y promedios de los semivariogramas estimados residuales (línea de puntos) y corregidos (línea discontinua), para $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $a = 0{,}6$ y $c_0 = 0{,}2$.

Tabla 4.5: Resúmenes de errores cuadráticos de los estimadores del variograma, para $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$, $a = 0{,}6$ y distintos tamaños muestrales.

Estimador	Residual			Corregido		
n	10×10	15×15	20×20	10×10	15×15	20×20
Media	0.035	0.041	0.040	0.007	0.006	0.006
Mediana	0.008	0.009	0.009	0.002	0.002	0.002
Desv. Est.	0.057	0.066	0.069	0.012	0.010	0.010

De manera general, los distintos estudios numéricos llevados a cabo muestran el mejor comportamiento del método propuesto, pues proporciona mejores estimaciones conjuntas de la varianza y el semivariograma en relación al método tradicional, con todas las tendencias y varianzas teóricas consideradas. Por ejemplo en la Tabla 4.6 se presentan los promedios de los errores cuadráticos para ambos métodos, considerando la función no polinómica $\mu_3(\cdot)$ como la tendencia teórica.

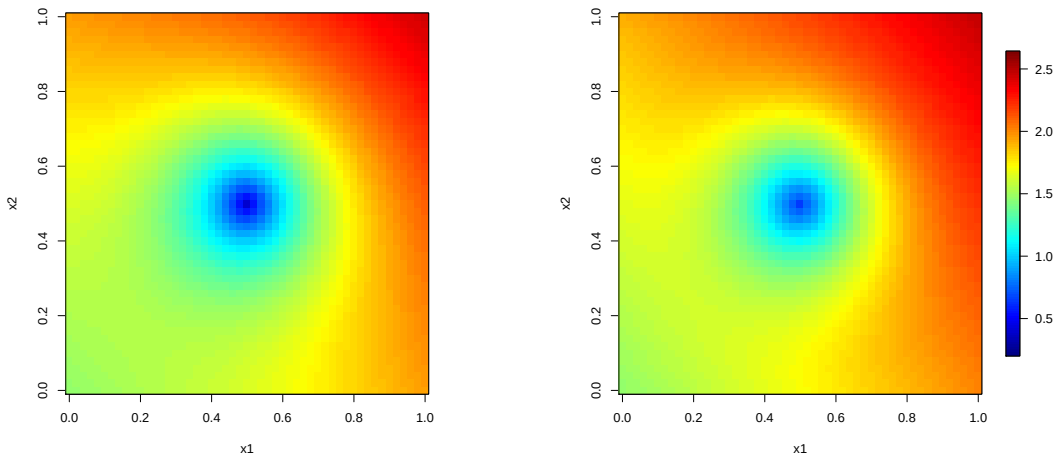
Tabla 4.6: Promedio de errores cuadráticos para los estimadores de la función varianza y variograma, para $\mu_3(\cdot)$, $n = 20 \times 20$, $c_0 = 0{,}2$, $a = 0{,}6$, con distintas varianzas teóricas.

Estimador	Varianza		Variograma	
Método	Residual	Corregido	Residual	Corregido
$\sigma_1^2(\cdot)$ (constante)	0.175	0.073	0.044	0.006
$\sigma_2^2(\cdot)$ (lineal)	0.183	0.085	0.044	0.006
$\sigma_3^2(\cdot)$ (no-lineal)	0.040	0.023	0.053	0.005

Para analizar el comportamiento del método propuesto bajo diseño aleatorio, se consideraron muestras simuladas con posiciones espaciales generadas en posiciones irregularmente espaciadas, siguiendo una distribución uniforme bidimensional sobre el cuadrado unitario. Un ejemplo de los resultados obtenidos para este caso se muestra en la Tabla 4.7. Se puede observar que los errores cuadráticos estimados para la varianza y el variograma en el caso irregular, son similares a los obtenidos bajo diseño fijo (ver p.e. las Tablas 4.2 y 4.5) y por tanto, las conclusiones se mantienen válidas en ambas situaciones.

Tabla 4.7: Resúmenes de errores cuadráticos para los estimadores de la varianza y el variograma, para $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $n = 20 \times 20$, $c_0 = 0{,}2$ y $a = 0{,}6$, bajo muestreo irregular (con distribución uniforme sobre el cuadrado unidad).

Estimator	Varianza		Variograma	
Método	Residual	Corregido	Residual	Corregido
Media	0.183	0.123	0.041	0.009
Mediana	0.162	0.030	0.010	0.003
Desv. Est.	0.120	0.932	0.062	0.016



(a) (b)

Figura 4.4: (a) Semivariograma no estacionario heterocedástico teórico $\gamma(\mathbf{x_0}, \mathbf{x})$ y (b) promedio de semivariograma no estacionario heterocedástico estimado con el método propuesto de corrección, en $\mathbf{x_0} = (0{,}5, 0{,}5)$, para $n = 20 \times 20$, $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $c_0 = 0{,}2$ y $a = 0{,}6$.

Finalmente, se utilizaron los estimadores anteriores para construir las correspondientes estimaciones del variograma no estacionario del proceso heterocedástico, definido en (4.9). El buen comportamiento de las estimaciones corregidas se puede observar en las Figuras 4.4(a) y 4.4(b), donde los promedios de la estimación $\hat{\gamma}(\mathbf{x_0}, \mathbf{x})$, en $\mathbf{x_0} = (0{,}5, 0{,}5)$ son similares a los obtenidos para la función teórica (4.4). Considerando al error cuadrático de las estimaciones del variograma no estacionario como una medida global de precisión de los métodos, se construyeron los resultados que se presentan en la Tabla 4.8, con $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $n = 20 \times 20$,

$c_0 = 0{,}2$ y $a = 0{,}6$. Esta tabla pone nuevamente de manifiesto el mejor comportamiento del método propuesto respecto al método residual, tanto en el caso de muestreo regular como irregular.

Tabla 4.8: Resúmenes de los errores cuadráticos para la estimación del variograma heterocedástico del proceso $\hat{\gamma}(\mathbf{x_0}, \mathbf{x_0} + \mathbf{u})$ con $\mathbf{x_0} = (0{,}5, 0{,}5)$, para $\mu_3(\cdot)$, $\sigma_2^2(\cdot)$, $n = 20 \times 20$, $c_0 = 0{,}2$ y $a = 0{,}6$, considerando muestreo regular e irregular

Muestreo	Regular		Irregular	
Método	Residual	Corregido	Residual	Corregido
Media	0.356	0.010	0.370	0.013
Mediana	0.345	0.010	0.360	0.013
Desv. Est.	0.222	0.007	0.230	0.009

4.4. Aplicación a datos reales

En esta sección se describe una aplicación práctica del método propuesto en el presente capítulo. Para esto, se ha considerado el mismo conjunto de datos utilizado en la Sección 3.5, en la cual se representan las precipitaciones totales (en pulgadas de lluvias y posteriormente transformados mediante raíz cuadrada), registradas sobre 1053 localizaciones situadas en la parte continental de Estados Unidos de Norteamérica. Sobre este conjunto de datos y suponiendo heterocedásticidad, se aplicaron tanto el método tradicional basado en residuos (descrito en la Sección 4.1.2) como el método no paramétrico propuesto.

Para estimar las componentes del modelo (4.1) se realizó en procedimiento en dos etapas. En primer lugar, se obtuvo una estimación preliminar de la tendencia y el variograma suponiendo homocedasticidad (mediante 2 iteraciones del Algoritmo 2.2, de forma análoga a lo efectuado en la Sección 3.5). En un segundo paso, usando la matriz de varianzas y covarianzas derivada del paso anterior, se aplicó el criterio $CGCV$ para obtener la matriz ventana $\mathbf{H}$ para la estimación de la tendencia y posteriormente aplicar el Algoritmo 4.2. En lugar de repetir este

proceso iterativamente, se decidió mantener las matrices ventana seleccionadas, de manera que ambos métodos (residual y corregido) utilicen la misma ventana **H** en la estimación de la tendencia, evitando de esta forma la influencia que podría tener en los resultados el hecho de considerar distintas estimaciones de la tendencia. Asimismo, esta tendencia estimada fue utilizada en el método tradicional para aproximar la función varianza (por suavizado lineal de los residuos cuadrados respectivos). Por otra parte, la ventana g del variograma se obtuvo minimizando el correspondiente error cuadrático relativo de validación cruzada (el mismo que fue empleado en los casos anteriores).

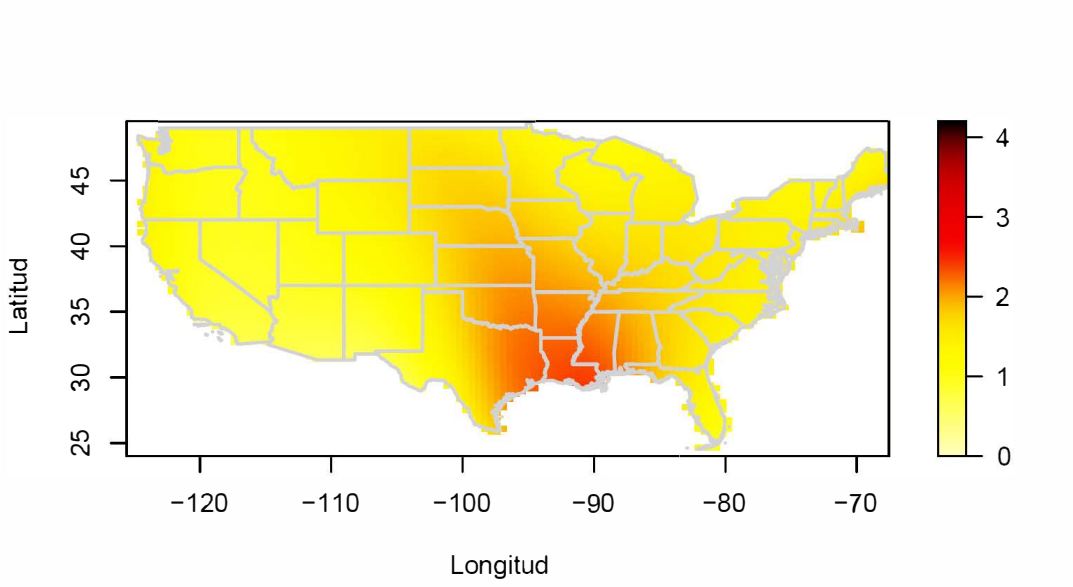
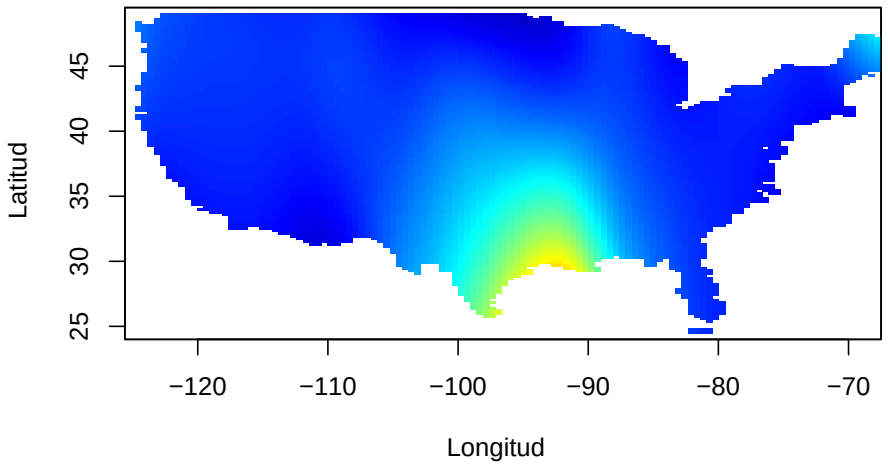


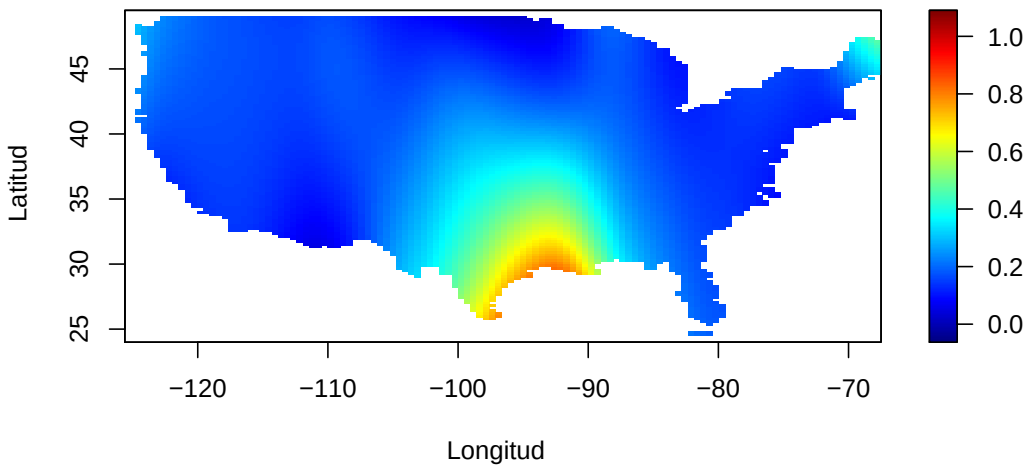
Figura 4.5: Estimación final de la tendencia asumiendo heterocedasticidad.

La Figura 4.5 muestra la tendencia final estimada, utilizando una matriz ventana $\mathbf{H} = diag(10{,}0{,}18{,}2)$. Cabe mencionar que la ventana obtenida mediante este procedimiento resulta casi idéntica a la obtenida en la Sección 3.5 (ver p.e. Figura 3.8). Las estimaciones finales de la función varianza se pueden observar en las Figuras 4.6(a) y 4.6(b) respectivamente. La ventana $\mathbf{H}_2 = diag(5{,}1, 15{,}4)$ se utilizó en ambos casos para evitar distorsiones. Como cabría esperar, las es-

timaciones no corregidas aparentemente subestiman la variabilidad del proceso espacial.



(a)



(b)

Figura 4.6: Varianzas estimadas obtenidas con el (a) método de residuos cuadrados, y (b) el método corregido.

Estos resultados muestran que existen áreas geográficas, en especial al sur

de Estados Unidos (Texas y Lousiana), donde tanto la tendencia como la varianza presentan valores muy altos, en comparación con las otras zonas. Estas estimaciones son coherentes con el comportamiento de los datos observados y con los comentarios realizados al final del capítulo anterior, respecto a los mapas de riesgo no paramétrico, construidos a partir de este mismo conjunto de datos. Sin embargo, en este caso se observa que la variabilidad del proceso es aparentemente mayor en las zonas con mayores precipitaciones.

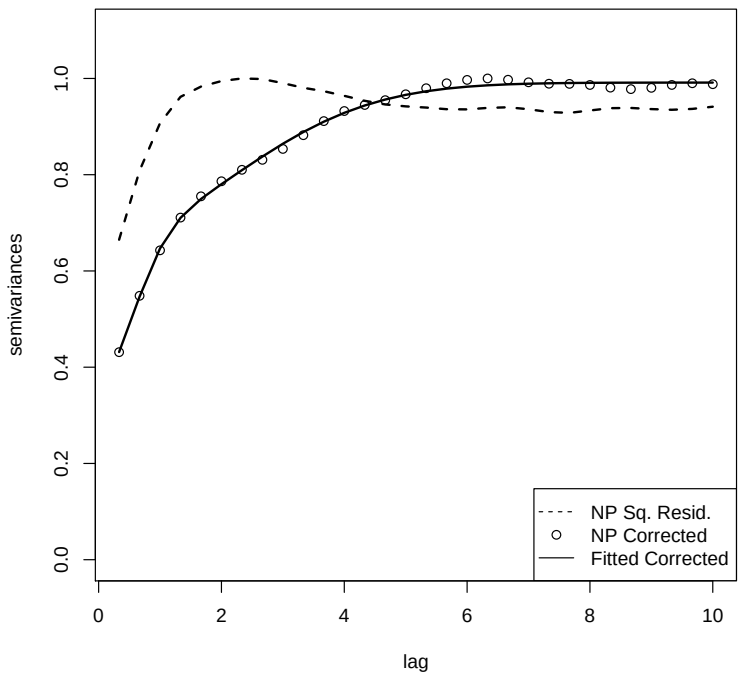


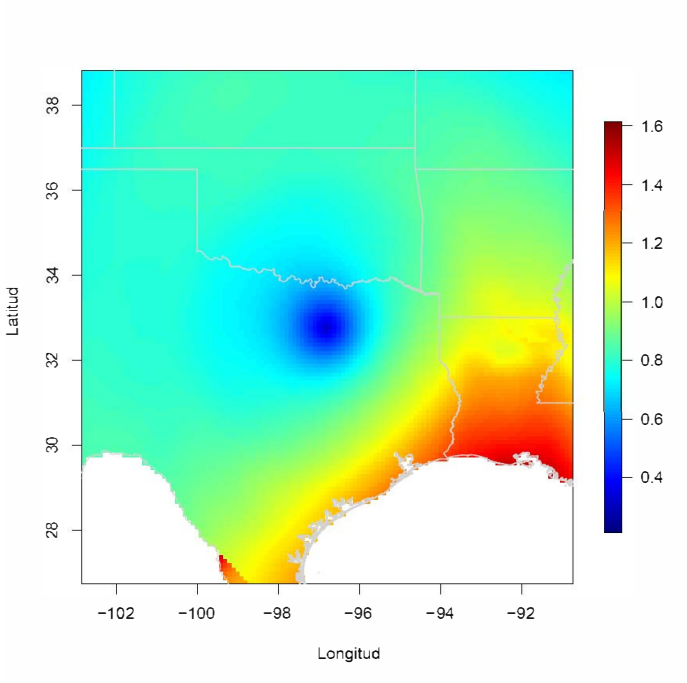
Figura 4.7: Estimaciones no paramétricas del variograma de error obtenidas mediante el método tradicional (línea discontinua) y con el método propuesto (lína de puntos). La línea continua corresponde al modelo de Shapiro-Botha ajustado a las estimaciones corregidas.

Respecto a la estimación del semivariograma, la Figura 4.7 presenta las estimaciones no paramétricas obtenidas mediante el estimador residual (línea discontinua) y por el método corregido (línea de círculos). Estas estimaciones son consistentes con los resultados obtenidos en los estudios de simulación, donde

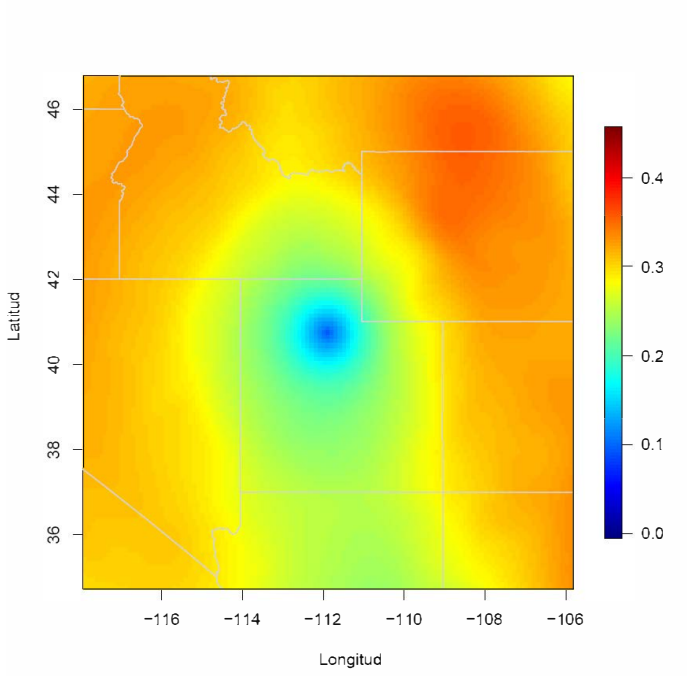
se observó que el estimador no corregido no aproxima de forma adecuada la dependencia espacial. Cabe recordar que las estimaciones del semivariograma obtenidas a partir del método de corrección propuesto son estimaciones piloto, en el sentido que no necesariamente cumplen la condición de ser condicionalmente definidas negativas. Queda a elección del usuario seleccionar el modelo y método más apropiado para ajustar un semivariograma válido. Nosotros proponemos utilizar el correspondiente modelo de Shapiro Botha ajustado (línea contina).

Como ilustración final del método propuesto, se construyeron estimaciones $\hat{\gamma}(\mathbf{x_0}, \mathbf{x})$ del variograma no estacionario del proceso heterocedástico definido por (4.4), combinando las estimaciones de la varianza y del modelo de variograma ajustado y sustituyéndolas en la ecuación (4.9). En este ejemplo, se fijaron como punto central $\mathbf{x_0}$ las coordenadas geográficas de las ciudades de Dallas (Texas) y Salt Lake City (Utah). Se puede comprobar que el semivariograma heterocedástico centrado en Dallas (Figura 4.8(a)) tiene valores más altos comparados con la estimación correspondiente en Salt Lake City (Figura 4.8(b)). Estas diferencias se pueden explicar debido a las distintas estimaciones de la varianza en localizaciones vecinas a dichos puntos centrales, además que ambas ciudades se encuentran en zonas con precipicitaciones altas y bajas respectivamente, de acuerdo con el registro obtenido para el mes considerado.

Estos gráficos constituyen solo un ejemplo sencillo de las posibles aplicaciones que se pueden derivar a partir de las estimaciones no paramétricas obtenidas mediante el algoritmo propuesto. La extensión de este método permitiría además construir predicciones kriging, o mapas de riesgo heterocedásticos, entre otras aplicaciones importantes. Incluso se podría recurrir a este algoritmo para el análisis de procesos espacio temporales, en los cuales la variabilidad temporal sea recogida también a través de la función varianza.



(a)



(b)

Figura 4.8: Estimación corregida del variograma heterocedástico $\hat{\gamma}(\mathbf{x_0}, \mathbf{x})$, fijando $\mathbf{x_0}$ como las coordenadas geográficas de (a) Dallas (Texas), y (b) Salt Lake City (Utah).

Conclusiones

En el presente trabajo se han propuesto distintos procedimientos no paramétricos para su aplicación a datos espaciales. Para ello, en el primer capítulo se introdujeron las definiciones y metodología utilizadas en el caso estacionario, haciendo especial referencia al análisis estructural para la obtención de modelos de estimadores válidos del variograma. También se describen los modelos de Shapiro-Botha que se utilizaron posteriormente en diversos apartados de este trabajo. Al final de este capítulo, nos centramos en el estudio de procesos con tendencia no constante, en especial en el método de estimación basado en residuos. Se evidenció el problema circular existente entre la estimación de la tendencia y la dependencia, junto con el algoritmo iterativo de Neuman y Jacobson (1984) que se suele utilizar para solventar este problema. Sin embargo, este procedimiento no tiene en cuenta el sesgo existente en la estimación del variograma debido al uso directo de los residuos, y su enfoque paramétrico presenta diversas limitaciones, en especial debido a la posible mala especificación de los modelos seleccionados para la tendencia o el variograma.

Considerando un enfoque no paramétrico, en el segundo capítulo se propone utilizar el estimador lineal local de la tendencia. Sin embargo, la precisión de las estimaciones obtenidas de esta manera dependen de la adecuada selección de la matriz ventana. Para el caso de datos espaciales, los criterios de selección de ventana debe tener presente la estructura de dependencia de los datos, dando lugar

nuevamente al problema circular mencionado anteriormente. Para solventar este inconveniente, en la Sección 2.3 se propone en primer lugar, una ligera modificación del procedimiento de corrección de sesgo del variograma estimado a partir de residuos, propuesto por Fernández-Casal y Francisco-Fernández (2014). Este Algoritmo 2.1, utiliza pseudocovarianzas construidas a partir del estimador piloto del variograma, sin necesidad de recurrir al ajuste de modelos en cada iteración, reduciendo el tiempo de computación (hasta en un 84 % para el conjunto de datos utilizado en las Secciones 3.5 y 4.4). Posteriormente, se propone el Algoritmo iterativo 2.2 para la estimación conjunta no paramétrica de la tendencia y del variograma, el cual proporciona estimaciones de la tendencia teniendo en cuenta la estimación corregida del variograma (mediante el algoritmo anterior), evitando de esta manera el efecto del sesgo debido al uso de residuos.

Respecto a la selección de la ventana para la estimación de la tendencia, en la Sección 2.4 se proponen nuevos criterios para obtener la ventana óptima bajo dependencia, a partir de una aproximación para la esperanza matemática de los criterios de validación cruzada. Estos nuevos criterios corregidos CCV y $CMCV$ tratan de aproximar el criterio $MASE$ incluyendo un término que depende de la matriz de varianzas y covarianzas de los datos (omitiendo observaciones para un vecindario predeterminado). El comportamiento de estos nuevos selectores de ventana, conjuntamente con los criterios tradicionales se analizaron mediante estudios numéricos y una aplicación a datos reales. Estos análisis indican de manera general que los criterios $CGCV$ y CCV son más eficientes bajo correlación espacial alta o moderada, en comparación con selectores de ventana que no tienen en cuenta la estructura de dependencia. Los resultados también ponen en evidencia el buen comportamiento de los algoritmos de correción de sesgo y de estimación conjunta, pues se observa que utilizando las matrices de varianzas y covarianzas corregidas, los promedio de los errores cuadráticos de la estimación

de la tendencia disminuyen.

La contribución principal del tercer capítulo se centra en el desarrollo de un método bootstrap no paramétrico NPB, especialmente diseñado para reproducir la variabilidad en procesos espacial con tendencia no constante. Este método se compara con otros procedimientos utilizados en el caso espacial, como el bootstrap por bloques BB o el semiparamétrico SPB. Tanto en el caso estacionario, como bajo la presencia de una tendencia espacial determinística, el método propuesto presenta varias ventajas: no depende de la región espacial como el método BB y tampoco se encuentra expuesto a problemas de mala especificación de modelos, como ocurre con el SPB. Sin embargo, la principal ventaja del método propuesto es que permite obtener réplicas bootstrap teniendo en cuenta el efecto del sesgo debido al uso directo de los residuos, pues considera tanto la matriz de varianzas y covarianzas residuales como su versión corregida, ambas aproximadas utilizando los algoritmos descritos en el capítulo anterior.

Se llevaron a cabo diversos estudios de simulación para aproximar mediante bootstrap el sesgo y la varianza del estimador empírico y lineal local del variograma, considerando distintos escenarios de dependencia espacial, diversos modelos de tendencia, bajo muestreo regular e irregular. Los resultados obtenidos muestran que el comportamiento del método NPB es superior al BB en procesos estacionarios. Con el método SPB se obtienen resultados similares, aunque también se observa que este método es sensible a problemas de mala especificación del variograma. Los resultados bajo este escenario dan una clara ventaja en términos de error cuadrático al método no paramétrico propuesto. Cuando se admite la presencia de una tendencia determinística, el efecto del sesgo debido a los residuos afecta claramente a los resultados del SPB, mientras que el NPB reproduce mucho mejor el comportamiento teórico para ambos estimadores del variograma, aunque las aproximaciones boostrap son mejores al utilizar el estimador lineal

local del variograma.

El método NPB puede resultar adecuado para realizar inferencia sobre el proceso espacial, por ejemplo para construir intervalos de confianza (como se presenta en la Sección 3.5), o en contrastes de hipótesis sobre el variograma. Una adaptación de este método, para la construcción de mapas de riesgo geoestadístico se realizó en la Sección 3.4.1. El Algoritmo 3.3 propuesto permite estimar la probabilidad (incondicional) de que una variable supere un valor crítico para cada localización de la región de observación. Este procedimiento es una modificación al método semiparamétrico propuesto por Francisco-Fernández *et al.* (2011) para el estudio de terremotos. Los diversos estudios de simulación realizados al respecto, así como su aplicación a un conjunto de datos reales relacionados al total de precipitaciones en EEUU, ilustran el buen comportamiento de este método. Esto se debe a que el método propuesto tiene en cuenta la variabilidad corregida, reduciendo en 40 % los errores cuadráticos obtenidos al utilizar los residuos sin corrección. Asimismo, el nuevo algoritmo proporciona promedios de errores que son aproximadamente un 55 % más bajos que los obtenidos por el método semiparamétrico original.

En el cuarto capítulo se estudian los procesos espaciales heterocedásticos y se propone un método de estimación conjunta de las componentes de dicho modelo con un enfoque totalmente no paramétrico empleando el estimador lineal local. Nuevamente se analiza el efecto que tiene el sesgo debido al uso de los residuos, en este caso, tanto en la estimación del variograma como de la función varianza. Además, se extendió al caso heterocerocedástico el procedimiento para la corrección del variograma descrito en el Capítulo 2 (donde se asumía homocedasticidad). El método propuesto corrige iterativamente de forma conjunta los residuos cuadrados utilizados para aproximar la varianza, así como los sesgos de las semivarianzas de los residuos estandarizados para la estimación del variogra-

ma. Este enfoque permite obtener estimaciones no paramétricas de la tendencia y de la función varianza, conjuntamente con estimaciones piloto del variograma del error.

El método propuesto fue estudiado numéricamente considerando distintos grados de dependencia espacial, así como diversos modelos para las funciones de tendencia y varianza teóricas, bajo diseño regular y aleatorio. Su comportamiento fueron comparado con el correspondiente al método tradicional de estimación basados en residuos sin corregir, a la hora de aproximar las características del proceso heterocedástico. Los resultados de la Sección 4.3 muestran que tanto las estimaciones de la función varianza, como de los variogramas del error (y por tanto la estimación del variograma no estacionario del proceso) obtenidas mediante el método propuesto, son más eficientes que las obtenidas con el método tradicional. Este algoritmo proporciona menores errores cuadráticos cuando el grado de dependencia espacial es moderado o alto (es decir, cuando el nugget es pequeño el rango es grande).

Además se demostró la aplicabilidad en la práctica del método propuesto, utilizando nuevamente el conjunto de datos de las precipitaciones totales mensuales en EEUU, obteniéndose resultados coherentes con los anteriores (presentados en la Sección 3.5). Sin embargo, con este procedimiento se observa que aparentemente la variabilidad del proceso es mayor en las zonas con mayores precipitaciones.

Es importante indicar que los métodos propuestos en esta memoria permiten obtener estimaciones pilotos de las componentes del modelo estacionario o heterocedástico, según sea el caso. Queda a criterio del usuario seleccionar los modelos que considere más adecuados, para su posterior ajuste a las estimaciones obtenidas mediante los métodos propuestos. Otra ventaja de este enfoque no paramétrico es que no requiere la intervención del usuario para realizar procedimientos de ajuste de modelos, lo cual facilita su automatización.

A partir del estudio de los distintos algoritmos propuestos a lo largo del presente estudio y su extensión al caso heterocedástico, se pueden plantear nuevas líneas de trabajo futuro. Respecto a la selección de la ventana para la estimación de la tendencia bajo dependencia espacial, se podrían estudiar nuevos criterios, como por ejemplo, la ventana por bootstrap suavizado, que se podría construir utilizando los algoritmos de estimación NP para obtener matrices de varianzas y covarianzas estimadas corregidas, las cuales se aplicarían directamente en lugar de las correspondientes matrices teóricas en el criterio $MASE$. Otro aspecto importante de investigación, particularmente para el caso de procesos heterocedásticos, sería el uso de criterios locales para la selección de ventanas para la tendencia y el variograma, de modo que dichas ventanas dependiesen de la variabilidad en el punto de estimación.

Por otra parte, un objetivo de trabajo futuro será poner de manifiesto la utilidad del método NPB para la construcción de intervalos y bandas de confianza, e incluso efectuar contrastes de hipótesis sobre la estructura de dependencia del proceso espacial. Asimismo, actualmente se está trabajando en la extensión de este método al caso heterocedástico, cuyos resultados serán presentados próximamente.

Respecto a la construcción de mapas de riesgo, se podría considerar la extensión del algoritmo propuesto al caso condicional, basado en los métodos de simulación condicional (los cuales combinan la simulación incondicional obtenida mediante el NPB y la predicción kriging). Estos métodos se podrían comparar con otras técnicas utilizadas con el mismo fin, como el kriging indicador o el simplicial indicator kriging, entre otros.

El método de estimación del Capítulo 4 también se podría utilizar para realizar inferencias sobre los procesos espaciales heterocedásticos, como por ejemplo, en predicción kriging, mapas de riesgo, entre otras aplicaciones. Los modelos he-

terocedásticos pueden ser de especial utilidad en el caso espacio temporal. Por ejemplo, puede ser razonable suponer que la variabilidad del proceso es función de la componente temporal. En ese caso, para el conjunto de datos considerado es posible obtener estimaciones de la variabilidad de pequeña escala de las precipitaciones en un mes determinado, utilizando las mediciones de meses anteriores para estimar la función varianza.

Finalmente, es necesario mencionar que todos los procedimientos no paramétricos utilizados a lo largo del presente estudio, se aplicaron en los estudios numéricos, con datos reales y simulados, utilizando el software *R*. Particularmente se hizo uso del paquete npsp de Fernández-Casal (2014), en el que ya fueron implementadas las funciones necesarias para la aplicación de los criterios de selección de ventana, *h.cv* y *hcv.data* para el caso de matrices ventana diagonales (en una nueva versión del paquete se incluirá la opción para distintas configuraciones de ventana), así como el algoritmo 2.1 de corrección N.P. (función *np.svariso.corr*). Los códigos relacionados con el algoritmo NPB y la construcción de mapas de riesgos expuestos en el Capítulo 3 así como los procedimientos propuestos en el capítulo 4, se implementaron de manera conjunta entre el autor del paquete *npsp* y el doctorando, a lo largo de la presente investigación. Actualmente estamos trabajando en la incorporación de estas nuevas funciones en este paquete y algunos de los códigos preliminares ya se encuentran disponibles en la plataforma *GitHub*, de manera que se espera que en un futuro estos métodos sean accesibles de forma libre para los usuarios e investigadores en Geoestadística.

Bibliografía

Abramowitz, M. y Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volumen 55. Courier Corporation.

Anderes, E. B. y Stein, M. L. (2008). Estimating deformations of isotropic gaussian random fields on the plane. *The Annals of Statistics*, pp. 719–741.

Antunes, I. y Albuquerque, M. (2013). Using indicator kriging for the evaluation of arsenic potential contamination in an abandoned mining area (portugal). *Science of the Total Environment*, 442:545–552.

Armstrong, M. (1998). *Basic Linear Geostatistics*. Springer-Verlag.

Barry, R. P., Jay, M., y Hoef, V. (1996). Blackbox kriging: spatial prediction without specifying variogram models. *Journal of Agricultural, Biological, and Environmental Statistics*, pp. 297–322.

Beckers, F. y Bogaert, P. (1998). Nonstationary of the mean and unbiased variogram estimation: extension of the weighted least-squares method. *Mathematical Geology*, 30:223–240.

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. University of California Press.

Burrough, P. y McDonnell, R. (1998). *Principles of Geographical Information Systems*. Oxford University Press.

Cameletti, M., Ignaccolo, R., y Sylvan, D. (2013). Assessment and visualization of threshold exceedance probabilities in complex space–time settings: A case study of air quality in northern italy. *Spat. Stat.*, 5:57–68.

Cao, R. (1999). An overview of bootstrap methods for estimating and predicting in time series. *Test*, 8(1):95–116.

Carmack, P. S., Schucany, W. R., Spence, J. S., Gunst, R. F., Lin, Q., y Haley, R. W. (2009). Far casting cross-validation. *Journal of Computational and Graphical Statistics*, 18(4):879–893.

Carmack, P. S., Spence, J. S., y Schucany, W. R. (2012a). Generalised correlated cross-validation. *Journal of Nonparametric Statistics*, 24(2):269–282.

Carmack, P. S., Spence, J. S., Schucany, W. R., Gunst, R. F., Lin, Q., y Haley, R. W. (2012b). A new class of semiparametric semivariogram and nugget estimators. *Computational Statistics & Data Analysis*, 56(6):1737–1747.

Castillo-Páez, S., Fernández-Casal, R., y García-Soidán, P. (2017a). Bandwidth selection for local linear trend estimation. *Pre-print*.

Castillo-Páez, S., Fernández-Casal, R., y García-Soidán, P. (2017b). Bootstrap methods for inference on the variogram. *Pre-print*.

Cherry, S. (1996). An evaluation of a non-parametric method of estimating semi-variograms of isotropic spatial processes. *Journal of Applied Statistics*, 23(4):435–449.

Chiang, J.-L., Liou, J.-J., Wei, C., y Cheng, K.-S. (2014). A feature-space indicator kriging approach for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7):4046–4055.

Chilès, J. y Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley & Sons, New York, second edicin.

Chu, C. y Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Annals of Statistics*, 19(4):1906–1918.

Clark, R. G. y Allingham, S. (2011). Robust Resampling Confidence Intervals for Empirical Variograms. *Mathematical Geosciences*, 43:243–259.

Craven, P. y Wahba, G. (1978). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31(4):377–403.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology*, 17:563–586.

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, 22:239–252.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley and Sons.

Da Barrosa, M. R., Salles, A. V., y Ribeiro, C. d. O. (2016). Portfolio optimization through kriging methods. *Applied Economics*, pp. 1–12.

Davison, A. y Hinkley, D. (1997). *Bootstrap Methods and Their Application. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Draghicescu, D. e Ignaccolo, R. (2009). Modeling threshold exceedance probabilities of spatially correlated time series. *Electron. J. Stat.*, 3:149–164.

Efron, B. y Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Fan, J. y Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*, volumen 66. CRC Press.

Fan, J., J. y Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, 85(3):645–660.

Fernández-Casal, R. (2003). *Geoestadística Espacio-Temporal: Modelos flexibles de variogramas anisotrópicos no separables.* Tesis doctoral, Universidad de Santiago de Compostela.

Fernández-Casal, R. (2014). *npsp: Nonparametric spatial (geo)statistics*. R package version 0.3-6.

Fernández-Casal, R., Castillo-Páez, S., y Francisco-Fernández, M. (2017a). Nonparametric geostatistical risk mapping. *Stochastic Environmental Research and Risk Assessment*. Publicado online 27 Marzo 2017.

Fernández-Casal, R., Castillo-Páez, S., y García-Soidán, P. (2017b). Nonparametric estimation of the small-scale variability of heteroscedastic spatial processes. *Spat. Stat.* Aceptado para publicación.

Fernández-Casal, R. y Francisco-Fernández, M. (2014). Nonparametric bias-corrected variogram estimation under non-constant trend. *Stochastic Environmental Research and Risk Assessment*, 28(5):1247–1259.

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

Fernández-Casal, R., González-Manteiga, W., y M., F.-B. (2003b). Space-time dependency modeling using general classes of flexible stationary variogram models. *Journal of Geophysical Research: Atmospheres*, 108:8779.

Francisco-Fernández, M. y Opsomer, J. D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics*, 33:539–558.

Francisco-Fernández, M., Quintela-del Río, A., y Fernández-Casal, R. (2011). Nonparametric methods for spatial regression. an application to seismic events. *Environmetrics*, 23:85–93.

Francisco-Fernández, M. y Vilar-Fernández, J. M. (2001). Local polynomial regression estimation with correlated erors. *Communications in Statistics - Theory and Methods*, 30:1271–1293.

Franks, S. W. y Kuczera, G. (2002). Flood frequency analysis: Evidence and implications of secular climate variability, new south wales. *Water Resour. Res.*, 38(5):20–1–20–7.

Fuentes, M. (2001). A high frequency kriging approach for non-stationary environmental processes. *Environmetrics*, 12(5):469–483.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes. *Biometrika*, 89(1):197–210.

García-Soidán, P., Febrero-Bande, M., y González-Manteiga, W. (2004). Nonparametric kernel estimation of an isotropic variogram. *Journal of Statistical Planning and Inference*, 121:65–92.

García-Soidán, P., González-Manteiga, W., y Febrero-Bande, M. (2003). Local

linear regression estimation of the variogram. *Statistics & Probability Letters*, 64:169–179.

García-Soidán, P., Menezes, R., y Rubiños, O. (2012). An approach for valid covariance estimation via the fourier series. *Environmental Earth Sciences*, 66:615–624.

García-Soidán, P., Menezes, R., y Rubiños, O. (2014). Bootstrap approaches for spatial data. *Stochastic Environmental Research and Risk Assessment*, 28:1207–1219.

Gelfand, A., Diggle, P., Guttorp, P., y Fuentes, M. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press.

Goncalves, S. y Politis, D. (2011). Discussion: Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society*, 40:383–386.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Applied geostatistics series. Oxford University Press.

Goovaerts, P., Webster, R., y Dubois, J.-P. (1997). Assessing the risk of soil contamination in the swiss jura using indicator geostatistics. *Environmental and Ecological Statistics*, 4(1):49–64.

Guardiola-Albert, C. y Pardo-Igúzquiza, E. (2011). Compositional bayesian indicator estimation. *Stoch. Environ. Res. Risk. Assess.*, 25(6):835–849.

Hall, P., Horowitz, J. L., y Jing, B. Y. (1995). On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574.

Hanna-Attisha, M., LaChance, J., Sadler, R. C., y Champney Schnepp, A. (2016). Elevated blood lead levels in children associated with the flint drinking water crisis: a spatial analysis of risk and public health response. *American journal of public health*, 106(2):283–290.

Härdle, W. (1990). *Applied Nonparametric Regression*. Econometric Society Monographs. Cambridge University Press.

Härdle, W. y Tsybakov, A. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *Journal of Econometrics*, 81(1):223 – 242.

Iranpanah, N., Mohammadzadeh, M., y Taylor, C. (2011). A comparison of block and semi-parametric bootstrap methods for variance estimation in spatial statistics. *Computational Statistics & Data Analysis*, 55(1):578–587.

Isaaks, E. y Srivastava, R. (1989). *Applied Geostatistics*. Oxford University Press.

Journel, A. (1983). Nonparametric estimation of spatial distributions. *Journal of the International Association for Mathematical Geology*, 15(3):445–468.

Journel, A. G. y Huijbregts, C. J. (1978). *Mining geostatistics*. Academic press.

Jowett, G. H. (1952). The accuracy of systematic sampling from conveyor belts. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 1(1):50–59.

Kim, H. J. y Boos, D. D. (2004). Variance estimation in spatial regression using a non-parametric semivariogram based on residuals. *Scandinavian Journal of Statistics*, 31:387–401.

Krzysztofowicz, R. y Sigrest, A. A. (1997). Local climatic guidance for probabilistic quantitative precipitation forecasting. *Mon. Weather Rev.*, 125(3):305–316.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer Science & Business Media.

Lark, R. y Ferguson, R. (2004). Mapping risk of soil nutrient deficiency or excess by disjunctive and indicator kriging. *Geoderma*, 118:39 – 53.

Lele, S. (1995). Inner product matrices, kriging, and nonparametric estimation of variogram. *Mathematical geology*, 27(5):673–692.

Li, W., Zhang, C., Dey, D., y Wang, S. (2010). Estimating threshold-exceeding probability maps of environmental variables with markov chain random fields. *Stochastic Environmental Research and Risk Assessment*, 24(8):1113–1126.

Liu, R. Y. y Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. *Exploring the limits of bootstrap*, 225:248.

Liu, X. (2001). *Kernel smoothing for spatially correlated data*. Tesis doctoral, Department of Statistics, Iowa State University.

Matheron, G. (1962). *Traité de géostatistique appliquée*. Éditions Technip.

Matheron, G. (1971). *The theory of regionalized variables and its applications*, volumen 5. École national supérieure des mines.

Neuman, S. P. y Jacobson, E. A. (1984). Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. *Mathematical Geology*, 16:499–521.

Nordman, D. J., Lahiri, S. N., y Fridley, B. L. (2007). Optimal block size of variance estimation by a spatial block bootstrap method. *Sankhya: The Indian Journal of Statistics*, 69:468–493.

Olea, R. y Pardo-Igúzquiza, E. (2011). Generalized bootstrap method for assessment of uncertainty in semivariogram inference. *Mathematical Geosciences*, 43(2):203–228.

Oliver, M. A., Webster, R., y Mcgrath, S. P. (1996). Disjunctive kriging for environmental management. *Environmetrics*, 7(3):333–357.

Opsomer, J. D., Ruppert, D., Wand, M. P., Holst, U., y Hssjer, O. (1999). Kriging with nonparametric variance function estimation. *Biometrics*, 55(3):704–710.

Opsomer, J. D., Wang, Y., y Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, 16:134–153.

Pardo-Igúzquiza, E. y Olea, R. A. (2012). Varboot: A spatial bootstrap program for semivariogram uncertainty assessment. *Computers & Geosciences*, 41(0):188 – 198.

Pebesma, E. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30:683–691.

Pérez-González, A., Vilar-Fernández, J., y González-Manteiga, W. (2010). Nonparametric variance function estimation with missing data. *Journal of Multivariate Analysis*, 101(5):1123 – 1142.

Politis, D. y Romano, J. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313.

Robinson, P. M. y Thawornkaiwong, S. (2012). Statistical inference on regression with spatial dependence. *Journal of Econometrics*, 167:521–542.

Rupert, D. y Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics*, 22:1346–1370.

Ruppert, D., Wand, M. P., Holst, U., y Hösjer, O. (1997). Local polynomial variance-function estimation. *Technometrics*, 39(3):262–273.

Sampson, P. D. y Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Shapiro, A. y Botha, J. D. (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis*, 11(1):87–96.

Solow, A. (1985). Bootstrapping correlated data. *Journal of the International Association for Mathematical Geology*, 17(7):769–775.

Steere, L., Rosenzweig, N., y Kirk, W. (2016). Using conditional probability and a nonlinear kriging technique to predict potato early die caused by verticllium dahliae. En *Geographical Information Systems Theory, Applications and Management*, pp. 142–151. Springer.

Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Annals of Statistics*, 16:55–63.

Tolosana-Delgado, R., Pawlowsky-Glahn, V., y Egozcue, J.-J. (2008). Indicator kriging without order relation violations. *Math Geosci.*, 40(3):327–347.

Vilar-Fernández, J. M. y Francisco-Fernández, M. (2006). Nonparametric estimation of the conditional variance function with correlated errors. *Journal of Nonparametric Statistics*, 18(4-6):375–391.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. Springer Berlin Heidelberg.

Wand, M. P. y Jones, M. C. (1995). *Kernel smoothing*. Monographs on statistics and applied probability. Chapman and Hall/CRC, Boca Raton (Fla.), London, New York.

Webster, R. y Oliver, M. A. (2007). *Geostatistics for Environmental Scientists*. Jhon Wiley & Sons, Ltd.

Yaglom, A. (1986). *Correlation Theory of Stationary and Related Random Functions, Volume I: Basic Results*. Springer-Verlag.