Trabajo Fin de Máster

Comparación de estimadores de la función de distribución para datos espaciales

Susana Iglesias Rey

Máster en Técnicas Estadísticas

Curso 2015-2016

II

Propuesta de Trabajo Fin de Máster

Título en galego: Comparación de estimadores da función de distribución para datos espaciais

Título en español: Comparación de estimadores de la función de distribución para datos espaciales

English title: Distribution function estimators comparison for spatial data

Modalidad: Modalidad A

Autora: Susana Iglesias Rey, Universidad de Santiago de Compostela

Directores: María del Pilar García Soidán, Universidad de Vigo; Tomás Cotos Yáñez, Universidad de Vigo

Breve resumen del trabajo:

En este trabajo se estudiarán alternativas no paramétricas de la función de distribución para la construcción de un predictor no paramétrico. En particular, se propondrá la utilización de una media ponderada de funciones indicadoras, cuyos pesos dependerán de la separación entre las posiciones observadas y la localización objeto de estudio. La implementación práctica del estimador propuesto requiere la selección de un parámetro ventana o de una matriz ventana. Para esto último se sugerirán distintos métodos, si bien la inclusión de un único parámetro de suavización facilita la selección del parámetro óptimo desde un punto de vista computacional.

Otras observaciones:

Se realizarán estudios numéricos con datos simulados, evaluando el comportamiento de las alternativas no paramétricas con los métodos existentes. También se realizará una aplicación a datos reales, lo cual permitirá construir mapas de riesgo en una región de observación considerada.

Agradecimientos

El presente trabajo no podría haber sido realizado sin los conocimientos adquiridos en el Máster en Técnicas Estadísticas, por ello dar las gracias al conjunto del profesorado.

En especial a mi tutora María del Pilar García Soidán y a mi tutor Tomás Cotos Yáñez, sin cuya ayuda tampoco hubiese sido posible.

Índice general

Resumen

Resumen en español

La estimación de la función de distribución de un proceso espacial resulta de gran interés en el ámbito de la estadística espacial. Por ello, en el presente trabajo se ha propuesto un predictor no paramétrico alternativo empleando una media ponderada de funciones indicadoras, cuyos pesos dependen de la separación entre las posiciones observadas y las localizaciones objeto de estudio.

El comportamiento del estimador propuesto se evaluó y comparó con la aproximación que proporciona el kriging indicador, tradicionalmente utilizado en este contexto, por medio de simulaciones realizadas bajo distintos escenarios. En particular, se han diseñado procesos espaciales con distintas estructuras para la tendencia y la dependencia espacial. La implementación del nuevo estimador requiere la selección de un parámetro ventana o de una matriz ventana, que también ha sido abordada en esta investigación.

Finalmente, se realizó una aplicación de ambos estimadores a un caso práctico con datos reales, descritos en Cressie (1993) y relativos al nivel de carbón en una mina de Pennsylvania, con objeto de construir mapas de riesgo relativos a la concentración de dicho mineral.

English abstract

The estimation of the distribution function values has become more important in spatial statistics. Owing to this, in this study a new non parametric predictor has been proposed, based on a weighted mean of indicator functions, whose weights depend on the distance between observed and objective locations.

The behavior of the estimator proposed was studied and compared with the kriging indicator estimation by means of simulations in different scenarios. In particular, spatial processes with different trends and correlation structures have been generated. Furthermore, a bandwidth selection according to different criteria was necessary for the predictor implementation.

To conclude, a real application was conducted with the coal-ash data obtained in Cressie (1993), with the aim of constructing risk maps relative to mineral concentration.

Introducción

La *estadística espacial* o *geoestadística* fue descrita por primera vez por Matheron (1963) como el conjunto de aplicaciones de la teoría de las variables regionalizadas a la estimación de los depósitos minerales. En la actualidad, se designa la *estadística espacial* como un conjunto de técnicas cuyo objetivo es la determinación de la estructura de autocorrelación entre datos asociados a una posición espacial y la predicción en nuevas localizaciones. Cuando las observaciones del fenómeno de interés vienen asociadas a una posición del espacio es necesario emplear técnicas estadísticas específicas que exploten de manera adecuada dicha propiedad, pues, generalmente, no se satisface la hipótesis de partida de que las observaciones del fenómeno han sido tomadas bajo idénticas condiciones y de forma independiente. Es en esta situación, ante el estudio de datos que tienen una componente espacial asociada a ellos de forma que se presupone que datos próximos presentarán propiedades parecidas, donde se hace necesaria la implementación y desarrollo de las técnicas específicas de la *estadística espacial*.

Aunque hoy en día el uso de estas técnicas se ha extendido a prácticamente todas las ciencias, epidemiología, geología, ecología, astronomía, climatología, y, en general, a todas aquellas que trabajen con fenómenos en los que la influencia de la localización de los datos debe ser tenida en cuenta, el desarrollo de la *estadística espacial* es todavía reciente. Las primeras muestras del estudio de datos espaciales datan del siglo XVII por medio de mapas de datos, pero no sería hasta la década de 1980 cuando surgiría la *geoestadística*, término acuñado por Matheron, como una ciencia híbrida entre las matemáticas, geología, ingeniería de minas y estadística, cuyos modelos estadísticos sí incluían la dependencia a pequeña y gran escala que presentan los datos espaciales, asociados a las localizaciones en las que han sido observados.

Uno de los objetivos clave de la *geoestadística* es la reconstrucción de un fenómeno sobre la región de observación a partir de un conjunto finito de datos muestreados. De nuevo Matheron, formalizaría y generalizaría el uso de las técnicas *kriging* con este objetivo. El nombre de dichas técnicas de predicción se le debe al geólogo sudafricano D. G. Krige, quien desarrolló una serie de metodologías para hacer predicciones en la evaluación de reservas de las minas de oro en Sudáfrica en 1951.

Parte de la investigación estadística de la última década ha estado centrada en el desarrollo de métodos más apropiados para la cuantificación de la dependencia espacial entre datos, así como en la generalización a procesos en los que, además de tener en cuenta la dependencia espacial y temporal, se considera también la evolución espacio-temporal conjunta, dando lugar a modelos mucho más complejos.

La estimación y modelización de la correlación espacial, que recibe el nombre de *análisis estructural*, será expuesta en la primera parte del Capítulo 1, junto con conceptos básicos sobre los procesos estocásticos espaciales. Una vez caracterizada la estructura de dependencia, la predicción de valores en puntos no muestrales se puede hacer empleando la técnica *kriging* ya mencionada. Dicha metodología se presentará al final del capítulo.

Si bien las estimaciones suelen realizarse en torno al proceso estocástico espacial, a veces resulta de interés estimar la probabilidad de que la variable espacial sea mayor (o menor) que un valor determinado, para lo cual, se requiere estimar la función de distribución del proceso espacial. Esta estimación permite, por ejemplo, construir mapas de riesgo, en los que figura la probabilidad de que una determinada substancia química sobrepase los niveles admitidos para el consumo humano, y será

sobre lo que se centre el objetivo del presente trabajo fin de máster. En el Capítulo 2 se presentarán las técnicas ya conocidas para realizar la estimación de la función de distribución y se propondrá una alternativa no paramétrica indicadora, mediante la utilización de una media ponderada de funciones indicadoras, cuyos pesos dependerán de la separación entre las posiciones observadas y la localización objeto de estudio.

Para comprobar el funcionamiento de los estimadores propuestos, se realizarán estudios numéricos bajo diversos escenarios, cuyos resultados se presentarán en el Capítulo 3.

Finalmente, en el último Capítulo 4 se incluyen las conclusiones del trabajo fin de máster.

Capítulo 1

Generalidades sobre los procesos estocásticos

En este capítulo se recogen las principales nociones y resultados necesarios sobre los procesos estocásticos espaciales, así como sobre los procesos de inferencia relativos a los mismos.

1.1. Conceptos básicos

La *estadística espacial* aglutina un conjunto de técnicas usadas para analizar y predecir valores de una propiedad distribuida en el espacio. A diferencia de la estadística clásica, en el contexto espacial el problema que se plantea es la existencia de relación entre los datos observados en un área determinada. Esto se traduce en que los modelos espaciales deben incorporar la dependencia presente en todas las direcciones y que se vuelve cada vez más débil conforme las posiciones espaciales muestreadas están más dispersas, ya que se incrementa la dificultad de encontrar observaciones semejantes en localizaciones próximas.

La noción de que los datos cercanos en tiempo o espacio están presumiblemente correlacionados y, por lo tanto, no pueden ser modelados independientemente, es natural y ha sido largamente empleada en la historia de la estadística.

Partiendo de la noción de proceso estocástico, se empleará la formulación de Cressie (1993) para la definición matemática de un proceso espacial.

Definición 1.1. Un proceso estocástico es una colección de variables aleatorias $\{Z(\tau) \in S/\tau \in T\}$, indexada por un parámetro τ, que toma valores en el espacio de parámetros T, mientras las variables aleatorias $Z(\tau)$ toman valores en el espacio de estados S.

Si particularizamos al contexto espacial, la indexación vendría determinada por las localizaciones s de una región de observación D en un espacio euclídeo d-dimensional $\mathbb{R}^d$. La variable aleatoria asociada a la localización s puede ser unidimensional o multidimensional y suele designarse como $Z(s)$. De este modo:

$$\{Z(s) \in \mathbb{R}/s \in D\} \tag{1.1}$$

representa un *proceso estocástico espacial* (también llamado función aleatoria, campo espacial aleatorio, variable aleatoria georeferenciada o variable regionalizada), denotado a menudo simplemente por $Z(s)$.

El proceso aleatorio (1.1) es el objeto de estudio de la estadística espacial. Particularmente, la geoestadística estudiará aquellos procesos en los que el índice s varíe de forma continua en la región de observación $D \subset \mathbb{R}^d$. Si $d = 2$, $Z(s)$ puede asociarse a una variable medida en un punto s del plano

(Giraldo 2002).
De forma general, los estados de un proceso estocástico espacial se suelen descomponer en una parte determinista y una parte aleatoria de la forma:

$$Z(s) = \mu(s) + Y(s),$$

tal que $\mathbb{E}[Z(s)] = \mu(s)$, $\forall s \in D$.

La componente de variación determinista $\mu(s)$ (también llamada variación a gran escala, variación no estocástica o tendencia global) suele ser el resultado de una tendencia global que en la práctica se trata de modelar. La variación estocástica (también llamada variación a pequeña escala) es el residuo resultante de eliminar la tendencia. Una vez ésta ha sido eliminada, se pueden encontrar patrones de comportamiento que obedecen al entorno en el que se encuentran las localizaciones. Esta variación espacial puede modelarse atendiendo a criterios espaciales.

Si $\{Z_i(s), 1 \leq i \leq m, s \in D\}$ son m procesos espaciales univariantes, el conjunto de vectores aleatorios $\mathbf{Z}(s) = (Z_1(s), ..., Z_m(s))'$ se denominará proceso espacial multivariante (también proceso espacial vectorial, campo vectorial espacial, vector aleatorio georeferenciado o vector regionalizado).

Una realización del proceso estocástico espacial (1.1) en las localizaciones $s \in D$ se denotará por $\{z(s)/s \in D\}$. En este trabajo, se supondrá que D es una región de observación, en la que s puede variar de forma continua, aunque tal y como se indica en Cressie (1993, pp. 8-9) se podría considerar de forma más general que D es un conjunto aleatorio.

Sea $\{s_1, ..., s_n\} \subset D$ un conjunto de n localizaciones muestrales, para las cuales se observará el proceso $\{Z(s_1), ..., Z(s_n)\}$. Se representará por $\{z(s_1), ..., z(s_n)\}$ una realización del mismo. El proceso estocástico espacial (1.1) generalmente se define a través de las funciones de distribución de dimensión finita:

$$F_{s_1...s_n}(x_1....x_n) = \mathbb{P}(Z(s_1) \leq x_1...Z(s_n) \leq x_n)$$

para cualquier conjunto de localizaciones $\{s_1, ..., s_n\} \subset D$ y cualquier conjunto de valores reales $\{x_1, ..., x_n\} \subset \mathbb{R}$, para todo $n \in \mathbb{N}$.

Dichas funciones de distribución deben verificar las condiciones de Kolmogorov de simetría y consistencia:

- Simetría:

$$F_{s_1...s_n}(x_1....x_n) = F_{s_{i_1}...s_{i_n}}(x_1....x_n).$$

  donde $i_1, ..., i_n$ es una permutación de los índices $1, ..., n$.

- Consistencia:

$$F_{s_1...s_n, s_{n+1}, ...s_{n+m}}(x_1...x_n, \infty...\infty) = F_{s_1...s_n}(x_1....x_n).$$

En la práctica, resulta difícil caracterizar adecuadamente la función de distribución del proceso espacial, de ahí que en general se trata de imponer condiciones sobre los primeros momentos de la distribución de $Z(s)$, cuya estimación es menos compleja. Se denotará por $\mu(\cdot)$ la tendencia del proceso, que representa el momento de primer orden, es decir, $\mathbb{E}[Z(s)] = \mu(s)$, $\forall s \in D$. A lo largo de este trabajo se supondrá que $\mu(s) < \infty$, para cada localización $s \in D$. Además, se utilizará la siguiente notación para la varianza, $Var[Z(s)] = \sigma^2(s)$.

Puesto que generalmente se dispone de una única realización discreta del proceso (1.1) es necesario asumir ciertas hipótesis simplificadoras sobre $Z(s)$, de forma que asegure cierta regularidad en los datos y se pueda realizar inferencia sobre el modelo a partir de ellos. Para ello, se suele recurrir a la imposición de algún tipo de estacionariedad, lo que permite que algunas características del proceso se repitan en la región de observación, proporcionando así la regularidad necesaria para llevar a cabo la estimación e inferencia.

Definición 1.2. Se dirá que el proceso espacial $\{Z(s)/s \in D\}$ es *estrictamente estacionario* (también llamado estacionario en sentido fuerte) si para cualquier conjunto de localizaciones muestrales $\{s_1, ..., s_n\} \subset D$, la función de distribución del proceso permanece invariable ante traslaciones, es decir:

$$F_{s_1 \dots s_n}(z_1 \dots . z_n) = F_{s_1+t \dots s_n+t}(z_1 \dots . z_n), \qquad \forall t \in \mathbb{R}^d.$$

De esta forma, al trasladar un conjunto de localizaciones respecto a cualquier vector t, la distribución conjunta no varía. Esta condición es demasiado restrictiva en la práctica, por lo que se suele recurrir a otros tipos de estacionariedad obtenidos como relajaciones de la estricta.

Definición 1.3. Se dirá que el proceso espacial $\{Z(s)/s \in D\}$ es *estacionario de segundo orden* (también llamado proceso estacionario homogéneo, débilmente estacionario o simplemente estacionario) si verifica que:

1. $\mathbb{E}[Z(s)] = \mu$, $\forall s \in D$; es decir, la tendencia del proceso existe y no depende de s.
2. $Cov(Z(s), Z(s')) = C(s - s') < \infty$, $\forall s, s' \in D$; es decir, la función de covarianza existe y sólo depende del vector de separación $s - s'$ entre las localizaciones involucradas. Esta condición permite obtener predictores lineales óptimos, Cressie (1993).
   La función $C(\cdot)$ recibe el nombre de *covariograma* (también llamado *función de covarianza* o *autocovarianza*).

Además, cuando $C(s-s')$ es función sólo de $\|s-s'\|$ para cualesquiera localizaciones $s, s' \in D$, donde $\|\cdot\|$ denota la norma euclídea, se dirá que $C(\cdot)$ es una función *isotrópica*. En caso contrario se dirá que $C(\cdot)$ es *anisotrópica*. Por lo tanto, una variable regionalizada es isotrópica si la dependencia espacial del proceso entre dos localizaciones cualesquiera depende únicamente de la distancia existente entre ellas y no de su localización. El caso contrario, las anisotropías, están causadas por procesos físicos subyacentes que se comportan de forma diferente en el espacio, como puede ser el campo gravitatorio, cuyo comportamiento difiere de la componente vertical a la horizontal (Martínez Ruiz, 2008).

En algunas ocasiones el uso del covariograma se sustituye por el *correlograma*, definido por:

$$\rho(s - s') = \frac{C(s - s')}{C(0)} \in [-1, 1], \tag{1.2}$$

suponiendo que $C(0) \neq 0$.

La estacionariedad de segundo orden implica que la varianza del proceso existe, es finita y no depende de la localización pues

$$Var[Z(s)] = Cov[Z(s), Z(s)] = C(0) = \sigma^2, \quad \forall s \in D.$$

Cabe destacar que los procesos estacionarios de segundo orden son más generales que los procesos estrictamente estacionarios, es decir, un proceso estrictamente estacionario con varianza finita es un proceso estacionario de segundo orden. El recíproco no tiene por qué cumplirse en general, aunque sí es cierto para procesos gaussianos.

Definición 1.4. Se dirá que el proceso espacial $\{Z(s)/s \in D\}$ es *gaussiano* si, para cualquier conjunto de localizaciones $\{s_1, ..., s_n\} \subset D$, el vector aleatorio $(Z(s_1), ..., Z(s_n))'$ sigue una distribución normal multivariante.

Por consiguiente, teniendo en cuenta las propiedades de un vector aleatorio normal multivariante, si el proceso $\{Z(s)/s \in D\}$ es gaussiano, las variables aleatorias $Z(s_i)$ $i \in \{1...n\}$ son gaussianas.
La implicación de que un proceso gaussiano estacionario de segundo orden sea estrictamente estacionario es cierta, pues la distribución de estos procesos está completamente caracterizada por su media y covariograma.

Existen fenómenos físicos que poseen capacidad infinita de dispersión, por lo que no existe ni la varianza ni la covarianza de la función aleatoria. Sin embargo, en algunos casos los incrementos de estas funciones aleatorias son estacionarios de segundo orden (Cressie 1993) y, por lo tanto, tienen varianza finita. De esta forma resulta útil definir la siguiente noción de estacionariedad:

Definición 1.5. Se dirá que el proceso espacial $\{Z(s)/s \in D\}$ es *intrínsecamente estacionario* (también llamado *proceso de incrementos estacionarios u homogéneos*) si verifica que:

1. $\mathbb{E}[Z(s)] = \mu$, $\forall s \in D$, es decir, la tendencia del proceso existe y no depende de s.

2. $Var[Z(s) - Z(s')] = 2\gamma(s - s')$, $\forall s, s' \in D$, es decir, la varianza de los incrementos existe y sólo depende del vector de separación $s - s'$ entre las localizaciones involucradas.
   La función $\gamma(\cdot)$ recibe el nombre de *semivariograma* y $2\gamma(\cdot)$ *variograma*.

De manera análoga a la estacionariedad de segundo orden, cuando $Var[Z(s) - Z(s')]$ depende sólo de $\|s - s'\|$ para cualesquiera localizaciones $s, s' \in D$, se dirá que $\gamma(\cdot)$ es *isotrópico*. En caso contrario, se dirá que $\gamma(\cdot)$ es *anisotrópico*. En la literatura a menudo se designa a un proceso intrínseco e isotrópico como un proceso *homogéneo*.

La estacionariedad de segundo orden implica estacionariedad intrínseca. Si $Z(s)$ es un proceso estacionario de segundo orden, entonces:

$$\begin{aligned}
\gamma(s - s') &= \frac{1}{2} Var[Z(s) - Z(s')] = \frac{1}{2}\mathbb{E}[(Z(s) - Z(s'))^2] \\
&= \frac{1}{2} Var[Z(s)] + \frac{1}{2} Var[Z(s')] - Cov[Z(s), Z(s')] \\
&= \sigma^2 - C(s - s'),
\end{aligned}$$

y, por consiguiente, es un proceso intrínsecamente estacionario.

El recíproco no tiene porqué ser cierto. Haciendo uso del ejemplo presente en Cressie (1993), el movimiento Browniano d-dimensional isotrópico constituye un ejemplo para el cual el semivariograma está definido, pero no la función de covarianzas $C(\cdot)$. Si $\{W(s)/s \in \mathbb{R}^d\}$ es el proceso mencionado, entonces:

$$Var(W(s + t) - W(s)) = \|t\|, \qquad t \in \mathbb{R}^d.$$

Sin embargo, $Cov(W(u) - W(v)) = \frac{1}{2}(\|u\| + \|v\| - \|u - v\|)$; $u, v \in \mathbb{R}^d$, que no es función de $u - v$.

Para que un proceso intrínsecamente estacionario lo sea también de segundo orden es necesario que el variograma esté acotado, es decir,

$$\lim_{t \to \infty} \gamma(t) = M < \infty,$$

pues de esta forma se puede obtener el covariograma correspondiente,

$$C(t) = \sigma^2 - \gamma(t).$$

Además, de forma análoga se tiene la igualdad, $\rho(t) = 1 - \dfrac{\gamma(t)}{C(0)}$.

Vistos los tipos de estacionariedad que se considerarán a lo largo de las siguientes páginas, cabe definir el recíproco, como se plantea a continuación.

Definición 1.6. Se dirá que la variable regionalizada $\{Z(s)/s \in D\}$ es *no estacionaria* si la media del proceso depende de la localización, es decir:

$$\mathbb{E}[Z(s)] = \mu(s).$$

Existe un enfoque que considera a las variables regionalizadas no estacionarias como intrínsecas de orden k, es decir, que si se toman los incrementos de un orden k adecuado estos son estacionarios (Díaz Viera 2002).

Dentro de las funciones estacionarias de segundo orden, existe una subclase que posee la cualidad de ser *ergódica*. Un proceso se dice *ergódico* o que posee la propiedad de *ergodicidad* si, a partir de una única realización del mismo, se pueden estimar sus parámetros de interés (Fernández Casal, 2003). Esta propiedad garantiza la convergencia en media cuadrática de los promedios muestrales de interés a sus correspondientes teóricos. Así, se puede hablar, por ejemplo, de ergodicidad en media cuando la media muestral converge en media cuadrática a la media teórica. De forma análoga, se puede hablar de ergodicidad en la covarianza o en el semivariograma.

La noción de ergodicidad se ha generalizado para procesos espaciales en los que el dominio donde se realiza el promedio tiende a infinito. Por ejemplo,

Definición 1.7. Se dice que el proceso espacial $\{Z(s)/s \in D\}$ es *ergódico en media* si verifica que:

1. $\mathbb{E}[Z(s)] = \mu$, $\forall s \in D$, es decir, la tendencia del proceso existe y no depende de s.
2. Además,
$$\lim_{L\to\infty} \frac{1}{(2L)^d} \int_{U_L} Z(s)ds = \mu,$$
donde el dominio de integración es el cubo d-dimensional $U_L = \{s = (s_1, ..., s_d)/|s_i| < L, i = 1...d\}$.

Las funciones empleadas habitualmente en el modelado de la dependecia espacial serán las ya presentadas, el variograma y el covariograma, lo que hace necesario el estudio de dichas funciones. En la siguiente sección se presentarán estimadores de ambas funciones que, para ser válidos, deberán cumplir una serie de propiedades que se incluyen a continuación.

Proposición 1. *Dado un proceso espacial $\{Z(s)/s \in D\}$ estacionario de segundo orden con función de covarianza $C(\cdot)$ se verifica que:*

1. $C(0) = Var[Z(s)] = \sigma^2 \geq 0$, $\forall s \in D$.
2. $C(\cdot)$ *es una función simétrica, es decir,* $C(t) = C(-t)$.
3. $C(\cdot)$ *es una función semidefinida positiva, es decir,*
$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(s_i - s_j) \geq 0, \quad \forall n \geq 1, \forall s_i \in D, \forall a_i \in \mathbb{R}, i \in \{1, ..., n\}.$$

La última propiedad se deriva de que:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(s_i - s_j) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov[Z(s_i), Z(s_j)] =^{1} \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (Z(s_i) - \mu)(Z(s_j) - \mu)] \\ &= \mathbb{E}[\sum_{i=1}^{n} a_i^2 (Z(s_i) - \mu)^2] \geq 0.\end{aligned}$$

donde la igualdad ^{1} se ha obtenido teniendo en cuenta la linealidad de la esperanza matemática, de modo que $\mathbb{E}[X] + \mathbb{E}[Y] = \mathbb{E}[X + Y]$.

Además, haciendo uso de la propiedad 1 y del Teorema de Cauchy-Schwarz 1.8 (Evans y Rosenthal, 2005, pp. 207), se concluye que $|C(t)| \leq C(0)$, $\forall t \in \mathbb{R}^d$. Asimismo, el producto de funciones de covarianza es también una función de covarianza (Montero y Larraz, 2008).

Teorema 1.8 (Teorema de Cauchy-Schwarz)**.** *Sean X e Y dos variables aleatorias cualesquiera con varianza finita y no nula. Entonces*

$$|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)}.$$

La propiedad 3 es necesaria y suficiente para que exista un proceso estacionario de segundo orden con dicho covariograma $C(\cdot)$ (Cressie, 1993).

De la propiedades anteriores y de la definición de correlograma (1.2) se deduce a su vez que $\rho(t) = \rho(-t)$ y $\rho(0) = 1$.

De forma análoga a la función de covarianza se tienen las siguientes propiedades para el semivariograma.

Proposición 2. *Dado un proceso espacial $\{Z(s)/s \in D\}$ intrínsecamente estacionario con semivariograma $\gamma(\cdot)$ se verifica que:*

1. $\gamma(0) = 0$.
2. *$\gamma(\cdot)$ es una función simétrica, es decir, $\gamma(t) = \gamma(-t)$, $\forall t \in \mathbb{R}^d$.*
3. $\gamma(t) \geq 0$, $\forall t \in \mathbb{R}^d$.
4. *Es una función condicionalmente semidefinida negativa, es decir,*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(s_i - s_j) \geq 0, \quad \forall n \geq 1, \forall s_i \in D, \forall a_i \in \mathbb{R}, i \in \{1,...,n\}, tales\ que \sum_{i=1}^{n} a_i = 0.$$

   Esta condición es más débil que la condición de que el covariograma sea una función definida positiva, por consiguiente, la clase de modelos de semivariogramas válidos es mayor que la de covariogramas.

5. *Como consecuencia de la propiedad anterior, el semivariograma debe tener un ritmo de crecimiento inferior a t^2 (Matheron 1971), es decir,*

$$\lim_{n\to\infty} \frac{\gamma(t)}{t^2} = 0.$$

La demostración de la propiedad 5 puede encontrarse en Montero y Larraz (2008). Considérense tres localizaciones s, $s+t$ y $s-t$, asociadas a los pesos $a_1 = -2$, $a_2 = a_3 = 1$ respectivamente. Dado que el semivariograma es una función condicionalmente semidefinida negativa se tiene que,

$$-8\gamma(t) + 2\gamma(2t) \geq 0,$$

equivalentemente,

$$\gamma(2t) \geq 4\gamma(t).$$

Esta última desigualdad sólo se verifica si el variograma crece más despacio que la parábola, de donde se obtiene que

$$\lim_{n\to\infty} \frac{\gamma(t)}{t^2} = 0.$$

Las funciones que satisfacen las propiedades anteriores se denominan modelos válidos del semivariograma.

Generalmente el semivariograma crece con la distancia, pues en la mayoría de procesos físicos existen mayores similitudes entre los valores observados en localizaciones próximas que disminuyen al

aumentar la distancia (Martínez Ruiz, 2008).

Cabe destacar que cualquier combinación lineal de modelos válidos da lugar a un modelo válido, tanto para las funciones de covarianza como para los semivariogramas. Además, bajo el supuesto de estacionariedad tanto el covariograma, como el correlograma, como el semivariograma, pueden ser utilizados en la determinación de la correlación espacial de las variables regionalizadas. No obstante, de las tres funciones anteriores, el semivariograma es la única función cuya estimación no requiere la aproximación de los parámetros media y varianza y, por ello, es la función más utilizada en la práctica.

De forma general, los semivariogramas pueden dividirse en dos grandes grupos: acotados (también llamados transitivos) y no acotados. El crecimiento con la distancia de los semivariogramas transitivos se estabiliza alrededor de un determinado valor, que se denotará por c_1, conocido como *umbral* o *meseta* (en inglés, *sill*).

$$c_1 = \lim_{t\to\infty} \gamma(t).$$

Si el proceso es estacionario de segundo orden entonces el umbral coincide con la varianza σ^2. Si además $\lim_{t\to\infty} C(t) = 0$, entonces $\sigma^2 = C(0)$.
Se denominará *rango* o *alcance* (en inglés, *range*) a la distancia para la cual el semivariograma alcance su meseta, en caso de que exista. Es decir, es el valor real c_2 tal que para todo $\|t\| \geq c_2$, $\gamma(t) = c_1$. El rango también representa el valor a partir del cual el covariograma se anula, es decir, la distancia a partir de la cual las observaciones son independientes. Por consiguiente, cuanto menor sea el rango más cerca está el modelo de la independencia espacial. Puesto que el rango no tiene por qué existir, se define el *rango efectivo* (en inglés *effective range*) como la distancia para la cual el semivariograma alcanza el 95 % de su meseta, es decir, el valor real c_2' tal que para todo $\|t\| \geq c_2'$, $\gamma(t) = 0.95c_1$.

Por regla general debe verificarse que $\gamma(0) = 0$, pero en la práctica suele ocurrir que,

$$\lim_{t\to 0} \gamma(t) = c_0 > 0.$$

El valor de c_0 se denomina *efecto pepita* (en inglés, *nugget*). Esta discontinuidad puede deberse a diversas razones, como variaciones a pequeña escala o errores de medida. En la práctica solo se observa un conjunto discreto de datos $\{z(s_i), s_i \in D, 1 \leq i \leq n\}$, por lo que se desconoce el comportamiento del variograma para distancias menores que $m = \text{mín}\{\|s_i - s_j\|, 1 \leq i, j \leq n, i \neq j, s_i, s_j \in D\}$, de forma que c_0 puede ser indicativo de la existencia de estructura espacial concentrada a distancias inferiores a las observadas. Si se interpreta el efecto pepita como un error en las mediciones, para un modelo que explique bien la realidad la pepita no debería representar más del 50 % del umbral (Giraldo, 2002).
Cuando existe efecto pepita, se define la *meseta* o *umbral parcial* (en inglés, *partial sill*) como $c_1 - c_0$.

En la siguiente sección se estudiará el proceso de caracterización de las funciones de covarianza y variograma.

1.2. Caracterización de la estructura de dependencia

El proceso de estimación y modelación de la función que describe la correlación espacial de la variable regionalizada a partir de la imposición de hipótesis sobre su variabilidad, es conocido como *análisis estructural* (Díaz Viera 2002). En este marco es donde se obtiene un modelo matemático para el proceso estocástico espacial de estudio, lo cual desempeña un papel crucial en la predicción espacial que se tratará en la siguiente sección.
Como se ha mencionado anteriormente, tanto el covariograma, como el correlograma, como el semivariograma, pueden ser utilizados en la determinación de la correlación espacial de las variables regionalizadas. No obstante, el semivariograma es la función más utilizada en la práctica, dado que no requiere hacer una estimación de la media y la varianza y, además, los procesos intrínsecamente estacionarios son más generales que los estacionarios de segundo orden.

En la caracterización del variograma se suele proceder como sigue:

1. Primero, se realiza una aproximación inicial del variograma a partir de un estimador no paramétrico.

2. A continuación, puesto que el variograma empírico no tiene por qué ser válido, se selecciona un modelo paramétrico válido del variograma y se ajustan los parámetros a partir de los datos disponibles, tomando como referencia el estimador obtenido en el primer paso.

3. Finalmente, se realiza la diagnosis del variograma ajustado, generalmente mediante el método de validación cruzada.

Se presentará en esta sección la caracterización tanto del variograma como de la función de covarianza.

Sea $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ un proceso aleatorio espacial para el cual se conocen los valores $Z(s_1), ..., Z(s_n)$ en las localizaciones $s_1, ..., s_n$.
Teniendo en cuenta la definición de semivariograma dada en la sección anterior se tiene que:

$$\begin{aligned}\gamma(s-s') &= \frac{1}{2}Var[Z(s)-Z(s')] = \frac{1}{2}\mathbb{E}[(Z(s)-Z(s'))^2] - \frac{1}{2}[\mathbb{E}(Z(s)) - \mathbb{E}(Z(s'))]^2 \\ &= \frac{1}{2}\mathbb{E}[(Z(s)-Z(s'))^2].\end{aligned}$$

La estimación más sencilla del semivariograma se obtiene mediante el método de los momentos, denominado **estimador empírico** (también clásico o muestral) debido a Matheron (1963), definido por:

$$\hat{\gamma}(t) = \frac{1}{2|N(t)|}\sum_{(i,j)\in N(h)}[Z(s_i)-Z(s_j)]^2, \tag{1.3}$$

donde $N(t) = \{(i,j)/s_i - s_j = t\}$ y $|N(t)|$ es su cardinal. Puesto que $\hat{\gamma}(t)$ es una media muestral, es un estimador no robusto.

El estimador del variograma (1.3) debe ser adaptado en la práctica para el caso en el que los datos no se distribuyan de manera regular en el espacio. De esta forma, se puede definir una versión suavizada de (1.3) de la forma:

$$\hat{\gamma}(t) = \frac{1}{2|N'(t)|}\sum_{(i,j)\in N'(t)}[Z(s_i)-Z(s_j)]^2, \tag{1.4}$$

donde $N'(t) = \{(i,j)/s_i - s_j \in T(t)\}$, $T(t)$ una región de tolerancia en torno a t y $|N'(t)|$ el cardinal de $N'(t)$. La estimación del variograma debe de estar basada en un número suficiente de valores, para lo cual la elección de t se realiza de forma que $||t|| < d_{max}/2$, donde d_{max} es la distancia máxima observada.

El estimador de Matheron es no paramétrico y cuando los datos de distribuyen de manera regular en una rejilla y la distribución es normal, es el estimador óptimo.
Como se señala en Díaz Viera (2002), en la práctica el empleo del estimador clásico puede producir variogramas experimentales erráticos debidos a desviaciones del caso ideal para la aplicación del mismo. Esto puede deberse, entre otros factores, a alguno de los que se mencionan a continuación:

- Al hecho de que las variables regionalizadas se alejan de la distribución normal.

- A la presencia de heterocedasticidad, es decir que la magnitud del semivariograma esté asociada a la magnitud de los valores de los datos.

- A desviaciones en el muestreo (sesgo en las localizaciones seleccionadas).

- A la existencia de valores atípicos.

Si el proceso $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ es estacionario de segundo orden, entonces:

$$C(s-s') = Cov[Z(s), Z(s')] = \mathbb{E}[(Z(s)-\mu)(Z(s')-\mu)].$$

De forma análoga al caso del semivariograma, para la función de covarianza el estimador empírico se define de la forma:

$$\hat{C}(t) = \frac{1}{|N(t)|} \sum_{(i,j)\in N(t)} [Z(s_i)-\bar{Z}][Z(s_j)-\bar{Z}], \tag{1.5}$$

donde $N(t) = \{(i,j)/s_i - s_j = t\}$, $|N(t)|$ es su cardinal y $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(s_i)$ es un estimador de μ.
Se puede observar que $\hat{C}(t)$ requiere la estimación de la media del proceso, μ, siendo éste el principal problema de este estimador.

De nuevo, el estimador de la función de covarianza (1.5) debe ser adaptado en la práctica para el caso en el que los datos no se distribuyan de manera regular en el espacio. De esta forma, se puede definir una versión suavizada de 1.5 de la forma:

$$\hat{C}(t) = \frac{1}{|N'(t)|} \sum_{(i,j)\in N'(t)} [Z(s_i)-\bar{Z}][Z(s_j)-\bar{Z}], \tag{1.6}$$

donde $N'(t) = \{(i,j)/s_i - s_j \in T(t)\}$, $T(t)$ una región de tolerancia en torno a t, $|N'(t)|$ el cardinal de $N'(t)$ y $\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z(s_i)$ un estimador de μ.
Como destaca Martínez Ruiz (2008), conviene señalar que los estimadores del variograma y de la función de covarianza no verifican necesariamente que $2\hat{\gamma}(t) = 2[\hat{C}(0) - \hat{C}(t)]$.

Una alternativa robusta al estimador clásico del semivariograma es el propuesto por Cressie y Hawkins que toma la forma:

$$\hat{\gamma}(t) = \frac{1}{2(0.457 + 0.494/N(t))} \left[\frac{1}{|N(t)|} \sum_{(i,j)\in N(t)} |Z(s_i) - Z(s_j)|^{1/2} \right]^4,$$

o

$$\hat{\gamma}(t) = \frac{1}{2(0.457 + 0.494/N(t))} \left[\text{Mediana}\left(|Z(s_i) - Z(s_j)|^{1/2} / s_i - s_j \in N(t) \right) \right]^4,$$

donde $N(t) = \{(i,j)/s_i - s_j = t\}$ y $|N(t)|$ es su cardinal. El estimador de Cressie y Hawkins (1980) se considera óptimo en condiciones de normalidad, sin embargo, infravalora los datos atípicos.

Si el proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ es isotrópico, entonces la varianza entre los datos depende solamente de la distancia entre las posiciones y no de la dirección. En este caso los estimadores anteriores quedarían de la forma:

- Estimador empírico:

$$\hat{\gamma}(t) = \frac{1}{2|N^*(t)|} \sum_{(i,j)\in N^*(t)} [Z(s_i) - Z(s_j)]^2,$$

  donde $N^*(t) = \{(i,j)/\|s_i - s_j\| = t\}$ y $|N^*(t)|$ es su cardinal.

- Estimador empírico de Cressie y Hawkins 1:

$$\hat{\gamma}(t) = \frac{1}{2(0.457 + 0.494/N^*(t))} \left[\frac{1}{|N(t)|} \sum_{(i,j)\in N^*(t)} |Z(s_i) - Z(s_j)|^{1/2} \right]^4,$$

- Estimador empírico de Cressie y Hawkins 2:

$$\hat{\gamma}(t) = \frac{1}{2(0.457 + 0.494/N^*(t))} \left[\text{Mediana}\left(|Z(s_i) - Z(s_j)|^{1/2} / s_i - s_j \in N^*(t)\right)\right]^4.$$

Una alternativa a estos estimadores empíricos se puede encontrar en García-Soidán *et al.* (2004). Sea $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ un proceso intrínseco e isotrópico. Continuando con la idea de promediar los valores de las diferencias cuadráticas $(Z(s_i) - Z(s_j))^2$, se puede estimar el semivariograma utilizando la media ponderada de dichas diferencias, de la forma siguiente:

$$\hat{\gamma}(t) = \frac{\sum_{i,j=1}^{n} w_{i,j}(t) \left(Z(s_i) - Z(s_j)\right)^2}{2 \sum_{i,j=1}^{n} w_{i,j}(t)}, \tag{1.7}$$

donde $w_{i,j}(t) \geq 0$, $\forall i, j$, y $\sum_{i,j=1}^{n} w_{i,j}(t) > 0$. Si en (1.7) se toma $w_{i,j}(t) = I_{\{\|s_i - s_j\| = t\}}$ se tiene el estimador empírico clásico.
Adaptando el estimador de Nadaraya-Watson empleado en regresión, García-Soidán *et al.* (2004) proponen tomar

$$w_{i,j}(t) = K\left(\frac{\|s_i - s_j\| - t}{h}\right),$$

donde K es una función de densidad simétrica y h el parámetro ventana. Al estimador del semivariograma resultante, $\hat{\gamma}(t)$, se le denominará estimador de Nadaraya-Watson.

De forma análoga se puede definir el estimador lineal local del semivariograma (García-Soidán *et al.* 2003) tomando como pesos los valores

$$w_{i,j}(t) = K\left(\frac{\|s_i - s_j\| - t}{h}\right) \sum_{k,l=1}^{n} K\left(\frac{\|s_k - s_l\| - t}{h}\right) (\|s_i - s_j\| - \|s_k - s_l\|),$$

donde K es una función de densidad simétrica y h el parámetro ventana. Dicho estimador es asintóticamente insesgado y presenta la ventaja adicional de que proporciona un valor de la primera derivada del semivariograma, lo que facilita la elección de un modelo paramétrico válido en el segundo paso de su caracterización.

Por último, partiendo del estimador del variograma (1.4) y bajo las suposiciones de estacionariedad intrínseca e isotropía, Yu *et al.* (2006) proponen aproximar $\gamma(\cdot)$ mediante el estimador de vecinos más próximos variables (en inglés *variable nearest-neighbour estimator*):

$$\hat{\gamma}(t) = \frac{\sum_{i<j} \frac{1}{\delta_0(\|s_i - s_j\|)} K\left(\frac{t - \|s_i - s_j\|}{\delta \delta_0(\|s_i - s_j\|)}\right) \left(Z(s_i) - Z(s_j)\right)^2}{\sum_{i<j} \frac{1}{\delta_0(\|s_i - s_j\|)} K\left(\frac{t - \|s_i - s_j\|}{\delta \delta_0(\|s_i - s_j\|)}\right)},$$

donde δ es una constante positiva de suavizado y $\delta_0(\cdot) > 0$ una función de suavizado.

Puesto que la hipótesis de isotropía simplifica notablemente el modelado de la dependencia espacial, al establecer que la variabilidad depende sólo de la distancia y no de la dirección de los datos, las situaciones reales tratan de resolverse en la medida de lo posible bajo esta suposición. Sin embargo, en

la práctica existen numerosas situaciones en las que la variación es anisotrópica, que podrían resolverse ajustando un variograma isotrópico para cada dirección de un conjunto de direcciones seleccionadas o empleando versiones paramétricas anisotrópicas.

Por consiguiente, antes de comenzar el procedimiento de caracterización del variograma sería conveniente determinar si se puede admitir la hipótesis de isotropía o, en caso contrario, determinar qué tipo de anisotropía presentan los datos. Una aproximación a este problema consiste en dibujar las líneas del *isovariograma*, donde se representan las distancias para las cuales el variograma alcanza los mismos valores en distintas direcciones prefijadas de antemano. Si las líneas del isovariograma se corresponden aproximadamente con círculos concéntricos, entonces se puede admitir la condición de isotropía. En caso de que dé lugar a elipses concéntricas, se hablará de anisotropía geométrica.

Definición 1.9. Se dice que un proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ presenta *anisotropía geométrica* (o afín) si su variograma es de la forma:

$$2\gamma(t) = 2\gamma_0(\|At\|), t \in \mathbb{R}^d,$$

siendo γ_0 un variograma isotrópico y A una matriz $d \times d$ que representa una determinada transformación lineal en $\mathbb{R}^d$.

La anisotropía geométrica aparece cuando el umbral permanece constante mientras que el rango varía con la dirección. Además, es una generalización de la isotropía, tomando como A la matriz identidad.

Un tipo de anisotropía más complicado aparece cuando el umbral varía con la dirección. En ese caso se hablará de anisotropía zonal.

Definición 1.10. Se dice que un proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ presenta *anisotropía zonal* si su variograma es de la forma:

$$2\gamma(t) = \sum_{i=1}^{n} 2\gamma_0(\|A_i t\|), t \in \mathbb{R}^d,$$

siendo γ_0 un variograma isotrópico y A_i, $i \in \{1...n\}$, una matriz $d \times d$.

El tratamiento de la anisotropía zonal es más complejo, si bien suele ajustarse un variograma isotrópico en cada dirección, aunque de modo distinto para cada una, en función de que el alcance varíe o no con la dirección.

Una vez resuelta la cuestión sobre el carácter isotrópico o anisotrópico del variograma, se procedería a aplicar el método de caracterización mediante tres pasos, indicado anteriormente. No obstante, con independencia del carácter establecido, el estimador no paramétrico que se obtiene para el variograma en el primer paso generalmente no satisface la condición de ser condicionalmente definido negativo. Y esto podría suponer un problema en la práctica para su aplicación a la predicción, ya que la propiedad indicada garantiza la existencia de solución de los sistemas kriging lineales. Para solucionar lo anterior, se puede aplicar el segundo paso del procedimiento de estimación del variograma, que consiste en la selección de un modelo válido y el ajuste de los parámetros del mismo utilizando los datos disponibles.

A continuación se introducirán los modelos de semivariograma paramétricos más empleados en la práctica. Todos los semivariogramas que se mostrarán son isotrópicos, ya que constituyen el punto de arranque sobre el que contruir modelos más complejos. En la notación utilizada en las parametrizaciones $c_0 \geq 0$ representa el efecto pepita, $c_1 \geq 0$ el umbral parcial, en caso de que exista, y $c_2 > 0$ el rango o alcance, si existe.

- Modelo de efecto pepita:

Constituye el modelo más sencillo de semivariograma, indicando la ausencia de correlación espacial. Viene dado por la expresión:

$$\gamma(t)_{c_0} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0, \text{ si } t > 0. \end{cases}$$

- Modelo lineal:

$$\gamma(t)_{c_0,c_1} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 t, \text{ si } t > 0. \end{cases}$$

Se trata de un modelo no acotado, sin estacionaridad de segundo orden, que no presenta mucha utilidad en la práctica.

- Modelo esférico:

$$\gamma(t)_{c_0,c_1,c_2} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 \left(1.5 \frac{t}{c_2} - 0.5 \left(\frac{t}{c_2}\right)^3\right), \text{ si } 0 < t < c_2; \\ \\ c_0 + c_1, \text{ si } t \geq c_2. \end{cases}$$

Este modelo transitivo presenta un crecimiento rápido cerca del origen con incrementos marginales decrecientes para distancias grandes, de forma que para distancias superiores al rango los incrementos son nulos. Sólo es válido en $\mathbb{R}^d$ con $d = 1, 2, 3$.

- Modelo exponencial:

$$\gamma(t)_{c_0,c_1,c_2} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 \left(1 - \exp\left(-\frac{3t}{c_2}\right)\right), \text{ si } t > 0. \end{cases}$$

Este modelo se aplica cuando la dependencia espacial tiene un crecimiento exponencial respecto a la distancia entre las observaciones (Giraldo, 2002). Al igual que el modelo esférico es un modelo transitivo. Muy empleado en la práctica por su sencillez y rapidez a la hora de realizar simulaciones.

- Modelo gaussiano:

$$\gamma(t)_{c_0,c_1,c_2} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 \left(1 - \exp\left(-\frac{3t^2}{c_2^2}\right)\right), \text{ si } t > 0. \end{cases}$$

Al igual que en el exponencial, la dependencia espacial para este modelo transitivo desaparece para distancias que tienden a infinito.

- Modelo potencial:

$$\gamma(t)_{c_0,c_1,\lambda} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 t^{\lambda}, \text{ si } t > 0, \end{cases}$$

donde $0 \leq \lambda < 2$ es un factor de crecimiento que determina el comportamiento del semivariograma cerca del origen. Este modelo se engloba dentro de los intransitivos.

- Modelo exponencial-potencial:

$$\gamma(t)_{c_0,c_1,c_2,\lambda} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 \left(1 - \exp\left(-3\left(\frac{t}{c_2}\right)^{\lambda}\right)\right), \text{ si } t > 0, \end{cases}$$

donde $0 \leq \lambda < 2$.

- Modelo oscilatorio:

$$\gamma(t)_{c_0,c_1,c_2} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 \left(1 - \frac{\sin(c_2 t)}{c_2 t}\right), \text{ si } t > 0. \end{cases}$$

Este modelo también es conocido como modelo de efecto hoyo o de efecto agujero. Sólo es válido en $\mathbb{R}^d$ con $d = 1, 2, 3$. Es un ejemplo de semivariograma que no es monótonamente creciente, útil para procesos con un comportamiento periódico donde existe una sucesión entre zonas ricas y pobres (Martínez Ruiz, 2008).

- Modelo de Matern (o K-Bessel):

$$\gamma(t)_{c_0,c_1,c_2,\nu} = \begin{cases} 0, \text{ si } t = 0; \\ \\ c_0 + c_1 \left(1 - \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{t}{c_2}\right)^{\nu} K_{\nu}\left(\frac{t}{c_2}\right)\right), \text{ si } t > 0, \end{cases}$$

donde K_ν es una función de Bessel de segunda clase y orden ν modificada. En la actualidad es uno de los más utilizados gracias a su flexibilidad a la hora de modelizar grandes cantidades de datos experimentales. Si se toma $\nu = \frac{1}{2}$ o $\nu = \infty$, entonces el modelo de Matern coincide con el exponencial y con el gaussiano, respectivamente.

Una vez seleccionado un variograma paramétrico válido para la caracterización de la estructura de dependencia, el siguiente paso será ajustar los parámetros a partir de los datos disponibles, tomando como referencia el estimador empírico obtenido. Para ello, se han propuesto en la literatura diversos criterios de ajuste, en estas páginas se incluirán los métodos de máxima verosimilitud, máxima verosimilitud restringida, mínimos cuadrados ordinarios, mínimos cuadrados ponderados y mínimos cuadrados generalizados.

- Máxima verosimilitud:

  Se basa en suponer que la distribución del proceso intrínseco $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ es normal o gaussiana. De esta forma para las localizaciones $s_1, ..., s_n$:

$$Z = (Z(s_1)...Z(s_n))' \sim N(\vec{\mu}, \Sigma), \text{ con } \vec{\mu} = (\mu...\mu)' \text{ y } \Sigma = (\sigma_{ij}),$$

donde $\sigma_{ij} = C(s_i - s_j) = \sigma^2 - \gamma(s_i - s_j)_{c_0,c_1,c_2}$.

La estimación por máxima verosimilitud consiste en obtener de forma simultánea los estimadores de μ, σ, c_0, c_1 y c_2 que maximizan la función de distribución conjunta normal multivariante:

$$(2\pi)^{-n/2}|\Sigma_{\sigma,c_0,c_1,c_2}|^{-1/2}\exp\left[-\frac{1}{2}(Z-\mu)'\Sigma^{-1}_{\sigma,c_0,c_1,c_2}(Z-\mu)\right]. \tag{1.8}$$

Aplicando logaritmos, maximizar la función (1.8) equivale a minimizar la función:

$$n\ln(2\pi) + \ln|\Sigma_{\sigma,c_0,c_1,c_2}| + (Z-\mu)'\Sigma^{-1}_{\sigma,c_0,c_1,c_2}(Z-\mu). \tag{1.9}$$

Minimizar la función (1.9) puede resultar computacionalmente costoso. Otro problema adicional es la posible multimodalidad de la función (Mardia y Watkins, 1989).

Los estimadores obtenidos por el método de máxima verosimilitud son asintóticamente consistentes, pero pueden presentar un sesgo considerable para muestras pequeñas y especialmente cuando la tendencia no es constante. Este problema se resuelve utilizando una variante de este método de ajuste, que se presenta a continuación.

- Máxima verosimilitud restringida:

  La idea de este método de ajuste consiste en filtrar los datos de forma que la distribución conjunta no dependa de μ ni σ^2. De esta forma, suponiendo de nuevo que el proceso intrínseco $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ es normal o gaussiano, se define:

$$Y = \Lambda Z, \text{ con } \Lambda = \begin{pmatrix} 1 & -1 & & 0 \\ & \ddots & \ddots & \\ 0 & & 1 & -1 \end{pmatrix},$$

  por consiguiente,

$$Y = \Lambda Z \sim N(0, \Lambda\Sigma\Lambda'), \text{ con } \Sigma = (\sigma_{ij}) \text{ y } \sigma_{ij} = C(s_i - s_j) = \sigma^2 - \gamma(s_i - s_j)_{c_0,c_1,c_2}.$$

  Procediendo ahora de manera análoga al caso anterior, se trataría de maximizar la función:

$$(2\pi)^{-n/2}|(\Lambda\Sigma\Lambda')_{c_0,c_1,c_2}|^{-1/2}\exp\left[-\frac{1}{2}Y'(\Lambda\Sigma\Lambda')^{-1}_{c_0,c_1,c_2}Y\right]. \tag{1.10}$$

  Equivalentemente, aplicando logaritmos a (1.10), se trataría de minimizar:

$$n\ln(2\pi) + \ln|(\Lambda\Sigma\Lambda')_{c_0,c_1,c_2}| + Y'(\Lambda\Sigma\Lambda')^{-1}_{c_0,c_1,c_2}Y.$$

Aunque los procesos de máxima verosimilitud se basan en la hipótesis de que el proceso intrínseco sigue una distribución normal multivariante, los métodos son lo suficientemente robustos para que la distribución del error del modelo no se vea demasiado afectada por la violación de la hipótesis.

- Mínimos cuadrados:

  Supóngase que para ciertos valores t_i, $1 \le i \le k$, se han obtenido los valores del semivariograma empírico $\hat{\gamma}(t_i)$, y sea $\gamma(t_i)_{c_0,c_1,c_2}$ el variograma teórico. Siguiendo las recomendaciones de Journel y Huijbregts (2003), solamente se consideran en el ajuste distancias menores o iguales que la mitad de la distancia máxima, y tales que $|N(t_i)| \ge 30$, donde $|N(t_i)|$ es el número de pares a una distancia t_i.

El método de ajuste por mínimos cuadrados se basa en minimizar la función:

$$(\vec{\hat{\gamma}} - \vec{\gamma}_{c_0,c_1,c_2})' V_{c_0,c_1,c_2} (\vec{\hat{\gamma}} - \vec{\gamma}_{c_0,c_1,c_2}), \tag{1.11}$$

donde $\vec{\hat{\gamma}} = (\hat{\gamma}(t_1)...\hat{\gamma}(t_k))'$ y $\vec{\gamma}_{c_0,c_1,c_2} = (\gamma(t_1)_{c_0,c_1,c_2}...\gamma(t_k)'_{c_0,c_1,c_2})$.

Según la matriz V_{c_0,c_1,c_2} elegida se obtienen distintos criterios de bondad de ajuste:

- Mínimos cuadrados ordinarios:
  La matriz V_{c_0,c_1,c_2} es la matriz identidad. De esta forma la función (1.11) se reduce a:

  $$(\vec{\hat{\gamma}} - \vec{\gamma}_{c_0,c_1,c_2})' (\vec{\hat{\gamma}} - \vec{\gamma}_{c_0,c_1,c_2}),$$

  y el objetivo sería encontrar c_0, c_1, c_2 tales que la minimicen.
  Un problema que presenta este método es que en este caso las estimaciones están correladas y tienen varianzas diferentes.
- Mínimos cuadrados ponderados:
  En este caso la matriz V_{c_0,c_1,c_2} toma la forma,

  $$V_{c_0,c_1,c_2} = \begin{pmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_k \end{pmatrix},$$

  donde los pesos w_i son inversamente proporcionales a las varianzas de $\hat{\gamma}(h_i)$.
  Bajo la hipótesis de que el proceso es gaussiano y de que las estimaciones son incorreladas se puede aproximar la varianza del estimador empírico de la forma,

  $$Var\left[\hat{\gamma}(t_i)\right] \approx \frac{\gamma(t_i)^2}{|N(t_i)|},$$

  donde $|N(t_i)|$ es el número de elementos a una distancia t_i.
  Por consiguiente, se tomarán los pesos,

  $$w_i \approx \frac{|N(t_i)|}{\gamma(t_i)^2_{c_0,c_1,c_2}}.$$

  La estimación por mínimos cuadrados ponderados es la más empleada en la práctica debido a la facilidad de su implementación y a las ventajas computacionales que presenta (Martínez Ruiz, 2008).
- Mínimos cuadrados generalizados:
  En este caso, en vez de considerar una matriz diagonal, se toma V_{c_0,c_1,c_2} igual o proporcional a la inversa de la matriz de varianzas-covarianzas de $\hat{\gamma}$, cuyos términos a su vez dependen de los parámetros a estimar.

Una vez ajustado el variograma es necesario realizar una diagnosis del mismo. Una manera de determinar si los datos son compatibles con el variograma obtenido, es aplicar la técnica de validación cruzada (en inglés *cross-validation*).

Sea $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ un proceso aleatorio espacial para el cual se conocen los valores $Z(s_1), ..., Z(s_n)$ en las localizaciones $s_1, ..., s_n$; el método de validación cruzada consiste en dejar un dato fuera, $Z(s_i)$, y utilizar los restantes para construir un predictor de $Z(s_i)$ y un estimador del

error medio, $\sigma^2(s_i)$, mediante algún mecanismo de predicción, como el kriging, que se presentará en la siguiente sección. Se denotarán por $\hat{Z}_{-i}(s_i)$ y $\hat{\sigma}^2_{-i}(s_i)$, respectivamente, al predictor y al error obtenidos en la localización s_i utilizando todos los datos observados salvo el i-ésimo. Calculando los correspondientes valores para todas las localizaciones $s_1, ..., s_n$, la proximidad de los predictores a los valores de la muestra permitirá diagnosticar la validez del modelo ajustado de la forma:

- Analizar la proximidad de la media, $\frac{1}{n}\sum_{i=1}^{n}(Z(s_i) - \hat{Z}_{-i}(s_i))\hat{\sigma}^{-1}_{-i}(s_i)$, al valor 0.
- Analizar la proximidad de la desviación, $\left[\frac{1}{n}\sum_{i=1}^{n}(Z(s_i) - \hat{Z}_{-i}(s_i))^2\hat{\sigma}^{-2}_{-i}(s_i)\right]^{1/2}$, al valor 1.
- Dibujar el histograma o el diagrama de tallo y hojas de los valores $(Z(s_i) - \hat{Z}_{-i}(s_i))\hat{\sigma}^{-1}_{-i}(s_i)$ y analizar su desviación de la curva normal o la presencia de datos atípicos.

1.3. Predicción espacial

Uno de los principales objetivos de la estadística espacial es la reconstrucción de un fenómeno sobre la región de observación a partir de un número finito de datos. Con este objetivo, se definen los métodos kriging como algoritmos de predicción de mínimo error en media cuadrática que tienen en cuenta la estructura de segundo orden del proceso.

En función del tipo de predictor se puede distinguir entre kriging lineal, entre los que se encuentran el simple, ordinario y el universal, o kriging no lineal, entre los que se encuentra el kriging indicador que será presentado en el siguiente capítulo.

En la presente sección se hará una breve presentación sobre los tres tipos de kriging lineal mencionados.

Supóngase que $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ es un proceso aleatorio espacial para el cual se conocen los valores $Z(s_1), ..., Z(s_n)$ en las localizaciones $s_1, ..., s_n$. Se trata de obtener un predictor de $Z(s)$, que se denotará por $\hat{Z}(s)$, tal que:

a) Sea insesgado, es decir,
$$\mathbb{E}(\hat{Z}(s)) = \mathbb{E}(Z(s)) = \mu(s). \tag{1.12}$$

b) Minimice el error en media cuadrática de predicción, también designado como varianza de predicción, σ^2,
$$\mathbb{E}[(\hat{Z}(s) - Z(s))^2]. \tag{1.13}$$

Dependiendo de las suposiciones acerca de la tendencia del proceso $\mu(\cdot)$, se presentarán tres métodos de kriging lineal para variables unidimensionales:

1. Kriging simple (KS), en el que se supone que la media del proceso $\mu(s)$ es conocida $\forall s \in D$.
2. Kriging ordinario (KO), en el que se supone que la media del proceso es desconocida y constante, $\mu(s) = \mu$ para todo $s \in D$.
3. Kriging universal (KU), en el que se supone que la media del proceso es desconocida y no constante, pero puede modelizarse como combinación lineal de un conjunto de funciones conocidas f_i,
$$\mu(s) = \sum_{i=0}^{k} f_i(s)a_i, \forall s \in D,$$
donde a_i son parámetros reales desconocidos. Típicamente se toma $f_0(s) = 1$ para todo $s \in D$.

Los predictores kriging son interpoladores exactos (suponiendo que no hay error de medida) y se calculan de forma que son los mejores predictores lineales insesgados. Además, si el proceso es gaussiano entonces coinciden con el mejor predictor posible

1.3.1. Kriging simple

Supongamos que el proceso $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ admite una descomposición de la forma:

$$Z(s) = \mu(s) + Y(s),$$

donde la función $\mu(\cdot)$ es conocida e $Y(\cdot)$ es un proceso estacionario de segundo orden de media 0 y función de convarianza $C(\cdot)$ conocida o estimada.

Se desea construir un predictor lineal óptimo de la forma:

$$\hat{Z}(s) = \sum_{i=1}^{n} \lambda_i Z(s_i) + \lambda_0,$$

verificando las condiciones (1.12) y (1.13). Esto se traduce en que

1. $\mathbb{E}[\hat{Z}(s)] = \sum_{i=1}^{n} \lambda_i \mathbb{E}[Z(s)] + \lambda_0 = \sum_{i=1}^{n} \lambda_i \mu(s_i) + \lambda_0 = \mu(s) \Rightarrow \lambda_0 = \mu(s) - \sum_{i=1}^{n} \lambda_i \mu(s_i)$.

   Por lo tanto el predictor es de la forma,

$$\hat{Z}(s) = \mu(s) + \sum_{i=1}^{n} \lambda_i (Z(s_i) - \mu(s_i)) = \mu(s) + \sum_{i=1}^{n} \lambda_i Y(s_i).$$

2. $\mathbb{E}\left[(\hat{Z}(s) - Z(s))^2\right] = \mathbb{E}\left[(\sum_{i=1}^{n} \lambda_i Z(s_i) + \lambda_0 - Z(s))^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \lambda_i (Z(s_i) - \mu(s_i)) - (Z(s) - \mu(s))\right)^2\right] =$

$$\mathbb{E}\left[\left(\sum_{i=1}^{n} \lambda_i Y(s_i) - Y(s)\right)^2\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n} \lambda_i C(s_i - s) + C(0).$$

Derivando respecto de λ_i e igualando a cero se obtiene:

$$2\sum_{j=1}^{n} \lambda_j C(s_i - s_j) - 2C(s_i - s) = 0 \Leftrightarrow \sum_{j=1}^{n} \lambda_j C(s_i - s_j) = C(s_i - s).$$

Escribiendo el problema en forma matricial como $A\lambda = B$, se obtiene como solución $\lambda = A^{-1}B$, siempre que A sea no singular, y como varianza de predicción $\sigma^2_{KS} = \sigma^2 - \lambda' B$, donde:

$$A = \begin{pmatrix} C(0) & C(s_1 - s_2) & \dots & C(s_1 - s_n) \\ C(s_2 - s_1) & C(0) & \dots & C(s_2 - s_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(s_n - s_1) & C(s_n - s_2) & \dots & C(0) \end{pmatrix}$$

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}, \quad B = \begin{pmatrix} C(s_1 - s) \\ C(s_2 - s) \\ \vdots \\ C(s_n - s) \end{pmatrix}.$$

1.3.2. Kriging ordinario

Supongamos que el proceso $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ admite una descomposición de la forma:

$$Z(s) = \mu + Y(s), \tag{1.14}$$

donde la tendencia del proceso μ es desconocida y constante, e $Y(\cdot)$ es un proceso estacionario de segundo orden de media 0 y función de covarianza C.

Se desea construir un predictor lineal óptimo de la forma:

$$\hat{Z}(s) = \sum_{i=1}^{n} \lambda_i Z(s_i), \tag{1.15}$$

verificando las condiciones (1.12) y (1.13). Esto se traduce en que

1. $\mathbb{E}[\hat{Z}(s)] = \sum_{i=1}^{n} \lambda_i \mathbb{E}[Z(s)] = \sum_{i=1}^{n} \lambda_i \mu = \mu$. Una condición suficiente para lo anterior es que $\sum_{i=1}^{n} \lambda_i = 1$.

2. Minimice el error cuadrático medio de predicción, sujeto a la condición $\sum_{i=1}^{n} \lambda_i = 1$. Lo que equivale, mediante le método de multiplicadores de Lagrange, a minimizar:

$$\mathbb{E}\left[(\hat{Z}(s) - Z(s))^2\right] + m_0(\sum_{i=1}^{n} \lambda_i - 1) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n} \lambda_i C(s_i - s) + C(0) + m_0(\sum_{i=1}^{n} \lambda_i - 1).$$

Derivando respecto a λ_i y m_0 e igualando a cero se obtiene el sistema de ecuaciones:

$$\begin{aligned} -2\sum_{j=1}^{n} \lambda_j C(s_i - s_j) + 2C(s_i - s) + m_0 &= 0 \\ \sum_{i=1}^{n} \lambda_i - 1 &= 0. \end{aligned}$$

Tomando $m = -m_0/2$ el sistema anterior escrito en forma matricial queda:

$$\begin{pmatrix} \Sigma & d \\ d' & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ m \end{pmatrix} = \begin{pmatrix} b \\ 1 \end{pmatrix} \Leftrightarrow AX = B, \tag{1.16}$$

donde:

$$\Sigma = \begin{pmatrix} C(0) & C(s_1 - s_2) & \dots & C(s_1 - s_n) \\ C(s_2 - s_1) & C(0) & \dots & C(s_2 - s_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(s_n - s_1) & C(s_n - s_2) & \dots & C(0) \end{pmatrix}$$

$$d = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}, \quad b = \begin{pmatrix} C(s_1 - s) \\ C(s_2 - s) \\ \vdots \\ C(s_n - s) \end{pmatrix}.$$

Obteniéndose como solución del sistema $X = A^{-1}B$ y como varianza de la predicción $\sigma^2_{KO} = \sigma^2 - X'B$, donde X' denota la traspuesta de la matriz X.

Bajo el modelo (1.14), si se supone que el proceso $Y(\cdot)$ es intrínsecamente estacionario, la construcción del predictor (1.15) se realizaría de forma similar. En este caso se tendría un sistema de ecuaciones análogo a (1.16), obtenido al sustituir la función de covarianza C por el semivariograma γ, tanto en la matriz Σ como en el vector b. La solución sería también de la forma $X = A^{-1}B$, aunque no ocurriría así con la varianza resultante, que quedaría como $\sigma^2_{KO} = X'B$.

1.3.3. Kriging universal

Supongamos que el proceso $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ admite una descomposición de la forma:

$$Z(s) = \mu(s) + Y(s),$$

donde la función $\mu(\cdot)$ es desconocida pero se puede expresar como $\mu(s) = \sum_{k=0}^{K} f_k(s)a_k$, siendo f_k funciones conocidas y a_k parámetros reales desconocidos, e $Y(s)$ es un proceso estacionario de segundo orden de media 0 y función de covarianza $C(\cdot)$.

Se desea construir un predictor lineal óptimo de la forma:

$$\hat{Z}(s) = \sum_{i=1}^{n} \lambda_i Z(s_i),$$

verificando las condiciones (1.12) y (1.13). Esto se traduce en que

1. $\mathbb{E}(\hat{Z}(s)) = \sum_{i=1}^{n} \lambda_i \sum_{k=0}^{K} f_k(s_i)a_k = \sum_{k=0}^{K} f_k(s)a_k \Rightarrow \sum_{i=1}^{n} \lambda_i f_k(s_i) = f_k(s)$.

2. Minimice el error cuadrático medio de predicción sujeto a la condición $\sum_{i=1}^{n} \lambda_i f_k(s_i) = f_k(s)$. Lo que equivale, mediante le método de multiplicadores de Lagrange, a minimizar:

$$\mathbb{E}\left[(\hat{Z}(s) - Z(s))^2\right] + \sum_{k=1}^{K} m_k(\sum_{i=1}^{n} \lambda_i f_k(s_i) - f_k(s)) =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n} \lambda_i C(s_i - s) + C(0) + \sum_{k=1}^{K} m_k(\sum_{i=1}^{n} \lambda_i f_k(s_i) - f_k(s)).$$

Derivando e igualando a cero se obtiene el sistema matricial:

$$\begin{pmatrix} \Sigma & d_0 & \dots & d_K \\ d_0' & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ d_K' & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ M \end{pmatrix} = \begin{pmatrix} b \\ f_0(s) \\ \vdots \\ f_K(s) \end{pmatrix} \Leftrightarrow AX = B, \tag{1.17}$$

donde:

$$\Sigma = \begin{pmatrix} C(0) & C(s_1 - s_2) & \dots & C(s_1 - s_n) \\ C(s_2 - s_1) & C(0) & \dots & C(s_2 - s_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(s_n - s_1) & C(s_n - s_2) & \dots & C(0) \end{pmatrix}$$

$$d_k = \begin{pmatrix} f_k(s_1) \\ \vdots \\ f_k(s_n) \end{pmatrix}, M = \begin{pmatrix} m_0 \\ \vdots \\ m_k) \end{pmatrix}, \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}, \quad b = \begin{pmatrix} C(s_1 - s) \\ C(s_2 - s) \\ \vdots \\ C(s_n - s) \end{pmatrix}$$

Obteniéndose como solución del sistema $X = A^{-1}B$ y como varianza de la predicción $\sigma^2_{KU} = \sigma^2 - X'B$.

De forma similar a la planteada al final de la sección anterior, se podría proceder a la construcción del predictor kriging universal suponiendo estacionariedad intrínseca, en vez de estacionariedad de segundo orden, en el proceso $Y(\cdot)$.

Capítulo 2

Estimación de la función de distribución

En este capítulo se presentará la estimación de la función de distribución mediante el variograma indicador. Además se incluirá un estimador no paramétrico de la función de distribución, basado en la utilización de una media ponderada de funciones indicadoras, cuyos pesos dependerán de la separación entre las posiciones observadas y la localización objeto de estudio.

2.1. La función de distribución y el variograma indicador

En el capítulo anterior se ha presentado la estimación de la estructura de dependencia de un proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$, necesaria para proceder a la predicción del proceso en una nueva localización $s \in D$, mediante las técnicas kriging. Sin embargo, en ocasiones interesa estimar la probabilidad de que la variable $Z(s)$ sea mayor (o menor) que un valor determinado $x \in \mathbb{R}$, para lo cual, se requiere estimar la función de distribución del proceso espacial. Esta estimación permitiría, por ejemplo, construir mapas de riesgo, en los que figure la probabilidad de que una determinada substancia química sobrepase los niveles admitidos para el consumo humano.

Una manera de llevar a cabo la estimación de la distribución de un proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ se basa en utilizar la función indicadora, $I_Z(s,x)$, que para un determinado valor $x \in \mathbb{R}$ toma la forma:

$$I_Z(s,x) = I(Z(s) \leq x) = \begin{cases} 0, & \text{si } Z(s) > x; \\ 1, & \text{si } Z(s) \leq x. \end{cases}$$

Pues se tiene que,

$$F_s(x) = \mathbb{P}(Z(s) \leq x) = 1 \cdot \mathbb{P}(Z(s) \leq x) + 0 \cdot \mathbb{P}(Z(s) > x) = \mathbb{E}\left[I_Z(s,x)\right].$$

Con la variable $I_Z(s,x)$, se puede proceder de forma análoga a la planteada para la variable $Z(s)$ en el capítulo anterior, tanto para la caracterización de su estructura de dependencia como para la predicción de valores en posiciones no muestreadas. Para ello, se podría utilizar la covarianza o el semivariograma, en su versión indicadora. En este trabajo se utilizará la segunda función, si bien para la función de covarianza se procedería de forma similar.

Teniendo presente lo anterior, el *semivariograma indicador* $\gamma_I(h,x)$ se define como:

$$\gamma_I(s-s',x) = \gamma_{I,x}(s-s') = \frac{1}{2}Var[I_Z(s,x) - I_Z(s',x)] = \frac{1}{2}\mathbb{E}[(I_Z(s,x) - I_Z(s',x))^2], \forall s,s' \in D, \ \forall x \in \mathbb{R}$$

Dado $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ un proceso aleatorio espacial para el cual se conocen los valores $Z(s_1),...,Z(s_n)$ en las localizaciones $s_1,...,s_n$, la estimación del semivariograma se hará de manera

análoga al capítulo anterior, pero de nuevo aplicando los resultados sobre $I_Z(s,x)$. Los variogramas indicadores $\gamma_I(h,x)$ son estimados para cada valor de corte dado $x \in \mathbb{R}$.

Por consiguiente, los estimadores no paramétricos quedarían de la forma:

- Estimador empírico de Matheron:

$$\hat{\gamma}_I(t,x) = \frac{1}{2|N(t)|} \sum_{(i,j)\in N(t)} [I_Z(s_i,x) - I_Z(s_j,x)]^2,$$

donde $N(t) = \{(i,j)/s_i - s_j = t\}$ y $|N(t)|$ es su cardinal.

- Estimadores robustos de Cressie y Hawkins:

$$\hat{\gamma}_I(t,x) = \frac{1}{2(0.457 + 0.494/N(t))} \left[\frac{1}{|N(t)|} \sum_{(i,j)\in N(t)} |I_Z(s_i,x) - I_Z(s_j,x)|^{1/2} \right]^4,$$

o

$$\hat{\gamma}_I(t,x) = \frac{1}{2(0.457 + 0.494/N(t))} \left[\text{Mediana}\left(|I_Z(s_i,x) - I_Z(s_j,x)|^{1/2} / s_i - s_j \in N(t) \right) \right]^4,$$

- Alternativas no paramétricas:

Se estima el semivariograma empleando medias ponderadas.

$$\hat{\gamma}_I(t,x) = \frac{\sum_{i,j=1}^{n} w_{i,j}(t) \left(I_Z(s_i,x) - I_Z(s_j,x)\right)^2}{2 \sum_{i,j=1}^{n} w_{i,j}(t)},$$

obteniendo los estimadores no paramétricos:

1. Estimador de Nadaraya-Watson tomando $w_{i,j}(t) = K\left(\frac{(s_i - s_j) - t}{h}\right)$, donde K es una función de densidad simétrica y h el parámetro ventana.
2. Estimador lineal local del semivariograma tomando como pesos los valores

$$w_{i,j}(t) = K\left(\frac{(s_i - s_j) - t}{h}\right) \sum_{k,l=1}^{n} K\left(\frac{(s_k - s_l) - t}{h}\right) ((s_i - s_j) - (s_k - s_l)),$$

donde K es una función de densidad simétrica y h el parámetro ventana.

3. El estimador de Yu, Mateu y Porcu:

$$\hat{\gamma}_I(t,x) = \frac{\sum_{i<j} \frac{1}{\delta_0((s_i - s_j))} K\left(\frac{t - (s_i - s_j)}{\delta\delta_0((s_i - s_j))}\right) \left(I_Z(s_i,x) - I_Z(s_j,x)\right)^2}{\sum_{i<j} \frac{1}{\delta_0((s_i - s_j))} K\left(\frac{t - (s_i - s_j)}{\delta\delta_0((s_i - s_j))}\right)},$$

donde δ es una constante positiva de suavizado y $\delta_0(\cdot) > 0$ una función de suavizado.

De nuevo los estimadores anteriores presentan el inconveniente de que no cumplen necesariamente la propiedad de ser condicionalmente definidos positivos. De ahí que se recurra también en este caso a la selección de una familia de variogramas válidos y al ajuste de sus parámetros para obtener un

estimador final válido del variograma indicador, siguiendo lo expuesto en el Capítulo 2, Sección 2.

A partir de la obtención de un estimador del variograma indicador, puede plantearse como un problema de kriging lineal la estimación de la función de distribución $F_s(x)$ bajo ciertas condiciones, en el que las predicciones serán sobre la variable indicadora $I_Z(s,x)$.

Dado $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ un proceso aleatorio espacial para el cual se conocen los valores $Z(s_1),...,Z(s_n)$ en las localizaciones $s_1,...,s_n$, y fijado un valor de corte $x \in \mathbb{R}$, podría aproximarse $F_s(x)$ mediante kriging ordinario de la forma:

$$\hat{F}_s(x) = \hat{I}_Z(s,x) = \sum_{i=1}^{n} \lambda_i I_Z(s_i,x). \tag{2.1}$$

De nuevo deben verificarse las condiciones de insesgadez y minimización del error cuadrático medio de predicción, de forma que:

1. $\mathbb{E}[\hat{I}(s,x)] = \mathbb{E}[I(s,x)] = F_s(x)$.
2. Los valores λ_i son tales que minimicen el error cuadrático medio de predicción, también designado como varianza de predicción, $\mathbb{E}\left[\left(\hat{I}(s,x) - I(s,x)\right)^2\right]$.

La aplicación del kriging ordinario precisaba de la imposición de que la media del proceso fuese constante. En este caso ésto se traduce en que $\mathbb{E}[I(s,x)] = \mathbb{E}[I(s',x)]$, para todo $s, s' \in D$; y por consiguiente $F_s(x) = F_{s'}(x) = F(x)$ para $x \in D$. Lo anterior se verifica, por ejemplo, si el proceso es estrictamente estacionario. A partir de esta hipótesis y de la condición de insesgadez se deduce que:

$$\mathbb{E}\left[\hat{I}(s,x)\right] = \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i I_Z(s_i,x)\right] = \sum_{i=1}^{n} \lambda_i \mathbb{E}\left[I_Z(s_i,x)\right] = \sum_{i=1}^{n} \lambda_i F(x) = F(x) \Rightarrow$$

$$\Rightarrow \sum_{i=1}^{n} \lambda_i = 1.$$

Por lo tanto la estimación de la función de distribución $F(\cdot)$ consiste en minimizar el error de predicción sujeto a $\sum_{i=1}^{n} \lambda_i = 1$. Empleando multiplicadores de Lagrange, se trataría de encontrar valores λ_i, $i \in \{1...n\}$ y m_0 tales que minimicen:

$$\mathbb{E}\left[\left(\hat{I}(s,x) - I(s,x)\right)^2\right] + m_0\left(\sum_{i=1}^{n} \lambda_i - 1\right) =$$

$$= -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma_{I,x}(s_i - s_j) + 2\sum_{i=1}^{n} \lambda_i \gamma_{I,x}(s_i - s) + m_0\left(\sum_{i=1}^{n} \lambda_i - 1\right).$$

Derivando respecto a λ_i y m_0 e igualando a cero se obtiene:

$$-2\sum_{j=1}^{n} \lambda_j \gamma_{I,x}(s_i - s_j) + 2\gamma_{I,x}(s_i - s) + m_0 = 0$$

$$\sum_{i=1}^{n} \lambda_i - 1 = 0$$

Tomando $m = -m_0/2$ el sistema anterior escrito en forma matricial queda:

$$\begin{pmatrix} \Gamma_{I,x} & d \\ d' & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ m \end{pmatrix} = \begin{pmatrix} b \\ 1 \end{pmatrix}, \tag{2.2}$$

donde:

$$\Gamma_{I,x} = \begin{pmatrix} \gamma_{I,x}(0) & \gamma_{I,x}(s_1 - s_2) & \dots & \gamma_{I,x}(s_1 - s_n) \\ \gamma_{I,x}(s_2 - s_1) & \gamma_{I,x}(0) & \dots & \gamma_{I,x}(s_2 - s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{I,x}(s_n - s_1) & \gamma_{I,x}(s_n - s_2) & \dots & \gamma_{I,x}(0) \end{pmatrix}$$

$$d = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix}, \quad b = \begin{pmatrix} \gamma_{I,x}(s_1 - s) \\ \gamma_{I,x}(s_2 - s) \\ \vdots \\ \gamma_{I,x}(s_n - s) \end{pmatrix}$$

La solución del sistema (2.2) es:

$$\begin{pmatrix} \lambda \\ m \end{pmatrix} = \begin{pmatrix} \Gamma_{I,x} & d \\ d' & 0 \end{pmatrix}^{-1} \begin{pmatrix} b \\ 1 \end{pmatrix},$$

y como varianza de la predicción se obtiene:

$$\begin{aligned} \sigma^2_{KOI}(s,x) &= -\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma_{I,x}(s_i - s_j) + 2\sum_{i=1}^{n} \lambda_i \gamma_{I,x}(s_i - s) \\ &= \sum_{i=1}^{n} \lambda_i \gamma_{I,x}(s_i - s) + m \\ &= \begin{pmatrix} \lambda \\ m \end{pmatrix}' \begin{pmatrix} b \\ 1 \end{pmatrix}, \end{aligned}$$

ya que $\sum_{j=1}^{n} \lambda_j \gamma_{I,x}(s_i - s_j) = \gamma_{I,x}(s_i - s) - m$ en virtud del sistema anterior.

El método kriging ordinario aplicado al contexto de estimación de la función de distribución suele designarse como kriging indicador (ordinario). Un procedimiento alternativo para la estimación de la función de distribución $F_s(x)$ puede derivarse de la relación existente entre la meseta del variograma indicador, $S(\cdot)$, y la función de distribución, $F(\cdot)$, (Journel, 1983):

$$S(x) = \lim_{\|h\| \to \infty} \gamma_{I,x}(h) = F(x) - F(x)^2$$

De esta forma, los valores de la función de distribución $F(x)$ pueden obtenerse resolviendo la ecuación de segundo grado:

$$F^2(x) - F(x) + S(x) = 0. \tag{2.3}$$

La función $S(x)$ toma valores en el intervalo $[0, 0.25]$, alcanzando su máximo para la mediana de la distribución, que se denotará por x_M, es decir $F(x_M) = 0.5$. Además es una función creciente en $(-\infty, x_M]$ y decreciente en $[x_M, \infty)$.

Por consiguiente la resolución de la ecuación (2.3) toma la forma:

$$F(x) = 0.5\left(1 + \epsilon(x)\sqrt{1 - 4S(x)}\right), \tag{2.4}$$

donde $\epsilon(x) = sign(x - x_M)$.

A partir de la igualdad (2.4) y de la estimación de la meseta $S(x)$ se puede obtener una estimación de la función $F(x)$ para un valor de corte $x \in \mathbb{R}$ de la forma:

$$\hat{F}(x) = 0.5\left(1 + \hat{\epsilon}(x)\sqrt{1 - 4\hat{S}(x)}\right),$$

donde $\hat{\epsilon}(x) = sign(x - x_{\hat{M}})$ y $x_{\hat{M}}$ representa un estimador de x_M, que podría obtenerse como el valor para el cual la función $\hat{S}(x)$ alcanza el máximo, próximo a 0.25.

2.2. Estimación no paramétrica de la función de distribución

La aplicación del kriging ordinario para aproximar la función de distribución $F_s(x)$ de un proceso aleatorio espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$, requiere que Z sea estrictamente estacionario, lo que en la práctica suele ser muy restrictivo.

Una solución a este problema es descomponer Z en una parte determinista y una parte aleatoria de la forma:

$$Z(s) = \mu(s) + Y(s),$$

donde $\mu(\cdot)$ es la tendencia o media del proceso e $\{Y(s)/s \in D \subset \mathbb{R}^d\}$ es un proceso estrictamente estacionario de media 0 y función de distribución $G(x)$, para $x \in \mathbb{R}$.

Así, la estimación de la función de distribución de Z, $F_s(\cdot)$, puede transformarse en la estimación de la distribución de la variable estrictamente estacionaria Y, $G(\cdot)$, pues:

$$F_s(x) = \mathbb{P}[Z(s) \leq x] = \mathbb{P}[Y(s) \leq x - \mu(s)] = G_s(x - \mu(s)) = G(x - \mu(s)) = G(x_s),$$

donde $x_s = x - \mu(s)$.

Ahora, el problema de la estimación de la función de distribución $G(\cdot)$ puede abordarse mediante las técnicas ya mencionadas, siendo necesaria la estimación de la tendencia del proceso $\mu(\cdot)$, o mediante la construcción de un predictor no paramétrico indicador.

Un posible estimador de $G(\cdot)$ viene dado por el predictor kriging lineal como ya se ha presentado en la sección anterior:

$$\hat{G}(x) = \sum_{i=1}^{n} \lambda_i I_Y(s_i, x),$$

donde los coeficientes λ_i se obtienen resolviendo las ecuaciones kriging. Para ello se construiría el variograma empírico, por ejemplo el de Matheron o el de Nadayara-Watson, pero estos estimadores no son necesariamente válidos, por consiguiente sería necesario seleccionar una familia de variogramas paramétricos y ajustar sus valores. Finalmente se resolvería el sistema kriging.

De esta forma, la estimación de la función de distribución $F_s(x)$ constaría de los siguientes pasos:

1. Estimar la tendencia del proceso $\mu(\cdot)$.
2. Obtener los datos sin tendencia $Y(s_i) = Z(s_i) - \mu(s_i)$ y los valores indicadores $I_Y(s_i, x_s)$, para todo $s_i \in D, i \in \{1...n\}$ y $x_s = x - \mu(s)$ siendo x un valor de corte $x \in \mathbb{R}$.
3. Construir un estimador no paramétrico del semivariograma, $\hat{\gamma}_{I,x}(t)$.
4. Obtener un variograma válido, $\bar{\gamma}_{I,x}(t)$ a partir de $\hat{\gamma}_{I,x}(t)$.

5. Resolver un sistema kriging con pesos λ_i, $i \in \{1...n\}$, es decir $\hat{G}(x_s) = \sum_{i=1}^{n} \lambda_i I(s_i, x_s)$.

6. Finalmente, la estimación de la función de distribución $F_s(x)$ será $\hat{F}_s(x) = \hat{G}(x_s)$.

Otro posible estimador de $G(\cdot)$ viene dado por el método de la meseta propuesto por Journel (1983):

$$\hat{G}(x) = 0.5\left(1 + \hat{\epsilon}(x)\sqrt{1 - 4\hat{S}(x)}\right),$$

donde $\epsilon(x) = sign(x - \hat{x_M})$ y $\hat{S}(x)$ es el valor estimado de la meseta del variograma indicador.

De esta forma, la estimación de la función de distribución $F_s(x)$ constaría de los siguientes pasos:

1. Estimar la tendencia del proceso $\mu(\cdot)$.

2. Obtener los datos sin tendencia $Y(s_i) = Z(s_i) - \mu(s_i)$ y los valores indicadores $I_Y(s_i, x_s)$, para todo $s_i \in D, i \in \{1...n\}$ y $x_s = x - \mu(s_i)$ siendo x un valor de corte $x \in \mathbb{R}$.

3. Construir un estimador no paramétrico del semivariograma, $\hat{\gamma}_I(t, x_s)$.

4. Aproximar la meseta de $\hat{\gamma}_I(t, x_s)$, es decir $\hat{S}_Y(x_s) = \lim_{||t|| \to \infty} \hat{\gamma}_I(t, x_s)$.

5. Resolver la ecuación de segundo grado empleando la relación entre la meseta y la función de distribución:
$$\hat{G}(x_s) = 0.5\left(1 + \hat{\epsilon}(x_s)\sqrt{1 - 4\hat{S}(x_s)}\right).$$

6. Finalmente la estimación de la función de distribución $F_s(x)$ será $\hat{F}_s(x) = \hat{G}(x_s)$.

Un problema que se puede presentar en el kriging indicador y en el método de la meseta es que la función $\hat{G}(\cdot)$ resultante en cada caso no es necesariamente monótona creciente, es decir, que podría no verificarse que $\hat{G}(x) \leq \hat{G}(x')$ para algún par de valores x y x' tales que $x \leq x'$. Para solucionar las situaciones donde no se cumpla esta propiedad, consecuencia directa de la definición de la función de distribución, existen distintas opciones, como modificar las estimaciones de la distribución por el promedio de ambas, para cada par de valores consecutivos x y x' donde falle, o resolver las ecuaciones kriging utilizando el mismo variograma indicador (Cressie, 1993).

Otro inconveniente que presenta la adaptación, al contexto donde se admite la presencia de una tendencia determinista de los dos procedimientos indicados para la aproximación de la distribución, es que se requiere la caracterización de la propia tendencia $\mu(\cdot)$ puesto que

$$F_s(x) = G_s(x - \mu(s)) = G(x - \mu(s)) = G(x_s),$$

Ambos mecanismos demandan también la especificación del variograma indicador, aunque el método basado en el umbral, no precisa un estimador válido, de modo que bastaría para su aplicación la construcción de un variograma indicador no paramétrico.

Teniendo en cuenta lo anteriormente expuesto, otra opción sería construir directamente un predictor indicador no paramétrico de tipo núcleo para la función de distribución $F_s(x)$ de la forma:

$$\bar{F}_s(x) = \frac{\sum_{i=1}^{n} K\left(\frac{s - s_i}{h}\right) I_Z(s_i, x)}{\sum_{i=1}^{n} K\left(\frac{s - s_i}{h}\right)}, \tag{2.5}$$

siendo K una función de densidad de tipo núcleo y $h > 0$ el parámetro ventana.

Dado que $I_Z(h,x) \leq I_Z(h,x')$ si $x \leq x'$, $x, x' \in \mathbb{R}$, entonces $\bar{F}_s(\cdot)$ cumple la condición de ser no decreciente. Por otra parte, la aplicación del método tipo núcleo tiene una ventaja adicional, dado que no precisa la especificación del variograma indicador para cada valor x. No obstante, sí requiere una selección adecuada del parámetro ventana, para lo cual se propondrán distintos procedimientos.

Una versión más general consistiría en elegir como parámetro ventana una matriz $d \times d$, que se denotará por H, de forma que el estimador no paramétrico propuesto quedaría:

$$\bar{F}_s(x) = \frac{\sum_{i=1}^{n} K(H^{-1}(s-s_i)) I_Z(s_i,x)}{\sum_{i=1}^{n} K(H^{-1}(s-s_i))}. \tag{2.6}$$

Se podría simplificar la selección de la matriz ventana H reduciéndola a una matriz diagonal, $H = diag(h_1, ..., h_d)$, que permitiría incorporar diferentes grados de suavizado para las distintas coordenadas. En particular, si se toma $h_1 = ... = h_d = h$, se obtiene el selector univariante más simple, en función del cual se plantea el estimador no paramétrico (2.5).

Como alternativa, en la literatura estadística se propone tomar $H = hS^{-1/2}$, donde S denota la matriz de varianzas y covarianzas de las variables del proceso en las localizaciones muestreadas y h representa parámetro de suavización. La elección de la función de tipo núcleo gaussiana bidimensional con el parámetro ventana anterior, $H = hS^{-1/2}$ proporciona curvas de nivel que se orientan en la dirección de las posiciones. Si las posiciones tienen un coeficiente de correlación positivo, las curvas de nivel se orientan de forma positiva; de forma análoga si el coeficiente de correlación es negativo, las curvas tienen una orientación negativa. A su vez, modificaciones en la varianza producen achatamientos o engrosamientaos de las curvas de nivel.

Considerando el estimador (2.5), que requiere la especificación de un parámetro de suavizado unidimensional h, fijado un valor x, podría comenzarse por seleccionar una ventana óptima, local o global, de modo que minimizase el error cuadrático medio o el error cuadrático medio integrado, respectivamente. Esto supondría obtener el sesgo y la varianza de $\bar{F}_s(x)$. Para lo anterior, y con objeto de garantizar la consistencia del estimador, habría que imponer unas hipótesis sobre el proceso, similares a las introducidas por Hall y Patil (1994). En particular, se debería suponer que la región de observación es creciente, $D = \lambda_n D_0$, verificándose además los siguientes órdenes de convergencia:

$$\left\{\lambda^{-1} + n^{-2}\lambda^d h^{-d} + n^2\lambda^{-2d}h^d\right\} \to 0$$

Trabajando de forma similar a la planteada en García Soidán (2007) para el estudio del variograma no paramétrico de tipo núcleo, se deduciría que $\mathbb{E}[\bar{F}_s(x)] = \bar{F}_s(x) + O(h^2)$ y $Var[\bar{F}_s(x)] = O(n^{-2}\lambda^d h^{-d})$, de modo que los términos dominantes del sesgo y la varianza dependerían de valores desconocidos del proceso, que habría que estimar. Por lo tanto, se sugerirán otros mecanismos de selección del parámetro de suavizado h.

Entre los métodos alternativos de selección del parámetro de suavizado h podríamos utilizar los siguientes:

- Un mecanismo local de estimación basado en el método de los k puntos próximos, de modo que fijado $k \in \mathbb{N}$, se seleccionarían los datos asignados a las k posiciones muestrales más próximas a s. En particular, podría tomarse $k = 30$, por analogía con las recomendaciones de Journel y Huijbregts (2003) para la selección de las distancias en la estimación del variograma empírico.

- Un método global que utilice los datos localizados en un radio de búsqueda r, es decir, los valores observados del proceso correspondientes a las posiciones muestreadas que disten de la localización objetivo s menos o igual que un cierto valor r prefijado. Para la especificación del radio de búsqueda podría tomarse como r un porcentaje del rango (o una estimación del mismo) o de la máxima distancia entre las posiciones observadas.

- Un procedimiento global de validación cruzada, en el cual se seleccione la ventana h, dentro de un conjunto prefijado de valores, que minimice:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{F}_{s_i}^{(-i)}(x) - I_Z(s_i,x)\right)^2$$

  donde $\bar{F}_{s_i}^{(-i)}(x)$ denota la estimación no paramétrica de tipo núcleo de la distribución que se obtendría en la localización s_i al trabajar con todos los datos observados salvo el i-ésimo.

En el presente trabajo se pondrá a prueba el último método, validación cruzada, planteando una generalización del mismo a la estimación de la matriz diagonal bidimensional $H = diag(h_1, h_2)$. De esta forma se selecciona la matriz ventana H, dentro de un conjunto prefijado de valores, que minimice:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{F}_{s_i}^{(-i)}(x) - I_Z(s_i,x)\right)^2,$$

donde $\bar{F}_{s_i}^{(-i)}(x)$ denota la estimación no paramétrica de tipo núcleo (2.6).

Capítulo 3

Estudios numéricos

En este capítulo se describirán los estudios numéricos realizados para analizar el comportamiento de los distintos estimadores de la función de distribución en diferentes escenarios. Para ello se generarán datos de distintos procesos, en los que o bien se imponga la condición de estacionariedad o bien se relaje esta suposición admitiendo la presencia de una tendencia determinística, y donde la distribución subyacente del proceso podrá ser gaussiana o no gaussiana.

El principal objetivo es efectuar una comparativa entre el estimador indicador no paramétrico (NP) (2.6) y el estimador kriging indicador (KI) (2.1). A su vez, se estudiará la influencia del parámetro de suavizado en el estimador propuesto y el efecto de la variación de los parámetros del variograma sobre ambos estimadores.

Por último, se llevará a cabo una aplicación de los resultados a un caso real.

3.1. Datos simulados

En primer lugar, se describirán los resultados de los estudios numéricos realizados con datos simulados. Con este objetivo, se generarán valores de distintos procesos estocásticos sobre la región de observación $D = [0,1] \times [0,1] \subset \mathbb{R}^2$. Se trabajará bajo diseño fijo, empleando una rejilla regular para la selección de las localizaciones (datos equiespaciados), y también bajo diseño aleatorio, generando las posiciones muestreadas a través de la distribución uniforme en D, como se presenta en la Figura 3.1 y Figura 3.2, respectivamente.

En ambos casos, se han tomado distintos tamaños muestrales $n = 100$ y $n = 250$, para analizar el efecto que produce el incremento del número de datos en los diferentes estimadores.

De acuerdo con las recomendaciones de Goovaerts (1997), para reducir el coste computacional y la ocurrencia de fallos en la monotoneidad de la función de distribución estimada, el número de valores de corte x no debería de ser mayor que 15. Por otro lado, el número de valores de corte debería ser mayor que cinco para dar lugar a una discretización razonable de la función de distribución local. El conjunto de valores de x suele tomarse de forma que los valores del proceso estocástico Z se dividan en $x+1$ casos de igual frecuencia, es decir los nueve deciles de la distribución acumulativa de la muestra. Además, cabe comentar que los semivariogramas indicadores para valores extremos de x no están bien definidos, pues, en función del tamaño muestral, dependen de la distribución espacial de unos pocos pares de datos indicadores. Por lo tanto, valores menores que el primer decil o mayores que el noveno decil podrían ser inapropiados, aumentando el número de fallos en la monotoneidad de la función de distribución estimada.

Se tendrá en cuenta que en el kriging indicador la función $\hat{F}(\cdot)$ resultante en cada caso podría no sea necesariamente monótona creciente, es decir, que podría no verificarse que $\hat{F}_s(x) \leq \hat{F}_s(x')$ para algún par de valores x y x' tales que $x < x'$. Para solucionar esta cuestión, se ha recurrido a la opción

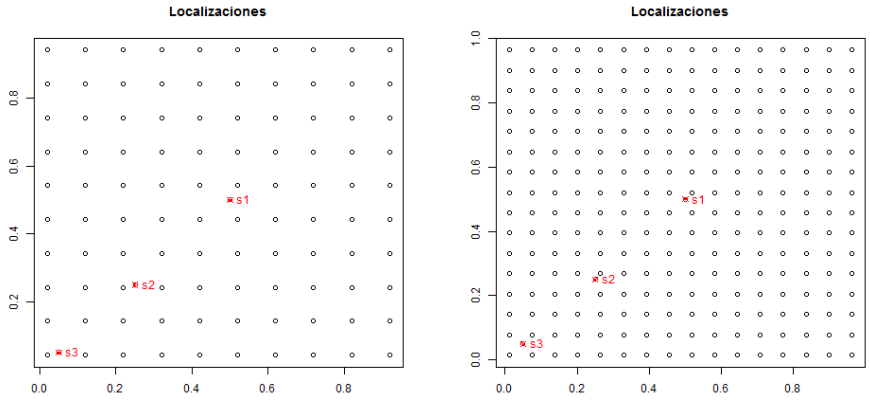

Figura 3.1: En negro las localizaciones de la muestra generada mediante diseño fijo con tamaños $n = 100$ (izquierda) y $n = 250$ (derecha). En rojo los tres puntos sobre los que se realizarán las estimaciones, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

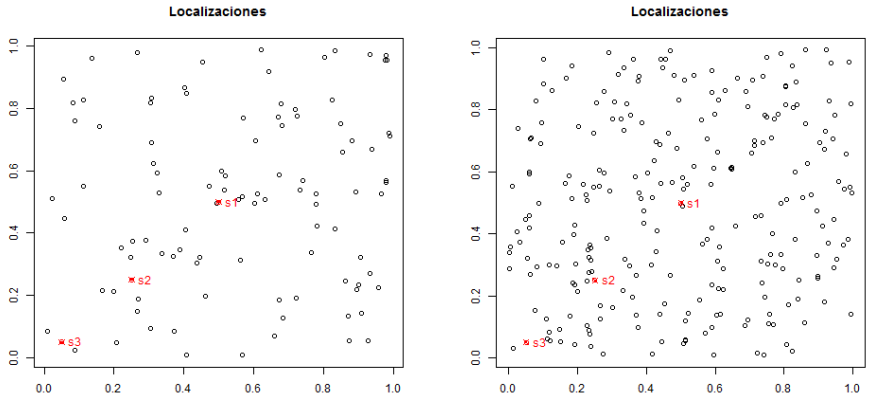

Figura 3.2: En negro ejemplo de las localizaciones de una muestra generada mediante diseño aleatorio con tamaños $n = 100$ (izquierda) y $n = 250$ (derecha). En rojo los tres puntos sobre los que se realizarán las estimaciones, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

de modificar las estimaciones de la distribución por el promedio de ambas, para cada par de valores consecutivos x y x' donde falle la propiedad de monotonía indicada. Para ello se ha empleado el algoritmo PAVA de Leeuw *et al.* (2009) implementado en el paquete *isotone* (2015) de R de los mismos autores. Además, se han truncado las predicciones obtenidas fuera del intervalo $[0, 1]$.

Para poder medir el ajuste de los estimadores de la función de distribución a los valores reales, se ha empleado el error cuadrático medio (en inglés *mean squared error*, MSE):

$$MSE_s(x) = \frac{1}{M} \sum_{j=1}^{M} (\tilde{F}_s^{(j)}(x) - F_s(x))^2,$$

donde M representa el número de muestras generadas y $\tilde{F}_s^{(j)}(x)$ representa la estimación obtenida, para cada uno de los métodos considerados, en la muestra j-ésima de $F_s(x)$, es decir, de la distribución

del proceso en la localización s para el valor x seleccionado.

En la construcción del estimador no paramétrico indicador de la función de distribución (2.6) se ha tomado $K(x,y) = K_e(x)K_e(y)$, siendo K_e la función de tipo núcleo de Epanechnikov,

$$K_e(u) = \frac{3}{4}(1-u^2)I\{|u| \leq 1\}.$$

Para la programación del estimador kriging indicador y la simulación de datos espaciales se ha recurrido a las librerías de R *sp* de Pebesma y Bivand (2016), *geoR* de Ribeirio y Diggle (2016), *gstat* de Pebesma y Graeler (2016) y *npsp* de Fernández Casal (2015).

3.1.1. Elección del parámetro/matriz ventana

Las primeras simulaciones realizadas han tenido por objetivo comparar el funcionamiento del estimador no paramétrico propuesto (NP) ante distintas elecciones del parámetro ventana h. En particular, se estudiará el comportamiento del parámetro de suavizado global obtenido por validación cruzada frente a un grid de valores para el radio de búsqueda.

Las simulaciones se han realizado para procesos espaciales no estacionarios, en concreto con tendencia lineal. Para el caso de estacionariedad, aunque las localizaciones más próximas tenderán a tomar valores parecidos, no variará sustancialmente el comportamiento de los datos en localizaciones más alejadas, al mantenerse el promedio de la variable objeto de estudio a lo largo de la región de observación. Por lo tanto, no se esperan grandes variaciones del MSE para diferentes valores de la ventana pero sí un decrecimiento conforme ésta aumenta, pues de esta forma el estimador no paramétrico propuesto se aproxima al valor de la función de distribución empírica,

$$\tilde{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(Z_i < x).$$

Teniendo en cuenta lo expuesto y que una de las ventajas que aporta el empleo del estimador no paramétrico propuesto frente al estimador kriging indicador de la función de distribución, es que no precisa la estimación de la tendencia del proceso, se ha considerado conveniente realizar el estudio para procesos espaciales no estacionarios.

A partir de las localizaciones espaciales seleccionadas s_i, $1 \leq i \leq n$, se supondrá que el proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ presenta tendencia lineal, es decir, es decir $\mu(s) = \mu(x,y) = ax + by + c$,

$$Z(s) = \mu(s) + Y(s).$$

En particular, se generarán n datos gaussianos $Z(s_i)$ en las localizaciones s_i consideradas, a partir de un proceso gaussiano de media dependiente de la localización, $\mu(x,y) = x + y + 2$, obteniendo la correspondiente estructura de dependencia a partir del modelo de variograma seleccionado. Concretamente, en estas simulaciones se ha recurrido al modelo isotrópico exponencial del semivariograma, de parámetros $c_0 = 0.4$, $c_2 = 0.4$ y $\sigma^2 = 1$:

$$\gamma(t) = \begin{cases} 0, \text{ si } t = 0; \\ 0.4 + 0.6\left(1 - \exp\left(-\frac{3t}{0.4}\right)\right), \text{ si } t > 0. \end{cases}$$

Con el objetivo de elegir un parámetro de suavizado adecuado, se han estimado los errores cuadráticos medios empleando la función indicadora en las tres localizaciones s^k para el valor de corte el cuantil P_{50} de una distribución normal de media 2.5 y desviación típica 1, en una muestra bajo diseño fijo de tamaño $n = 100$.

$$MSE_s^I(x) = \frac{1}{M}\sum_{j=1}^{M}(\bar{F}_s^{(j)}(x) - I_Z(s,x))^2,$$

donde M representa el número de muestras generadas y $\bar{F}_s^{(j)}(x)$ representa la estimación no paramétrica obtenida en la muestra j-ésima de $F_s(x)$, es decir, de la distribución del proceso en la localización s para el valor x seleccionado.

Se ha tomando una secuencia de 30 valores dentro del intervalo $[0.15, 1]$, construyendo para cada valor h la matriz suavizadora $H = diag(h, h)$, obteniendo así el selector univariante más simple que aporta el mismo grado de suavizado para las dos coordenadas, puesto que el efecto sobre la tendencia es similar para ambas y la construcción de $H = diag(h_1, h_2)$ no aporta diferencias notables en los valores del MSE^I. Los errores obtenidos se recogen en la Tabla 3.1 y de forma gráfica en la Figura 3.3. En ambos elementos se puede observar que el valor mínimo del error cuadrático medio empleando la función indicadora para las tres localizaciones, $s^{(j)}$ para $k = 1, 2, 3$, se ha alcanzado para los valores $h = 0.62$, $h = 0.27$ y $h = 0.33$ respectivamente. Además, si se comparan los valores obtenidos en las tres localizaciones objetivo $s^{(j)}$ $k = 1, 2, 3$, se puede observar que los errores aumentan conforme nos acercamos a la frontera, lo cual es esperable debido a la distribución de las localizaciones generadas.

	$h_1 = 0.15$	$h_2 = 0.18$	$h_3 = 0.21$	$h_4 = 0.24$	$h_5 = 0.27$	$h_6 = 0.30$	$h_7 = 0.33$	$h_8 = 0.36$	$h_9 = 0.38$	$h_{10} = 0.41$
$100 \times MSE^I_{s^{(1)}}$	29.5004	31.2148	29.2600	32.1820	27.1152	29.9216	30.9797	28.5728	27.2934	29.0506
$100 \times MSE^I_{s^{(2)}}$	39.9109	40.4826	36.7979	36.6637	**34.9077**	36.7989	36.8895	37.5742	35.3451	38.8989
$100 \times MSE^I_{s^{(3)}}$	39.9565	39.4717	37.1668	37.7096	36.8367	36.4751	**36.3417**	38.6526	38.5433	39.7094
	$h_{11} = 0.44$	$h_{12} = 0.47$	$h_{13} = 0.50$	$h_{14} = 0.53$	$h_{15} = 0.56$	$h_{16} = 0.59$	$h_{17} = 0.62$	$h_{18} = 0.65$	$h_{19} = 0.68$	$h_{20} = 0.71$
$100 \times MSE^I_{s^{(1)}}$	29.3172	28.7808	28.0271	27.8597	26.9386	26.9116	**24.6503**	28.8639	25.9563	28.1952
$100 \times MSE^I_{s^{(2)}}$	35.4846	36.3577	35.8088	37.8704	38.1098	37.5384	37.8520	39.7417	37.8003	39.6349
$100 \times MSE^I_{s^{(3)}}$	39.2962	39.3388	40.7576	40.5479	41.0926	42.4092	41.6460	43.1227	43.9477	43.7208
	$h_{21} = 0.74$	$h_{22} = 0.77$	$h_{23} = 0.79$	$h_{24} = 0.82$	$h_{25} = 0.85$	$h_{26} = 0.88$	$h_{27} = 0.91$	$h_{28} = 0.94$	$h_{29} = 0.97$	$h_{30} = 1$
$100 \times MSE^I_{s^{(1)}}$	27.3341	29.5333	26.9877	29.7009	26.4397	25.2662	25.9489	27.0199	26.1303	26.4088
$100 \times MSE^I_{s^{(2)}}$	38.7089	40.7678	40.1857	40.3359	41.4920	38.8929	40.9166	41.0200	41.3514	39.7770
$100 \times MSE^I_{s^{(3)}}$	44.4496	43.6874	46.6005	48.5761	48.8714	47.9156	47.9918	49.3413	49.8409	49.2990

Tabla 3.1: Error cuadrático medio empleando la función indicadora del estimador NP en las tres localizaciones $s^{(k)}$, bajo diseño fijo $n = 100$, valor de corte P_{50} de una distribución $N(2.5, 1)$, modelo exponencial, $M = 1000$, $\mu = x + y + 2$, $c_0 = 0.4$, $c_2 = 0.4$ y $\sigma^2 = 1$.

.

A continuación, se ha obtenido la predicción no paramétrica de la función de distribución empleando la matriz ventana, $H = diag(h_1, h_2)$, obtenida por medio de la validación cruzada del paquete *npsp* de Fernández Casal (2015), de modo que se selecciona la ventana H_{CV}, dentro de un conjunto prefijado de valores, que minimice:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\bar{F}_{s_i}^{(-i)}(x) - I_Z(s_i, x)\right)^2.$$

Para ello, fue necesario realizar previamente un binning de los datos indicadores. Empleando las mismas M simulaciones del proceso estocástico ya definido, el valor del error cuadrático medio obtenido mediante la estimación de la función de distribución empleando H_{CV} para cada una de las tres localizaciones se recogen en la Tabla 3.2.

Como se puede observar la diferencia entre el error obtenido empleando la matriz de suavizado

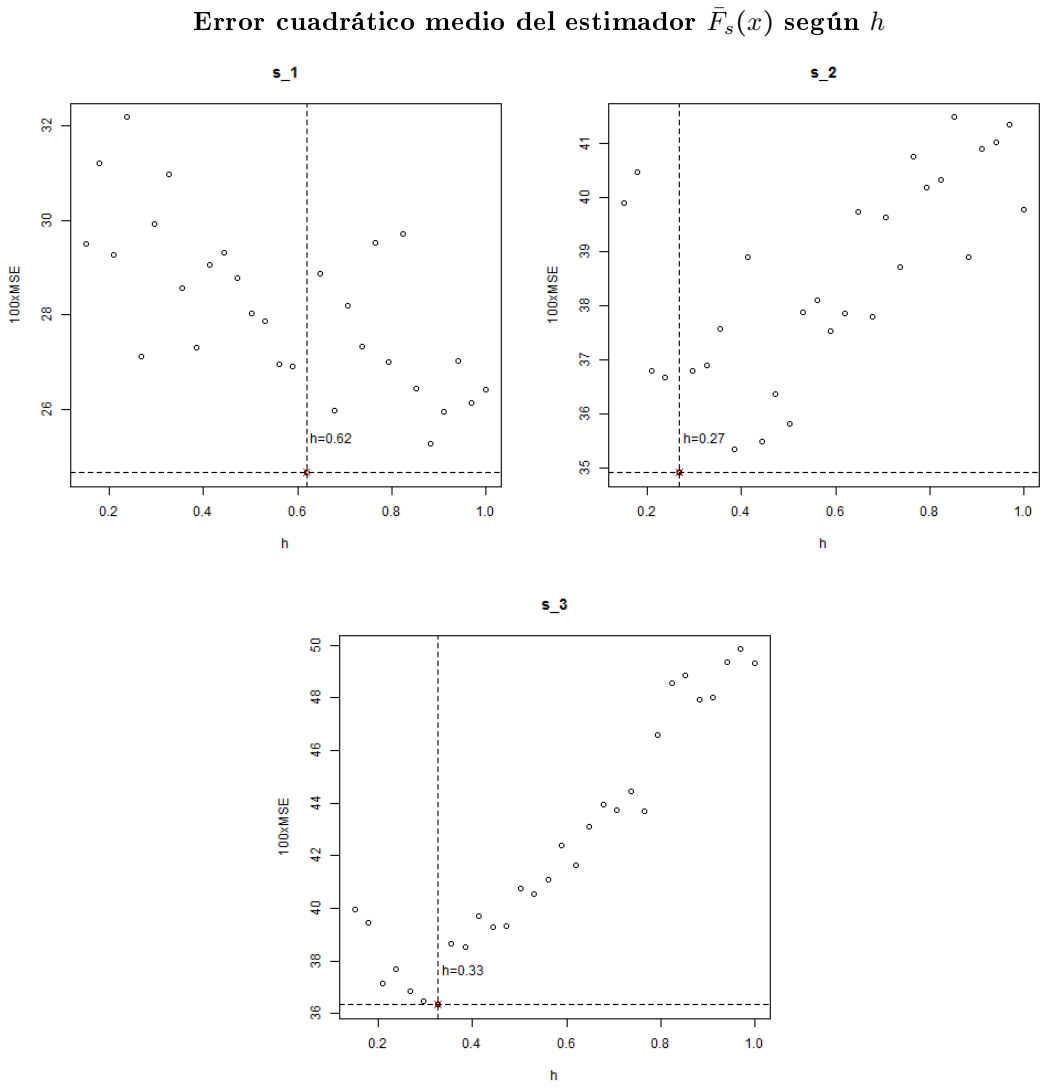

Figura 3.3: 100× Errores cuadráticos medios para el estimador indicador no paramétrico según el grid del parámetro ventana h, bajo diseño fijo, modelo exponencial y tamaño muestral $n = 100$.

H_{CV} y la matriz H que minimiza el MSE^I en las tres localizaciones es inferior a una décima. Es decir, ambos procedimientos conducen a valores similares para la ventana, si bien la ventana obtenida por validación cruzada es aplicable a casos reales, puesto que no requiere utilizar términos desconocidos y, además, es una ventana global que no depende de la localización objetivo. Por consiguiente, en las siguientes simulaciones se implementará la matriz ventana H_{CV} en la construcción del estimador no paramétrico de la función de distribución $\hat{F}_s(x)$, $x \in \mathbb{R}$ y $s \in D$.

$100 \times MSE_{s^{(1)}}$	$100 \times MSE_{s^{(2)}}$	$100 \times MSE_{s^{(3)}}$
28.5329	37.9099	43.5618

Tabla 3.2: Error cuadrático medio empleando la función indicadora del estimador NP en las localizaciones $s^{(k)}$ con matriz de suavizado obtenida mediante validación cruzada.

3.1.2. Proceso estacionario

Las siguientes simulaciones realizadas han tenido por objetivo comparar el funcionamiento del estimador no paramétrico propuesto (NP) de la función de distribución (2.6), frente al funcionamiento del estimador (2.1) obtenido mediante kriging indicador (KI). Para la selección de la matriz ventana $H = diag(h_1, h_2)$ para el estimador no paramétrico se ha empleado el método de validación cruzada sustentado en los resultados del punto anterior.

A partir de las localizaciones espaciales seleccionadas s_i, $1 \leq i \leq n$, se supondrá que el proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ es estrictamente estacionario, es decir, con tendencia constante $\mathbb{E}[Z(s)] = \mu$.

$$Z(s) = \mu + Y(s).$$

De este modo, se generarán n datos $Z(s_i)$ en las localizaciones de muestreo s_i, a partir de un proceso gaussiano de media $\mu = 5$ y la estructura de dependencia que se derive del modelo de variograma seleccionado. En particular, se han considerado dos modelos isotrópicos, exponencial y esférico, con distintos parámetros.

Como ya se indicó en el capítulo previo, para obtener las estimaciones del kriging indicador es necesario disponer de un estimador válido del semivariograma. En este estudio de simulación se ha partido del estimador no paramétrico empírico de Matheron y la obtención del semivariograma válido se ha llevado a cabo mediante el ajuste por mínimos cuadrados ponderados.

En primer lugar, se han generado datos de un proceso gaussiano, de media $\mu = 5$ y variograma isotrópico exponencial con parámetros $c_0 = 0.4$, $\sigma^2 = 1$ y $c_2 = 0.4$,

$$\gamma(t) = \begin{cases} 0, \text{ si } t = 0; \\ \\ 0.4 + 0.6\left(1 - \exp\left(-\frac{3t}{0.4}\right)\right), \text{ si } t > 0. \end{cases}$$

En total se han obtenido $M = 1000$ muestras, para las cuales se ha aproximado el valor de la distribución en las localizaciones objetivo $s^{(i)}$ para cinco valores de corte x, que se corresponden con los cuantiles 5 %, 25 %, 50 %, 75 % y 95 % de una distribución normal de media $\mu = 5$ y varianza $\sigma^2 = 1$; y con las dos metodologías (kriging y no paramétrica). Puesto que se conoce la distribución real del proceso estocástico se han obtenido los correspondientes valores reales de la función de distribución para los valores de corte dados y, por consiguiente, el error cuadrático medio para ambos estimadores.

La Tabla 3.3 recoge los resultados bajo diseño fijo para $n = 100$ y $n = 250$ localizaciones. Si se analiza la tabla se puede observar que para todos los valores de corte el estimador indicador no paramétrico (2.6) ha dado mejores resultados que el estimador kriging indicador (2.1). Además, el aumento del tamaño muestral ha dado lugar, por regla general, a errores cuadráticos más bajos para ambos estimadores como cabría esperar. Parece haber un crecimiento en el valor del MSE conforme las predicciones se localizan más próximas a la frontera de D, así como para los valores centrales de x. Teniendo en cuenta las recomendaciones ya mencionadas de Goovaerts (1997) los resultados correspondientes con los percentiles P_5 y P_{95} deben ser tomados con precaución.

En la Tabla 3.4 se recogen los resultados análogos bajo diseño aleatorio. Las conclusiones que se obtienen a partir de los datos son similares, el estimador propuesto ha obtenido mejores resultados

n=100	$100\times MSE_{s^{(1)}}$		$100\times MSE_{s^{(2)}}$		$100\times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	4.0603	0.3438	1.5047	0.4569	2.0140	0.6825
P_{25}	8.6305	2.7106	7.8270	3.1689	8.3357	3.9869
P_{50}	10.9491	4.1550	10.5365	4.7520	11.3582	5.7584
P_{75}	8.1186	2.7570	8.7816	3.1756	8.5426	3.8193
P_{95}	3.4305	0.3706	4.1177	0.4388	3.6785	0.5493

n=250	$100\times MSE_{s^{(1)}}$		$100\times MSE_{s^{(2)}}$		$100\times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	2.9576	0.3765	1.4285	0.4446	1.2385	0.5512
P_{25}	5.9015	2.4770	7.7760	2.8794	9.0402	3.5789
P_{50}	7.2243	3.6536	10.8551	4.2467	13.0646	5.2984
P_{75}	6.4594	2.4512	8.5296	2.8550	9.1399	3.5913
P_{95}	2.7481	0.3422	3.3192	0.3892	3.0208	0.5049

Tabla 3.3: Valores del error cuadrático medio para $n = 100$ (izquierda) y $n = 250$ (derecha), bajo diseño fijo, modelo exponencial, $M = 1000$, $\mu = 5$, $\sigma^2 = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

que el estimador obtenido mediante kriging indicador en todas las localizaciones $s^{(k)}$ y para todos los valores de corte x. De nuevo las estimaciones parecen mejorar al aumentar el tamaño muestral, obteniéndose los valores más elevados del MSE para los valores centrales de x.

n=100	$100\times MSE_{s^{(1)}}$		$100\times MSE_{s^{(2)}}$		$100\times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	5.4087	0.3888	1.0708	0.4801	0.8919	0.6044
P_{25}	8.2744	2.5809	5.5055	3.1393	6.1875	3.8047
P_{50}	9.2899	3.8423	7.1179	4.6981	9.4358	5.6458
P_{75}	8.9079	2.6194	6.9099	3.2483	9.0234	3.9681
P_{95}	5.6286	0.3686	4.9516	0.4501	5.7417	0.5575

n=250	$100\times MSE_{s^{(1)}}$		$100\times MSE_{s^{(2)}}$		$100\times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	5.5349	0.3864	1.6755	0.5077	1.2365	0.6691
P_{25}	7.6862	2.5866	7.0740	3.2143	7.4284	3.9661
P_{50}	8.6073	3.6908	9.5590	4.5523	9.6691	5.5354
P_{75}	7.9007	2.4234	7.7941	2.9590	7.7702	3.6308
P_{95}	4.9395	0.3852	4.9938	0.4959	4.4382	0.6364

Tabla 3.4: Valores del error cuadrático medio para $n = 100$ (izquierda) y $n = 250$ (derecha), bajo diseño aleatorio, modelo exponencial, $M = 1000$, $\mu = 5$, $\sigma^2 = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Comparando las simulaciones bajo diseño fijo o aleatorio se puede observar que el estimador propuesto se comporta de manera similar en ambos casos, mientras que para el estimador kriging indicador la nueva disposición de los datos ha hecho que aumente el error cuadrático medio en las localizaciones más extremas de x y disminuya en las centrales.

Para resumir los resultados de forma gráfica se ha recurrido a los boxplots de los errores cuadráticos para cada uno de los dos estimadores según la localización $s^{(k)}$, $k = 1, 2, 3$, y el valor de corte x, bajo diseño fijo, con tamaño muestral $n = 100$. Se han eliminado los valores atípicos para facilitar su interpretación (Figuras 3.4 y 3.5). Se puede observar que los diagramas de cajas obtenidos para el caso KI tienen mayor mediana y por consiguiente, mayor error, así como mayor dispersión, ya que el rango intercuartílico es mayor, comparando con los gráficos relativos al estimador NP.

A continuación, de manera análoga, se ha simulado un proceso estocástico espacial en las n locali-

zaciones a partir de un modelo isotrópico esférico, de parámetros $c_0 = 0.4$, $\sigma^2 = 1$ y $c_2 = 0.4$,

$$\gamma(t) = \begin{cases} 0, \text{ si } t = 0; \\ 0.4 + 0.6\left(1.5\frac{t}{0.4} - 0.5\left(\frac{t}{0.4}\right)^3\right), \text{ si } 0 < t < 0.4; \\ 1, \text{ si } t \geq 0.4. \end{cases}$$

En total se han efectuado $M = 1000$ simulaciones en las que se ha obtenido el valor estimado de la función de distribución en las localizaciones $s^{(k)}$, $k = 1, 2, 3$, para los valores de corte x anteriormente indicados, empleando las dos metodologías (kriging y no paramétrica). Puesto que se conoce la distribución real del proceso estocástico se han obtenido los correspondientes valores reales de la función de distribución para los valores de corte dados, así como el error cuadrático medio correspondiente.

La Tabla 3.5 recoge los resultados con variograma esférico, bajo diseño fijo, para muestras de tamaños $n = 100$ y $n = 250$ respectivamente. De nuevo se puede comprobar que el error cuadrático medio correspondiente a las predicciones $\bar{F}_s(x)$ es menor que el correspondiente al kriging indicador $\hat{F}_s(x)$, en todas las localizaciones $s^{(k)}$ y valores de corte x. Al aumentar el número de localizaciones, para ambos estimadores el MSE generalmente disminuye como cabría esperar. Si se comparan los resultados con el modelo exponencial de la Tabla 3.3, no se observan grandes diferencias en los valores del error.

n=100	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	2.3106	0.4595	2.1699	0.6108	1.6852	0.8353
P_{25}	8.3197	2.7837	8.6603	3.3369	8.8051	4.1846
P_{50}	11.0717	3.9427	10.8591	4.7402	12.4154	5.9044
P_{75}	8.6176	2.7220	7.8969	3.2382	8.8194	4.0715
P_{95}	1.6444	0.3996	1.8273	0.4883	1.2179	0.6510

n=250	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	2.4545	0.3969	1.8857	0.5022	1.0948	0.6220
P_{25}	6.9077	2.6079	7.3752	3.1282	7.2788	3.7875
P_{50}	9.4295	3.8927	10.3143	4.5848	10.4769	5.5430
P_{75}	7.4022	2.6386	7.8774	3.1524	7.8742	3.8389
P_{95}	2.1542	0.3742	2.3430	0.4460	2.3662	0.5793

Tabla 3.5: Valores del error cuadrático medio para $n = 100$ (izquierda) y $n = 250$ (derecha), bajo diseño fijo, modelo esférico, $M = 1000$, $\mu = 5$, $\sigma^2 = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

La Tabla 3.6 recoge los resultados con variograma esférico bajo diseño aleatorio para muestras de tamaños $n = 100$ y $n = 250$ respectivamente. Las conclusiones obtenidas son similares a las simulaciones anteriores. De nuevo los errores correspondientes al estimador $\hat{F}_s(x)$ han sido mayores que los derivados de $\bar{F}_s(x)$. El incremento del tamaño muestral ha supuesto una disminución del error del estimador NP y, por el contrario, un aumento para el estimador KI. La comparación con las tablas correspondientes al modelo exponencial no aporta diferencias destacables.

De nuevo, para resumir los resultados de forma gráfica se recurre a los boxplots de los errores cuadráticos para cada estimador en la localización objetivo $s^{(k)}$, $k = 1, 2, 3$, y el valor de corte x, bajo diseño aleatorio, con tamaño muestral $n = 100$. Se han eliminado los valores atípicos para facilitar su interpretación (Figuras 3.6 y 3.7). Se puede observar que los diagramas de cajas obtenidos para el caso KI en general tienen mayor mediana y por consiguiente, mayor error, así como mayor dispersión, ya que el rango intercuartílico es mayor. La única excepción se presenta en los gráficos correspondientes a los valores de corte P_5 y P_{95} cuyos resultados deben ser tomados con precaución siguiendo las indicaciones

n=100	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	2.3106	0.4595	2.1699	0.6108	1.6852	0.8353
P_{25}	8.3197	2.7837	8.6603	3.3369	8.8051	4.1846
P_{50}	11.0717	3.9427	10.8591	4.7402	12.4154	5.9044
P_{75}	8.6176	2.7220	7.8969	3.2382	8.8194	4.0715
P_{95}	1.6444	0.3996	1.8273	0.4883	1.2179	0.6510

n=250	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	2.4545	0.3969	1.8857	0.5022	1.0948	0.6220
P_{25}	6.9077	2.6079	7.3752	3.1282	7.2788	3.7875
P_{50}	9.4295	3.8927	10.3143	4.5848	10.4769	5.5430
P_{75}	7.4022	2.6386	7.8774	3.1524	7.8742	3.8389
P_{95}	2.1542	0.3742	2.3430	0.4460	2.3662	0.5793

Tabla 3.6: Valores del error cuadrático medio para $n = 100$ (izquierda) y $n = 250$ (derecha), bajo diseño aleatorio, modelo esférico, $M = 1000$, $\mu = 5$, $\sigma = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

de Goovaerts (1997).

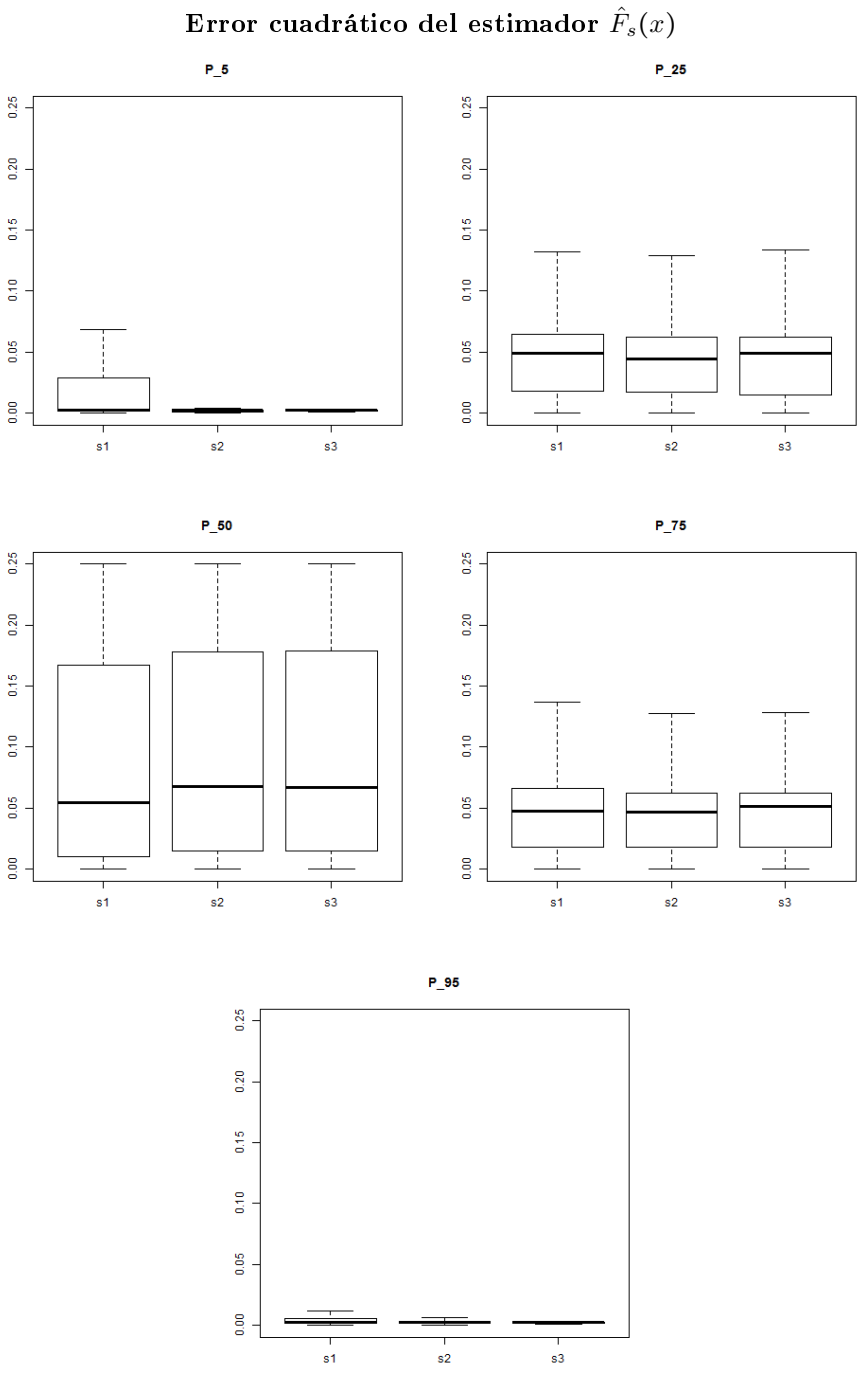

Figura 3.4: Boxplots de los errores cuadráticos para el estimador kriging indicador según el valor de corte y la localización, bajo diseño fijo, modelo exponencial y tamaño muestral $n = 100$. Los valores atípicos (*outliers*) han sido eliminados.

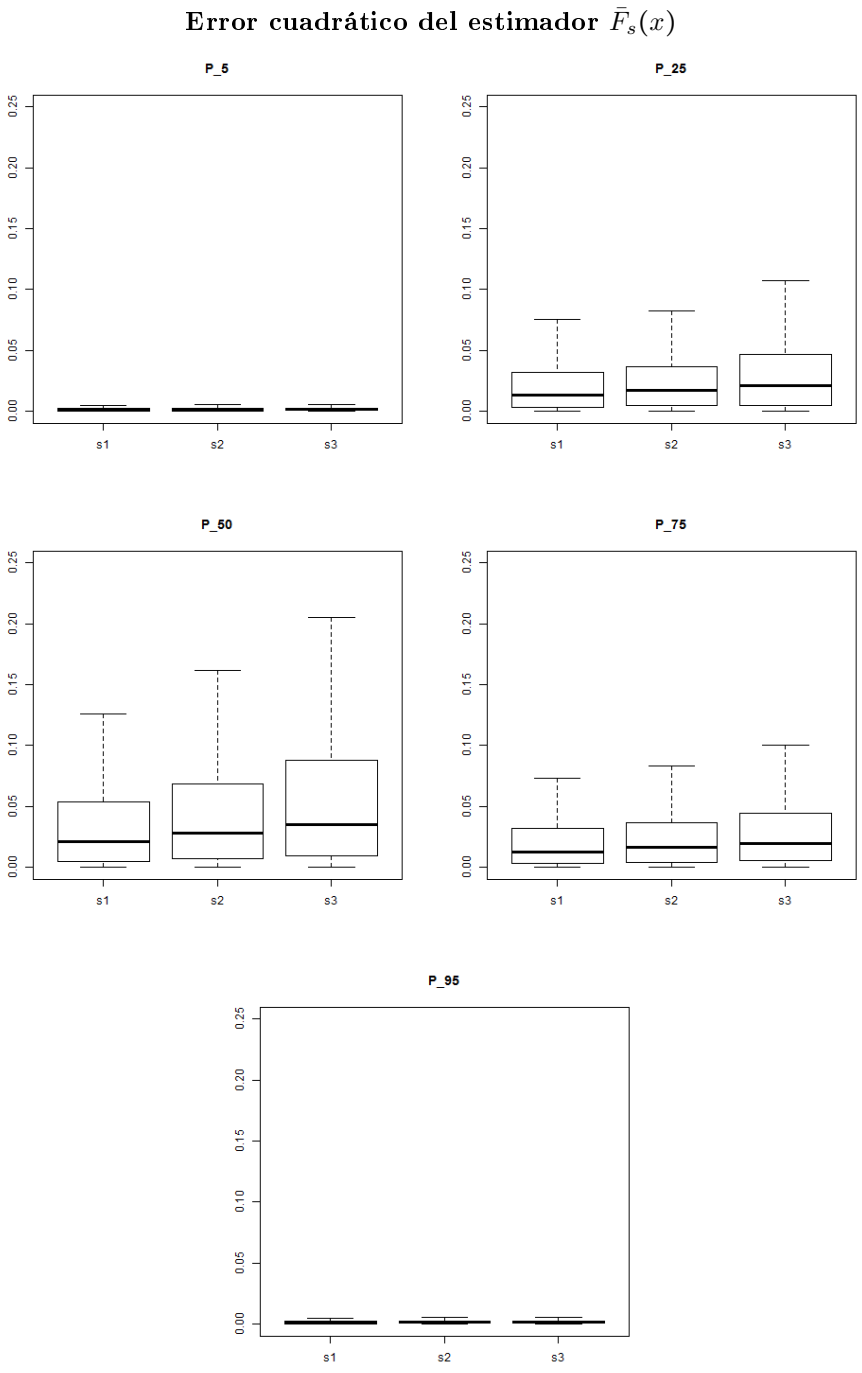

Figura 3.5: Boxplots de los errores cuadráticos para el estimador indicador no paramétrico según el valor de corte y la localización, bajo diseño fijo, modelo exponencial y tamaño muestral $n = 100$. Los valores atípicos (*outliers*) han sido eliminados.

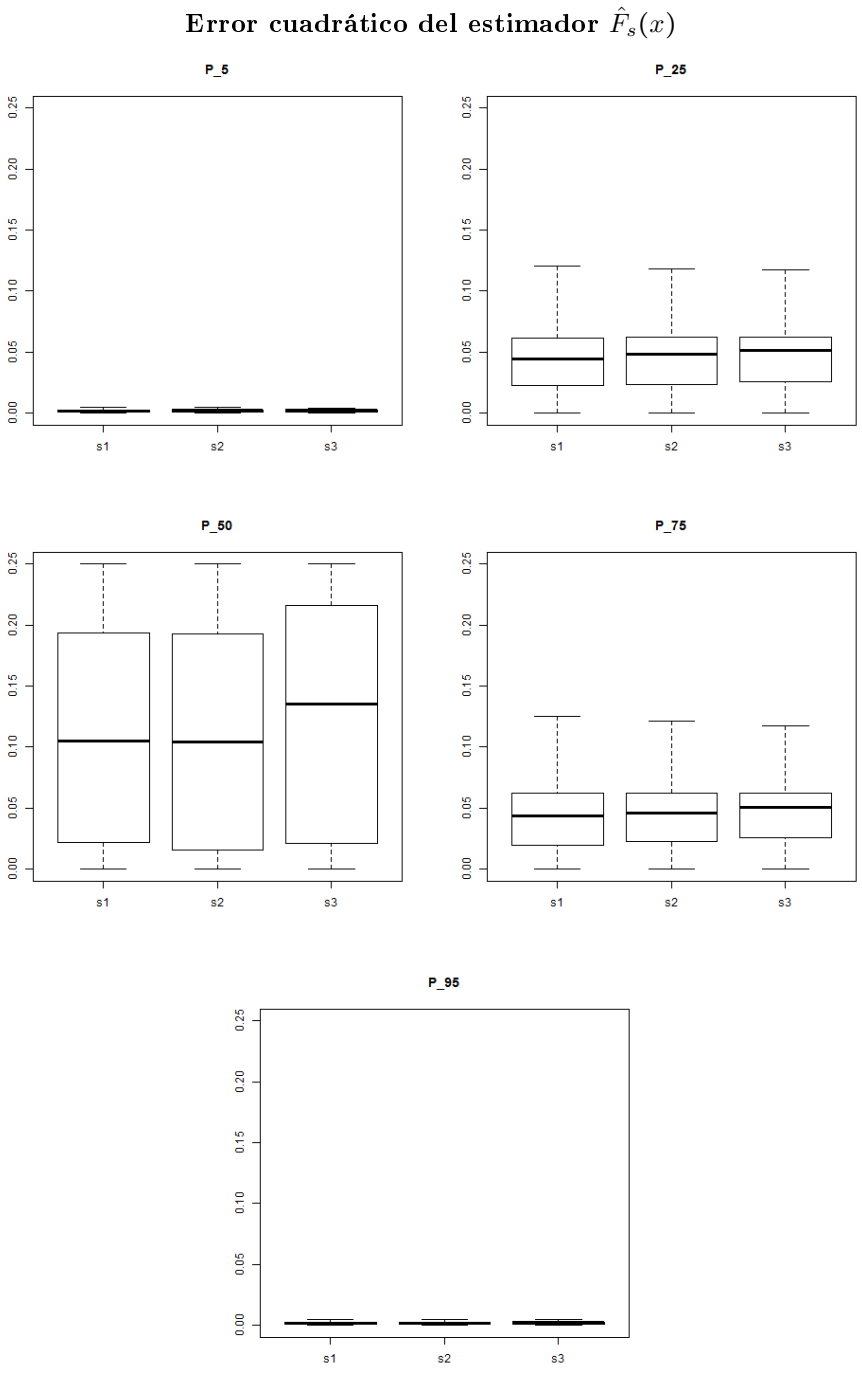

Figura 3.6: Boxplots de los errores cuadráticos para el estimador kriging indicador según el valor de corte y la localización, bajo diseño fijo, modelo esférico y tamaño muestral $n = 100$. Los valores atípicos (*outliers*) han sido eliminados.

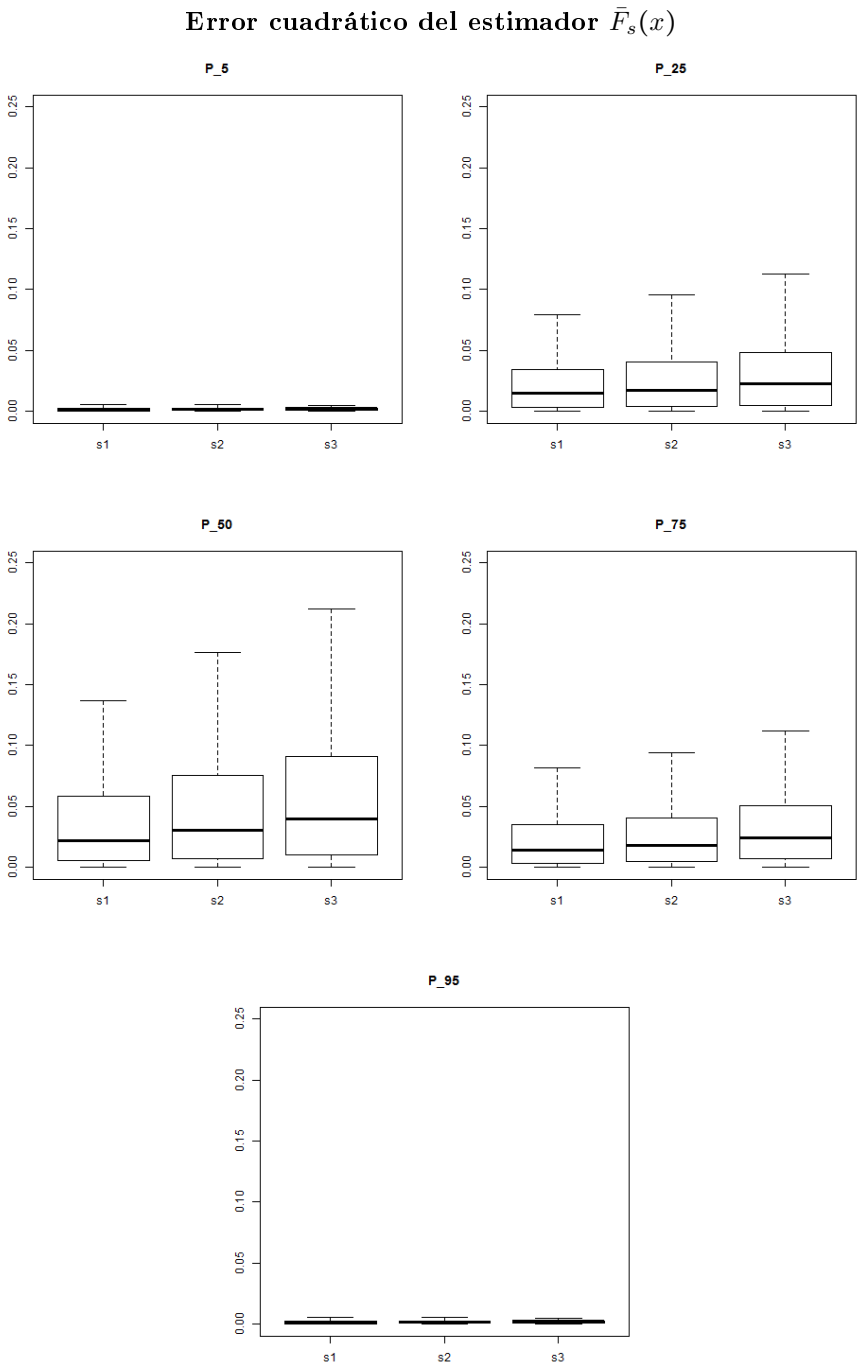

Figura 3.7: Boxplots de los errores cuadráticos para el estimador indicador no paramétrico según el valor de corte y la localización, bajo diseño fijo, modelo esférico y tamaño muestral $n = 100$. Los valores atípicos (*outliers*) han sido eliminados.

Las siguientes simulaciones han tenido como objetivo estudiar el comportamiento de los estimadores de la función de distribución al modificar la estructura de dependencia del proceso estocástico gaussiano definido. Para ello, se han modificado ciertos parámetros manteniendo constantes el resto. Cada simulación se ha obtenido bajo diseño fijo, con tamaño muestral $n = 100$, realizándose un total de $M = 1000$ simulaciones para cada escenario, y manteniendo los valores de x y las localizaciones de predicción $s^{(k)}$, $1 \leq k \leq 3$, de forma análoga a las simulaciones anteriores. Los resultados se recogen en las correspondientes tablas, indicadas a continuación.

- Modelo isotrópico exponencial:
  - Para valores del parámetro pepita $\{0, 0.15, 0.30, 0.45, 0.75\}$, con $\sigma^2 = 1$ y $c_2 = 0.4$, ver Tabla 3.7.
  - Para valores de la meseta, $\{0.2, 0.4, 0.6, 0.8, 1\}$, con $c_0 = 0.4$ y $\sigma^2 = 1$, ver Tabla 3.8.
  - Para valores de la varianza, $\{1, 1.5, 2, 2.5, 10\}$, con $c_0 = 0.4$ y $c_2 = 0.4$, ver Tabla 3.9.
- Modelo isotrópico esférico:
  - Para valores del parámetro pepita $\{0, 0.15, 0.30, 0.45, 0.75\}$, con $\sigma^2 = 1$ y $c_2 = 0.4$, ver Tabla 3.10.
  - Para valores de la meseta, $\{0.2, 0.4, 0.6, 0.8, 1\}$, con $c_0 = 0.4$ y $\sigma^2 = 1$, ver Tabla 3.11.
  - Para valores de la varianza, $\{1, 1.5, 2, 2.5, 10\}$, con $c_0 = 0.4$ y $c_2 = 0.4$, ver Tabla 3.12.

Las conclusiones sobre el comportamiento de los estimadores obtenidas a partir del estudio de los resultados se resumen a continuación.

El parámetro pepita se ha variado desde el valor 0 hasta un 75 % de la varianza, que se fijó a 1, de modo que el umbral parcial del semivariograma varía dentro del intervalo $[0, 0.25]$. Se puede observar en las Tablas 3.7 y 3.10, correspondientes al modelo exponencial y esférico respectivamente, que el aumento de dicho parámetro ha provocado que disminuya el error cuadrático medio de los estimadores para ambos modelos isotrópicos propuestos. Esta disminución se detecta para todos los puntos de corte y localizaciones objetivo. Ésto se debe a que conforme se aumenta el parámetro c_0, éste se aproxima a la varianza del proceso σ^2 (disminuye el umbral parcial) de ahí que el variograma tienda a ser más plano. Es decir, las variaciones entre los datos son menos dependientes de las distancias entre ellos. Esto se traduce en mayor similitud entre los datos y, por lo tanto, menores errores en las estimaciones, como se ha detectado en las tablas.

En cuanto a la comparación entre ambos estimadores, en todas las simulaciones correspondientes al parámetro pepita el estimador propuesto NP ha dado mejores resultados que el estimador KI.

El parámetro meseta, el valor a partir del cual se estabiliza el crecimiento del semivariograma, se ha variado entre 0.2 y 1. Para los distintos valores del error cuadrático medio en función de c_2 (Tabla 3.8 y Tabla 3.11) se llega a diferentes conclusiones para cada uno de los estimadores. En el caso del estimador KI no parece destacarse ningún patrón en el comportamiento del error ante las variaciones del parámetro, para todos los valores de corte y localizaciones objetivo. Sin embargo, para el estimador NP se detecta un aumento del MSE conforme aumenta c_2, lo que ocurre de manera general para las tres localizaciones objetivo y en todos los valores de corte. Ésto se debe a que al aumentar el rango los datos son más dependientes entre sí, de forma que cada vez se dispone de menos información en la construcción del estimador, esperando así un crecimiento en el error cuadrático medio.

Por otro lado, en cuanto a la comparación entre los dos estimadores análogamente al caso anterior, el MSE asociado al estimador NP ha sido menor para ambos modelos isotrópicos en todos los valores de corte y localizaciones objetivo.

Para concluir, se ha estudiado el efecto que produce un aumento en la varianza, variando sus valores entre 1 y 10. Dado que los procesos simulados son estacionarios de segundo orden, dicha varianza coincide con el umbral del semivariograma, el valor en el que se estabiliza su crecimiento. Se puede observar en las Tablas 3.9 y 3.12, correspondientes al modelo exponencial y esférico respectivamente,

que el aumento de σ^2 ha dado lugar a errores más altos para ambos estimadores en ambos modelos isotrópicos. Ésto se debe a que una mayor varianza implica más dispersión entre los datos, y por lo tanto es esperable obtener un mayor error cuadrático medio.

Al igual que en las anteriores simulaciones, de nuevo se puede observar que el estimador no paramétrico ha dado mejores resultados que el estimador kriging indicador para todos los valores de σ^2, puntos de corte y localizaciones objetivo.

Modelo exponencial del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_0 = 0$							$c_0 = 0.15$						
P_5	4.6139	0.7424	1.9871	0.9006	2.9550	1.1403		4.1663	0.6489	2.6735	0.8399	2.2253	1.0896
P_{25}	13.1338	4.5773	11.8076	5.5689	13.2255	6.9360		10.1975	3.8062	11.8144	4.7122	12.3762	5.9414
P_{50}	17.4044	6.7039	15.5923	8.3009	18.6451	10.2032		13.4785	5.5374	15.2567	6.7129	17.4413	8.4804
P_{75}	13.7076	4.6655	12.1067	5.9450	14.5820	7.3943		10.9828	3.7918	12.7324	4.5536	14.2825	5.9397
P_{95}	3.9794	0.7923	4.3335	1.0635	4.2564	1.4112		3.5233	0.6265	5.3238	0.8433	5.1901	1.1966
$c_0 = 0.30$							$c_0 = 0.45$						
P_5	3.8913	0.6081	2.5357	0.7526	2.7055	1.0477		3.7441	0.3526	1.5807	0.4264	1.0640	0.5331
P_{25}	9.7712	3.8698	10.6428	4.5106	12.1864	5.7313		8.3354	2.4940	7.6724	3.0166	7.9347	3.7891
P_{50}	13.0926	5.5734	13.9347	6.6089	17.1985	8.2221		9.9984	3.7622	10.3870	4.5438	11.5724	5.7684
P_{75}	10.7454	3.8235	11.9583	4.7025	13.4650	6.0321		7.8342	2.4929	8.4705	3.1188	9.3823	4.0020
P_{95}	4.0241	0.5877	4.6280	0.7216	4.4757	1.0002		3.4815	0.3233	4.2261	0.4749	3.8794	0.6901
$c_0 = 0.75$													
P_5	3.4650	0.1682	1.2777	0.2125	1.0055	0.2989							
P_{25}	4.6177	1.1719	6.5386	1.3993	7.1347	1.8667							
P_{50}	5.6937	1.7714	9.2476	2.1528	10.5863	2.9122							
P_{75}	5.1793	1.2123	6.9556	1.4818	8.5673	2.0156							
P_{95}	3.4781	0.1698	3.6177	0.2177	3.5171	0.3266							

Tabla 3.7: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = 5$, $\sigma^2 = 1$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo exponencial del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_2 = 0.2$							$c_2 = 0.4$						
P_5	3.5793	0.1816	1.3501	0.2401	1.4043	0.3548		4.0603	0.3438	1.5047	0.4569	2.0140	0.6825
P_{25}	6.1651	1.2091	8.3063	1.5372	9.9197	2.2785		8.6305	2.7106	7.8270	3.1689	8.3357	3.9869
P_{50}	8.0751	1.7994	11.7566	2.3408	13.8973	3.4489		10.9491	4.1550	10.5365	4.7520	11.3582	5.7584
P_{75}	6.9229	1.2411	9.0704	1.6219	10.2254	2.3601		8.1186	2.7570	8.7816	3.1756	8.5426	3.8193
P_{95}	3.0561	0.1761	4.3104	0.2289	3.9273	0.3802		3.4305	0.3706	4.1177	0.4388	3.6785	0.5493
$c_2 = 0.6$							$c_2 = 0.8$						
P_5	3.6387	0.5127	1.7250	0.5816	1.2115	0.6957		2.8942	0.6008	1.6418	0.7806	1.6923	0.9853
P_{25}	8.2265	3.4789	8.1509	3.9425	8.4634	4.6870		7.8459	3.9924	8.6999	4.5511	9.8267	5.2482
P_{50}	10.7422	5.0424	10.8914	5.7145	12.0605	6.8566		10.6696	5.9478	12.1111	6.7059	13.9205	7.7791
P_{75}	8.5522	3.5041	8.9700	4.0262	9.5897	4.7960		8.8667	4.1940	9.3437	4.7341	10.4120	5.4531
P_{95}	3.1853	0.5592	3.7685	0.6446	3.6964	0.8271		3.9512	0.6952	4.0319	0.8547	4.1562	1.0678
$c_2 = 1$													
P_5	2.7970	0.6841	2.1313	0.8657	1.8082	1.0637							
P_{25}	8.2333	4.3863	9.5131	4.8250	10.2481	5.5260							
P_{50}	10.7456	6.4879	12.6545	7.0070	13.4096	7.9558							
P_{75}	8.6896	4.5673	10.5118	4.9374	10.2434	5.5505							
P_{95}	3.9778	0.9896	4.6818	1.0650	3.9555	1.2120							

Tabla 3.8: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = 5$, $\sigma^2 = 1$, $c_0 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo exponencial del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
σ^2 =1							σ^2 =1.5						
P_5	4.0603	0.3438	1.5047	0.4569	2.0140	0.6825		5.8005	0.7401	1.9237	0.9941	2.1965	1.2867
P_{25}	8.6305	2.7106	7.8270	3.1689	8.3357	3.9869		13.2665	3.1907	8.8869	4.0842	9.7952	5.1085
P_{50}	10.9491	4.1550	10.5365	4.7520	11.3582	5.7584		15.4681	4.3043	10.8219	5.4123	13.0261	6.8358
P_{75}	8.1186	2.7570	8.7816	3.1756	8.5426	3.8193		12.9287	3.2808	8.8523	3.9239	9.8426	4.9191
P_{95}	3.4305	0.3706	4.1177	0.4388	3.6785	0.5493		5.6656	0.7494	3.7973	0.8932	4.2879	1.1171
σ^2 =2							σ^2 =2.5						
P_5	4.5069	1.0304	3.5801	1.2772	3.5665	1.7367		6.2370	1.4848	4.1324	1.7730	3.6289	2.1855
P_{25}	9.7235	3.6599	10.2592	4.2954	11.2707	5.3575		12.4435	4.1877	10.5270	4.8799	11.7611	5.9382
P_{50}	11.5150	4.4121	12.1210	5.1819	14.0380	6.4699		13.7500	4.8738	11.8474	5.6464	14.5380	6.9828
P_{75}	10.6534	3.4114	10.2258	4.0710	11.3768	5.1041		12.1697	3.8919	10.5230	4.4264	11.8616	5.4005
P_{95}	4.6528	0.9629	4.9643	1.2502	4.5396	1.5991		6.0648	1.2385	5.0355	1.3651	5.5441	1.6340
σ^2 =10													
P_5	9.4091	4.2480	10.0899	5.1244	11.1799	6.2497							
P_{25}	13.4316	5.8112	14.4894	7.0658	16.6668	8.8435							
P_{50}	14.5710	5.5095	14.5330	6.6991	16.5737	8.3394							
P_{75}	14.8621	6.0448	14.4748	7.0475	16.6641	8.5450							
P_{95}	10.6887	4.5523	10.8554	5.2353	11.4394	6.2768							

Tabla 3.9: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = 5$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo esférico del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_0 = 0$							$c_0 = 0.15$						
P_5	3.1629	0.7266	2.6458	1.0470	2.4223	1.3491		2.2735	0.5177	2.0952	0.6979	1.7933	0.9444
P_{25}	12.1320	4.5725	13.0595	5.6070	13.7670	6.9575		9.8403	3.5318	10.5889	4.3924	11.5497	5.5805
P_{50}	16.9970	6.5231	17.1927	7.8102	18.8753	9.5663		14.0789	5.1594	13.7499	6.1267	16.5387	7.7663
P_{75}	12.0396	4.4656	12.7274	5.3344	12.2808	6.6324		10.4544	3.5669	9.7779	4.1322	11.2545	5.1094
P_{95}	2.2537	0.7056	2.7750	0.9211	2.1637	1.2888		2.1108	0.5342	2.3501	0.6569	1.5707	0.8472
$c_0 = 0.30$							$c_0 = 0.45$						
P_5	2.1673	0.5494	2.3290	0.7182	1.3783	0.8987		1.3936	0.3216	1.6088	0.3581	0.9942	0.4630
P_{25}	9.2475	3.0691	8.7790	3.7017	9.2340	4.7143		6.3899	2.2012	9.1098	2.6251	9.0721	3.3825
P_{50}	12.3288	4.4718	12.0466	5.5208	14.3488	7.1030		9.6638	3.4465	13.2390	4.2682	14.2318	5.4456
P_{75}	9.0793	3.0566	8.9910	3.9639	9.8644	5.2048		6.8953	2.3428	10.2782	2.9460	10.1197	3.8212
P_{95}	1.9947	0.4661	1.9654	0.6644	1.3931	0.8865		1.5544	0.3424	2.5405	0.4832	1.4901	0.6881
$c_0 = 0.75$													
P_5	1.1832	0.1689	1.4479	0.2310	0.8645	0.3388							
P_{25}	5.0240	1.1121	6.3097	1.4606	6.1999	2.0100							
P_{50}	6.4891	1.7190	8.0291	2.1824	8.9890	2.9423							
P_{75}	4.5457	1.1107	5.2474	1.4137	6.0216	1.9421							
P_{95}	0.7309	0.1615	0.9649	0.2054	0.8290	0.3121							

Tabla 3.10: Valores del error cuadrático medio, bajo diseño fijo, modelo esférico, $M = 1000$, $n = 100$, $\mu = 5$, $\sigma^2 = 1$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo esférico del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
c_2 =0.2							c_2 =0.4						
P_5	1.8029	0.1856	1.2403	0.2850	0.9293	0.4333		2.3106	0.4595	2.1699	0.6108	1.6852	0.8353
P_{25}	8.4119	1.2560	6.8531	1.7748	7.1604	2.6541		8.3197	2.7837	8.6603	3.3369	8.8051	4.1846
P_{50}	11.8897	1.9364	10.1399	2.6915	10.7414	4.0435		11.0717	3.9427	10.8591	4.7402	12.4154	5.9044
P_{75}	8.2526	1.2202	7.7356	1.7352	7.6208	2.6588		8.6176	2.7220	7.8969	3.2382	8.8194	4.0715
P_{95}	1.5876	0.1688	1.5659	0.2542	0.6909	0.4269		1.6444	0.3996	1.8273	0.4883	1.2179	0.6510
c_2 =0.6							c_2 =0.8						
P_5	2.0319	0.6169	2.0928	0.7612	2.1643	0.9626		1.3584	0.6995	2.3899	0.8367	1.8299	1.0528
P_{25}	8.3711	3.5769	8.5242	4.2041	9.7613	5.0582		7.5720	4.0716	11.5866	4.6287	11.9324	5.4272
P_{50}	11.3625	5.4597	11.9027	6.3795	13.5407	7.4660		10.5615	6.0870	16.1458	6.7842	16.9152	7.7850
P_{75}	8.0106	3.7747	8.9324	4.4003	9.8063	5.3039		7.5280	4.1068	12.1518	4.6948	11.5813	5.3844
P_{95}	1.6784	0.5920	1.8000	0.7334	1.2323	0.9141		1.1083	0.5161	2.9098	0.6435	2.0597	0.8492
c_2 =1													
P_5	2.1256	0.7463	1.6235	0.8457	1.2012	0.9677							
P_{25}	9.6310	4.3746	7.8224	4.7842	8.3955	5.5028							
P_{50}	12.7666	6.4077	10.5486	6.8058	11.8816	7.7004							
P_{75}	9.1804	4.4458	7.7995	4.7451	7.7437	5.2578							
P_{95}	2.2801	0.8104	1.8396	0.8803	1.3212	0.9876							

Tabla 3.11: Valores del error cuadrático medio, bajo diseño fijo, modelo esférico, M = 1000, n = 100, μ = 5, σ^2 = 1, c_0 = 0.4, $s^{(1)}$ = $(0.5, 0.5)$, $s^{(2)}$ = $(0.25, 0.25)$ y $s^{(3)}$ = $(0.05, 0.05)$.

Modelo esférico del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
σ^2 =1							σ^2 =1.5						
P_5	2.3106	0.4595	2.1699	0.6108	1.6852	0.8353		3.0885	0.8276	2.9058	0.9322	1.7760	1.0821
P_{25}	8.3197	2.7837	8.6603	3.3369	8.8051	4.1846		9.4914	3.4020	9.4846	3.9064	10.0640	4.8519
P_{50}	11.0717	3.9427	10.8591	4.7402	12.4154	5.9044		12.1647	4.3852	12.0867	5.0871	13.3937	6.3915
P_{75}	8.6176	2.7220	7.8969	3.2382	8.8194	4.0715		8.5528	3.0813	9.3119	3.6838	10.7759	4.8174
P_{95}	1.6444	0.3996	1.8273	0.4883	1.2179	0.6510		2.0473	0.6386	2.6875	0.7747	1.8055	1.0185
σ^2 =2							σ^2 =2.5						
P_5	3.7704	0.9677	3.3548	1.1868	3.2978	1.5404		4.6205	1.5844	4.4349	1.7971	4.4615	2.1519
P_{25}	10.1685	3.5967	11.1369	4.2878	11.8181	5.3280		10.7390	4.2168	11.5568	4.8971	12.6442	6.0294
P_{50}	12.5053	4.6367	13.7597	5.4808	15.5333	6.7590		12.4883	4.9054	14.1334	5.7296	15.4349	7.0618
P_{75}	10.0951	3.7715	10.7506	4.3986	12.0592	5.4104		10.6011	3.9947	12.0853	4.7412	12.3817	5.8381
P_{95}	3.8371	1.1478	4.1463	1.3776	3.1710	1.7140		4.2759	1.4423	4.8318	1.7251	3.4496	2.1751
σ^2 =10													
P_5	10.2456	4.3283	9.7208	4.9029	8.7217	5.7003							
P_{25}	15.7262	6.1409	14.3384	6.9979	14.1404	8.4097							
P_{50}	15.7147	5.4416	14.5451	6.5473	14.9420	8.2208							
P_{75}	13.9966	5.5973	13.9804	6.7735	14.1740	8.4669							
P_{95}	8.8244	4.1759	9.6441	4.8580	9.3652	5.9137							

Tabla 3.12: Valores del error cuadrático medio, bajo diseño fijo, modelo esférico, M = 1000, n = 100, μ = 5, c_0 = 0.4, c_2 = 0.4, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

3.1.3. Proceso no estacionario

En esta sección las simulaciones realizadas han tenido por objetivo comparar el funcionamiento del estimador no paramétrico indicador de la función de distribución propuesto (2.6), frente al funcionamiento del estimador obtenido mediante kriging indicador (2.1), en presencia de tendencia espacial lineal.

A partir de las localizaciones espaciales seleccionadas s_i, $1 \leq i \leq n$, se supondrá que el proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ presenta tendencia lineal, es decir $\mu(s) = \mu(x,y) = ax + by + c$,

$$Z(s) = \mu(s) + Y(s).$$

Se generarán n datos gaussianos $Z(s_i)$ en las localizaciones s_i consideradas, $1 \leq i \leq n$, a partir de un proceso gaussiano de media dependiente de la localización, en particular $\mu(x,y) = x + y + 2$, obteniendo la correspondiente estructura de dependencia a partir del modelo de variograma seleccionado, en este caso el modelo isotrópico exponencial de parámetros $c_0 = 0.4$, $c_2 = 0.4$ y $\sigma^2 = 1$. Para el estimador no paramétrico propuesto se ha utilizado la matriz de suavizado obtenida mediante validación cruzada.

Como valores de corte x se han tomado los percentiles P_5, P_{25}, P_{50}, P_{75} y P_{95} de una distribución normal de media $\mu = 2.5$ y varianza $\sigma^2 = 1$. Puesto que se conoce la distribución real del proceso estocástico, se han obtenido los correspondientes valores reales de la función de distribución para los valores de corte dados en las localizaciones $s^{(k)}$. Dichos resultados se recogen en la Tabla 3.13.

$F_s(x)$	$s^{(1)}$	$s^{(2)}$	$s^{(3)}$
P_5	0.0160	0.0500	0.1066
P_{25}	0.1201	0.2500	0.3919
P_{50}	0.3085	0.5000	0.6554
P_{75}	0.5693	0.7500	0.8587
P_{95}	0.8739	0.9500	0.9796

Tabla 3.13: Valores reales de la función de distribución en las localizaciones $s^{(k)}$ para los distintos valores de corte x.

Como ya se indicó en el capítulo anterior, para obtener las estimaciones del kriging indicador es necesario efectuar los siguientes pasos:

1. Obtener los valores sin tendencia $Y(s_i) = Z(s_i) - \mu(s_i)$, $1 \leq i \leq n$, y los valores $x_s = x - \mu(s)$.

2. Construir un estimador no paramétrico del semivariograma, para lo cual se ha empleado el estimador empírico de Matheron.

3. Obtener un variograma válido, para lo cual se ha empleado el ajuste por mínimos cuadrados ponderados.

Por consiguiente, en virtud del primer paso, para obtener la predicción KI es necesario estimar la tendencia del proceso. Para ello se ha ajustado un modelo de regresión lineal múltiple con errores homocedásticos, usando como variables explicativas las localizaciones:

$$\mu(x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i + \epsilon_i,$$

donde (x_i, y_i) es el vector de localizaciones y los errores ϵ_i son independientes tales que $\mathbb{E}(\epsilon_i) = 0 \ \forall i$.

Las Tablas 3.14 y 3.15 recogen los resultados de ambos estimadores bajo diseño fijo y aleatorio respectivamente, para $n = 100$ y $n = 250$ localizaciones. Si se analizan las tablas se puede observar que para todos los valores de corte y localizaciones objetivo, el estimador indicador no paramétrico (2.6) de la función de distribución ha dado mejores resultados que el estimador kriging indicador (2.1). Además, para ambos estimadores al aumentar el tamaño muestral por regla general han disminuido los errores cuadráticos medios. Por otro lado, en el punto más próximo a la frontera $s^{(3)}$ se puede observar como para ambos escenarios y para los dos estimadores el error ha aumentado. Ésto se debe a que se dispone de menos datos para efectuar la predicción y al efecto frontera a la hora de estimar la tendencia.

n=100	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	1.7983	0.0976	1.1188	0.3074	2.6348	0.9291
P_{25}	4.1026	1.1356	9.3341	2.5820	12.8356	5.5959
P_{59}	7.9335	2.7642	12.3420	4.5416	12.8614	7.4151
P_{75}	10.1616	3.3321	8.4388	3.8020	6.9362	4.5676
P_{95}	5.3273	1.2550	2.3687	0.8012	1.8387	0.6245

n=250	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	2.5076	0.1032	1.5918	0.2893	3.1806	0.8432
P_{25}	4.6778	1.1928	8.6380	2.3927	11.1947	5.0540
P_{50}	8.6050	2.7166	10.9794	4.2602	10.3092	6.9083
P_{75}	10.2947	3.1758	8.0562	3.5730	4.6579	4.3096
P_{95}	4.6648	1.1852	2.1950	0.7698	1.1864	0.5804

Tabla 3.14: Valores del error cuadrático medio para $n = 100$ (izquierda) y $n = 250$ (derecha), bajo diseño fijo, modelo exponencial, $M = 1000$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

n=100	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	10.4259	0.1549	1.5002	0.3718	2.5340	0.9995
P_{25}	7.9413	1.4698	7.2731	2.8412	10.1280	6.0028
P_{50}	4.8079	3.2121	8.3909	5.1307	10.8237	8.6131
P_{75}	5.1956	3.6598	6.9666	4.4899	7.6910	5.6046
P_{95}	5.1505	1.3151	4.2074	1.0180	3.4630	0.8654

n=250	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	5.0648	0.1423	1.7884	0.3701	2.8847	0.9294
P_{25}	5.1722	1.2369	7.3774	2.4905	10.2111	5.2802
P_{50}	6.2961	2.7742	9.1883	4.4289	9.6890	7.3090
P_{75}	7.9582	3.2646	7.6229	3.9314	5.2278	4.8499
P_{95}	4.8853	1.2597	3.0266	0.8939	1.9067	0.7267

Tabla 3.15: Valores del error cuadrático medio para $n = 100$ (izquierda) y $n = 250$ (derecha), bajo diseño aleatorio, modelo exponencial, $M = 1000$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Para resumir los resultados de forma gráfica se recurre, de forma análoga al caso estacionario, a los boxplots de los errores cuadráticos para cada estimador según la localización $s^{(k)}$ y el valor de corte x, bajo diseño fijo con tamaño muestral $n = 100$. Se han eliminado los valores atípicos para facilitar su interpretación (Figuras 3.8 y 3.9). Como se puede observar la mediana de los errores cuadráticos medios correspondientes al estimador KI es más alta para todos los valores de corte y localizaciones que la correspondiente al NP. Además, el error del KI presenta mayor dispersión pues el rango intercuartílico es mayor en todos los gráficos, con excepción del diagrama de cajas correspondiente a la localización $s^{(3)}$ con valor de corte P_{75}.

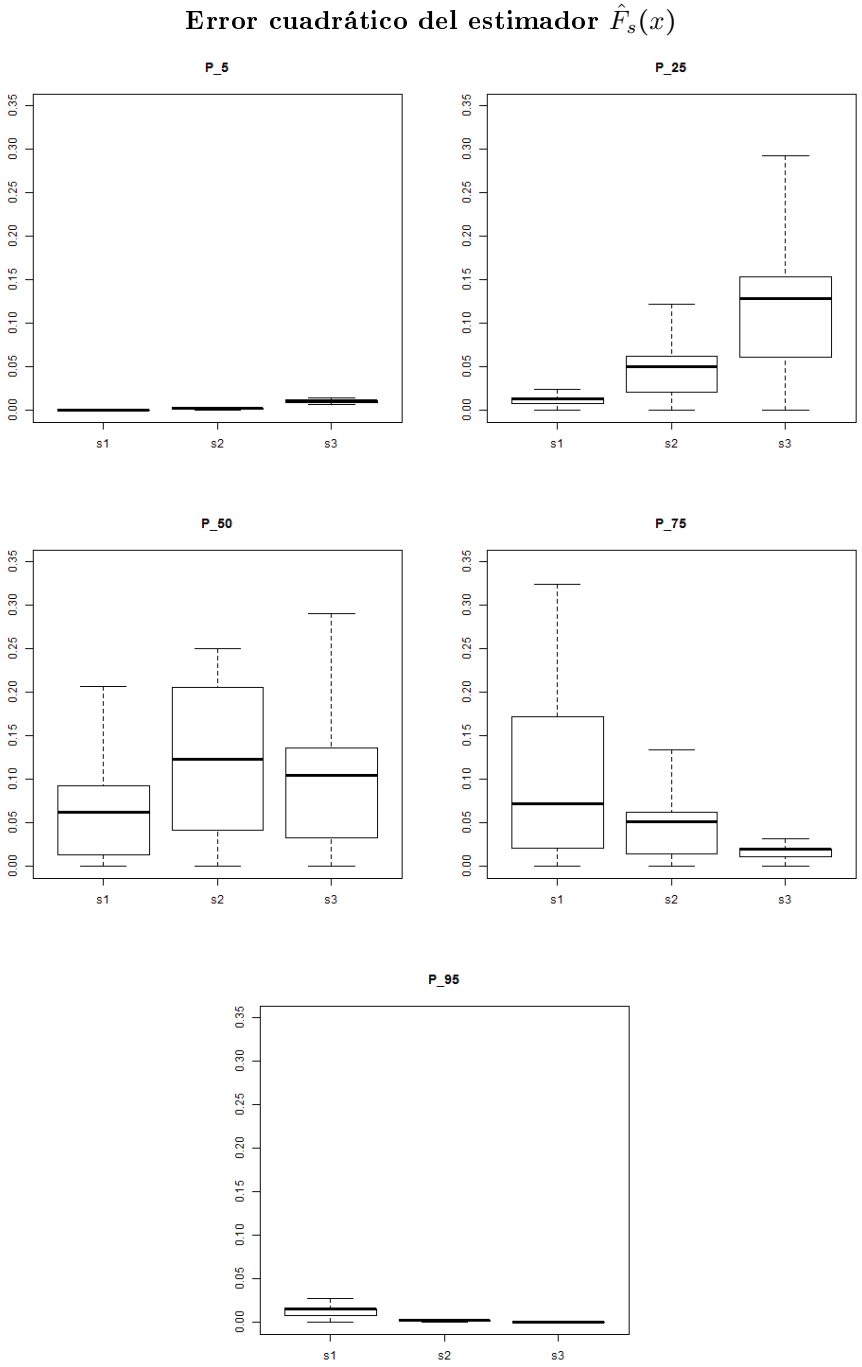

Figura 3.8: Boxplots de los errores cuadráticos para el estimador kriging indicador según el valor de corte y la localización, bajo diseño fijo y tamaño muestral $n = 100$. Los valores atípicos (*outliers*) han sido eliminados.

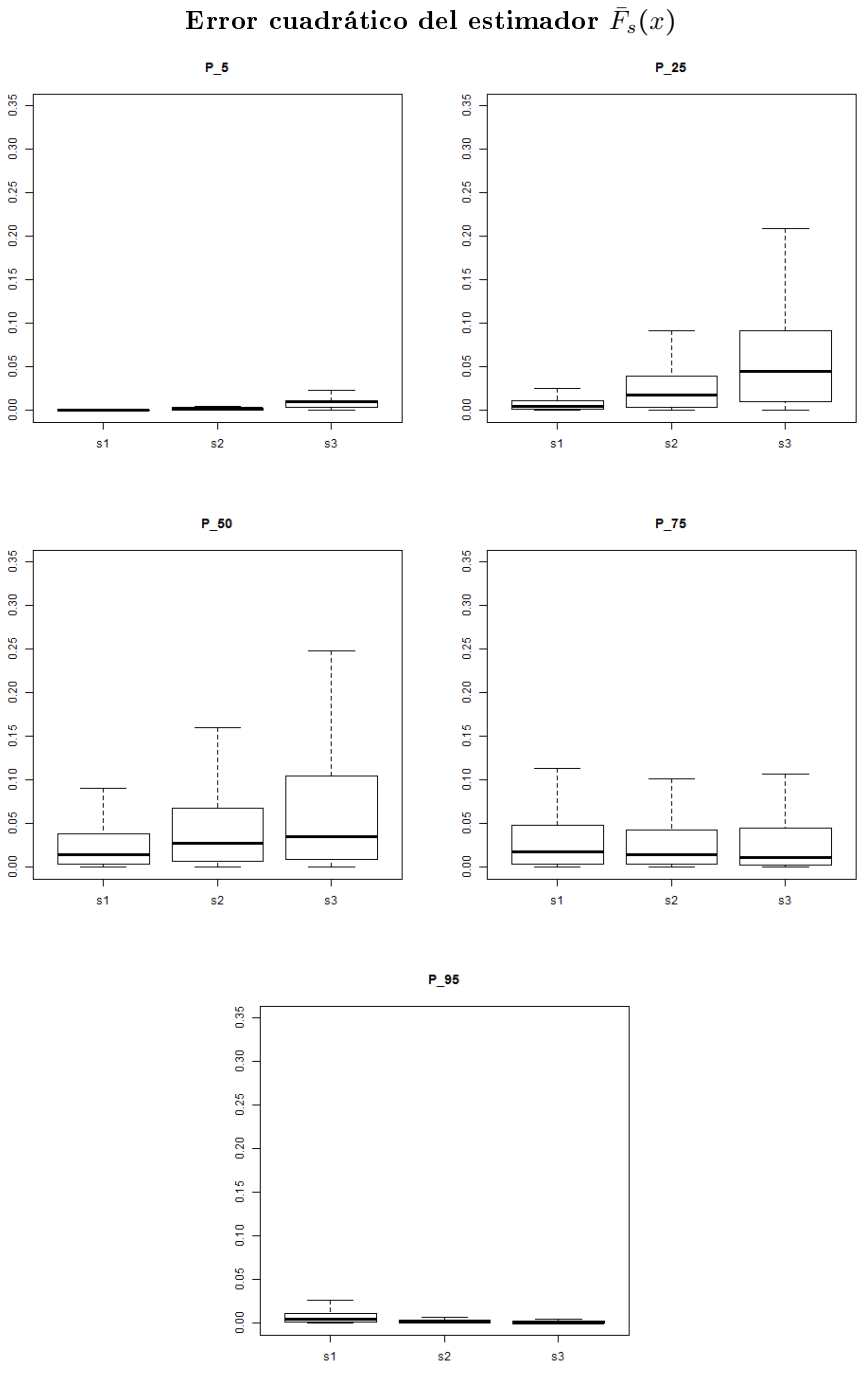

Figura 3.9: Boxplots de los errores cuadráticos para el estimador no paramétrico según el valor de corte y la localización, bajo diseño fijo y tamaño muestral $n = 100$. Los valores atípicos (*outliers*) han sido eliminados.

A continuación, se ha comparado el comportamiento del estimador no paramétrico y el estimador kriging indicador empleando los estimadores no paramétricos de Nadaraya-Watson y lineal local del semivariograma, en lugar del de Matheron. De esta manera, se han realizado $M = 1000$ simulaciones de un proceso estocástico espacial gaussiano en las n localizaciones a partir de un modelo isotrópico exponencial con tendencia lineal $\mu = x+y+2$, de parámetros, $c_0 = 0.4$, $\sigma^2 = 1$ y $c_2 = 0.4$. Además, se han tomado las mismas localizaciones objetivo $s^{(k)}$, $k = 1, 2, 3$, y valores de corte x que en las simulaciones anteriores.

A partir de los procesos simulados se han obtenido los estimadores NP y KI, para éste último se han empleado los variogramas empíricos mencionados, que posteriormente han sido ajustados a un modelo paramétrico mediante mínimos cuadrados ponderados. En este escenario para la construcción de ambos estimadores es necesario implementar un parámetro ventana, para lo cual de nuevo se ha recurrido a la matriz $H = diag(h_1, h_2)$ obtenida mediante el procedimiento de validación cruzada.

En la Tabla 3.16 se pueden consultar los resultados. En este caso el error cuadrático medio asociado al estimador KI, si se compara con los resultados de la Tabla 3.14 en los que se empleó el variograma empírico de Matheron en la estimación del semivariograma, ha disminuido en ambos casos (Nadaraya-Watson y lineal local).

Por otro lado, si se compara el estimador NP frente al KI obtenido empleando el estimador lineal local del semivariograma, el primero ha dado mejores resultados en todas las localizaciones y valores de corte. Sin embargo, no ocurre lo mismo en el caso del estimador del semivariograma de Nadaraya-Watson, para el cual el estimador NP ha dado mejores resultados que el KI con excepción del valor correspondiente a la localización $s^{(2)}$ y valor de corte P_{75}. No obstante, cabe destacar que la diferencia es del orden de 10^{-4} y en el resto de localizaciones el MSE ha sido mucho menor.

Si comparamos los errores correspondientes al estimador KI según las dos metodologías no paramétricas de estimación del semivariograma los resultados son similares. Por regla general el estimador de Nadaraya-Watson ha obtenido valores más bajos del MSE, si bien en la localización más próxima a la frontera, $s^{(3)}$, el estimador lineal local ha dado lugar a mejores resultados, lo que era de esperar pues dicho estimador reduce el sesgo en las proximidades de la frontera.

Por último, se ha analizado el tiempo de computación que lleva obtener una estimación de la función de distribución a partir de cada uno de los dos métodos (NP y KI) en una localización $s^{(k)}$, en particular $s^{(1)}$, para un valor de corte x, en particular $x = P_{50}$. Para la construcción del KI se ha recurrido al estimador del semivariograma lineal local. El tiempo de computación correspondiente al estimador NP ha sido de *user=0.11, system=0.01*, mientras que el correspondiente al estimador KI ha sido mayor, concretamente de *user=0.24, system=0.09*.

Nadaraya-Watson

n=100	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	8.0636	0.1263	0.4756	0.3905	1.3398	1.0601
P_{25}	5.7212	1.3393	4.2102	2.7551	7.8932	5.7411
P_{50}	3.5443	3.0929	5.3157	4.6542	9.8221	7.5030
P_{75}	3.8749	3.7295	3.9432	3.9611	5.9809	4.6796
P_{95}	1.7598	1.4375	1.0758	0.8924	1.2095	0.6972

Lineal Local

n=100	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP
P_5	9.3531	0.1187	0.7764	0.3894	2.1543	1.0881
P_{25}	6.9626	1.1780	5.6587	2.6807	8.4897	5.8201
P_{50}	4.7183	2.9000	6.3733	4.6930	8.6012	7.6109
P_{75}	5.3677	3.4402	4.2898	3.9399	4.7224	4.7314
P_{95}	2.1960	1.2874	0.8850	0.8092	0.8721	0.6587

Tabla 3.16: Valores del error cuadrático medio para el estimador de Nadaraya-Watson (izquierda) y el estimador lineal local (derecha), bajo diseño fijo, modelo exponencial, $M = 1000$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

De manera análoga a la sección anterior, las siguientes simulaciones han tenido como objetivo estudiar el comportamiento de los estimadores de la función de distribución al modificar la estructura de dependencia del proceso estocástico gaussiano definido, con tendencia lineal $\mu = x + y + 2$. Para ello, se han modificado ciertos parámetros manteniendo constantes el resto. Cada simulación se ha obtenido bajo diseño fijo, con tamaño muestral $n = 100$, realizándose un total de $M = 1000$ simulaciones para cada escenario, y manteniendo los valores de x y de las localizaciones objetivo $s^{(k)}$, $1 \leq k \leq 3$. Los resultados se recogen en las correspondientes tablas, indicadas a continuación. Las conclusiones que se extraen a partir de los mismos son similares al caso estacionario.

- Modelo isotrópico exponencial:
  - Para valores del parámetro pepita $\{0, 0.15, 0.30, 0.45, 0.75\}$, con $\sigma^2 = 1$ y $c_2 = 0.4$, ver Tabla 3.17.
  - Para valores de la meseta, $\{0.2, 0.4, 0.6, 0.8, 1\}$, con $c_0 = 0.4$ y $\sigma^2 = 1$, ver Tabla 3.18.
  - Para valores de la varianza, $\{1, 1.5, 2, 2.5, 10\}$, con $c_0 = 0.4$ y $c_2 = 0.4$, ver Tabla 3.19.
- Modelo isotrópico esférico:
  - Para valores del parámetro pepita $\{0, 0.15, 0.30, 0.45, 0.75\}$, con $\sigma^2 = 1$ y $c_2 = 0.4$, ver Tabla 3.20.
  - Para valores de la meseta, $\{0.2, 0.4, 0.6, 0.8, 1\}$, con $c_0 = 0.4$ y $\sigma^2 = 1$, ver Tabla 3.21.
  - Para valores de la varianza, $\{1, 1.5, 2, 2.5, 10\}$, con $c_0 = 0.4$ y $c_2 = 0.4$, ver Tabla 3.22.

Modelo exponencial del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_0 = 0$							$c_0 = 0.15$						
P_5	1.6680	0.2220	2.2380	0.6224	4.4241	1.5723		2.2815	0.2205	2.4652	0.6588	4.5888	1.5655
P_{25}	7.7227	2.4223	12.1662	4.8406	17.7085	8.9913		6.4959	1.8850	11.4112	4.0309	15.7869	7.6886
P_{50}	14.7878	5.0074	15.7051	7.8496	18.2377	11.9918		11.9908	4.2027	14.0822	6.7989	16.9637	10.3895
P_{75}	17.0208	5.7540	11.4644	6.6740	10.5037	7.9630		13.4498	4.8698	10.4086	5.7481	9.6118	7.0210
P_{95}	7.1277	2.1385	3.0781	1.3828	2.4422	1.2643		6.7912	1.8969	3.0222	1.2936	2.7296	1.0383
$c_0 = 0.30$							$c_0 = 0.45$						
P_5	1.9089	0.1304	1.5978	0.4374	3.2919	1.2108		1.8104	0.1028	1.5292	0.3652	3.0281	1.0715
P_{25}	5.3527	1.4100	9.3667	3.0820	13.5366	6.3577		4.8791	1.1159	7.4489	2.4944	11.4503	5.4830
P_{50}	9.1518	3.3368	11.9160	5.2232	12.7069	8.2946		9.4839	2.6056	9.5788	4.1169	11.2303	6.9045
P_{75}	11.5367	3.8670	9.0405	4.2190	6.9303	5.0069		11.3640	3.0640	7.1476	3.2618	5.8516	3.9844
P_{95}	5.3541	1.3917	2.8960	0.8991	1.6533	0.7148		5.0550	1.0913	2.1971	0.6420	1.3872	0.5023
$c_0 = 0.75$													
P_5	2.9087	0.0496	0.5677	0.1732	1.4910	0.6891							
P_{25}	4.1576	0.5614	5.3250	1.4367	8.4903	4.0470							
P_{50}	6.7690	1.3632	6.5688	2.5409	8.7163	5.2066							
P_{75}	7.8036	1.6480	4.6315	2.1672	5.3339	3.1270							
P_{95}	3.7798	0.6539	1.6167	0.4908	1.6332	0.4636							

Tabla 3.17: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo exponencial del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_2 = 0.2$							$c_2 = 0.4$						
P_5	2.6048	0.0700	0.7861	0.2357	2.1545	0.8055		1.7983	0.0976	1.1188	0.3074	2.6348	0.9291
P_{25}	5.9005	0.6689	6.9324	1.7495	11.6363	4.6492		4.1026	1.1356	9.3341	2.5820	12.8356	5.5959
P_{59}	10.7340	1.5575	8.9703	3.0559	12.6984	6.1337		7.9335	2.7642	12.3420	4.5416	12.8614	7.4151
P_{75}	11.8430	1.8229	6.7382	2.5884	7.0344	3.8618		10.1616	3.3321	8.4388	3.8020	6.9362	4.5676
P_{95}	5.1821	0.7007	1.8099	0.5990	1.3433	0.5803		5.3273	1.2550	2.3687	0.8012	1.8387	0.6245
$c_2 = 0.6$							$c_2 = 0.8$						
P_5	1.8897	0.1810	1.5814	0.5177	3.2576	1.2710		1.5738	0.1764	1.6123	0.4722	3.1551	1.1793
P_{25}	5.5214	1.7877	8.3998	3.6827	12.7115	7.0563		4.2332	1.8905	8.9716	3.7044	13.2455	6.9288
P_{50}	10.0119	3.9952	10.8151	6.1231	12.4809	9.1977		8.5004	4.4270	12.4345	6.5762	13.8822	9.5306
P_{75}	11.6500	4.6887	8.5882	5.0201	7.2612	5.7733		10.6937	5.3168	9.7195	5.6522	8.0126	6.3154
P_{95}	5.6283	1.7859	2.7825	1.0930	2.0430	0.8515		5.0577	1.9633	2.8643	1.2533	1.7969	1.0406
$c_2 = 1$													
P_5	2.1302	0.2016	1.6153	0.5758	3.5479	1.3480							
P_{25}	5.1296	2.2737	10.3674	4.2548	14.9215	7.5882							
P_{50}	9.2052	5.2220	13.8839	7.2949	15.7464	10.2332							
P_{75}	10.7795	6.1787	10.1386	6.3562	8.8449	6.9359							
P_{95}	5.1188	2.4643	3.7357	1.5668	2.0743	1.2206							

Tabla 3.18: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_0 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo exponencial del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$\sigma^2 = 1$							$\sigma^2 = 1.5$						
P_5	1.7983	0.0976	1.1188	0.3074	2.6348	0.9291		2.2815	0.2205	2.4652	0.6588	4.5888	1.5655
P_{25}	4.1026	1.1356	9.3341	2.5820	12.8356	5.5959		6.4959	1.8850	11.4112	4.0309	15.7869	7.6886
P_{59}	7.9335	2.7642	12.3420	4.5416	12.8614	7.4151		11.9908	4.2027	14.0822	6.7989	16.9637	10.3895
P_{75}	10.1616	3.3321	8.4388	3.8020	6.9362	4.5676		13.4498	4.8698	10.4086	5.7481	9.6118	7.0210
P_{95}	5.3273	1.2550	2.3687	0.8012	1.8387	0.6245		6.7912	1.8969	3.0222	1.2936	2.7296	1.0383
$\sigma^2 = 2$							$\sigma^2 = 2.5$						
P_5	1.4604	0.0747	1.0051	0.2931	1.7779	0.8115		1.3436	0.0700	0.8629	0.2568	1.2453	0.6654
P_{25}	5.3602	1.4736	9.2198	3.2926	12.8966	6.0956		5.6460	1.5118	8.3348	3.1160	11.4694	5.5361
P_{50}	10.6189	3.8006	12.4796	5.5620	14.9919	7.8113		12.4188	4.0174	12.9020	5.6072	13.1326	7.6503
P_{75}	10.0585	3.7356	8.2886	3.5878	7.1987	3.9540		11.9301	4.0649	8.7023	3.7326	6.3985	3.8062
P_{95}	3.1887	0.7584	2.0388	0.3951	1.8931	0.3214		3.8465	0.6551	2.3243	0.3424	1.5302	0.2590
$\sigma^2 = 10$													
P_5	0.3440	0.1242	0.2482	0.2443	0.4123	0.4077							
P_{25}	4.8293	2.5788	7.8851	3.9511	10.2051	5.7081							
P_{50}	13.8901	5.1277	14.2157	6.8838	17.7891	8.9200							
P_{75}	9.8684	4.1489	8.1308	3.8006	7.2217	3.7582							
P_{95}	1.1709	0.4272	0.9011	0.2359	0.6378	0.1500							

Tabla 3.19: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = 5$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo esférico del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_0 = 0$							$c_0 = 0.15$						
P_5	0.5806	0.3058	3.1743	0.8197	5.0245	1.8382		0.4829	0.1542	2.8499	0.5371	4.0679	1.3868
P_{25}	7.3955	2.6414	13.4284	5.0239	16.6629	8.8437		4.7620	1.8175	12.7807	3.9785	16.9224	7.7494
P_{50}	14.1324	5.3863	16.8764	7.9379	17.1985	11.3487		10.6807	4.1901	17.7954	6.8086	18.3755	10.6873
P_{75}	16.1762	5.9977	12.2223	6.3354	9.4345	7.2383		12.8953	4.8942	13.9200	5.7004	9.5392	7.0834
P_{95}	6.7457	2.3025	3.1010	1.5919	1.8539	1.4482		4.9085	1.9134	3.7070	1.2983	1.4133	1.1607
$c_0 = 0.30$							$c_0 = 0.45$						
P_5	0.7081	0.1269	1.2917	0.4039	2.2227	1.1581		0.4659	0.1197	1.5635	0.3449	2.3802	0.9295
P_{25}	5.9711	1.2998	7.6697	3.1281	12.9128	6.8073		3.7535	1.2037	7.4963	2.4479	11.4407	5.2547
P_{50}	13.1703	3.1902	11.4397	5.7579	15.8322	9.6583		8.7146	2.7473	10.5294	4.1163	11.5064	6.8034
P_{75}	15.2950	3.9525	8.8715	5.0365	9.1583	6.5361		10.4396	3.1584	7.7025	3.2942	5.1940	4.0455
P_{95}	6.1159	1.5066	2.0652	1.0729	1.1743	1.0273		4.0215	1.0769	1.5929	0.6563	0.6117	0.5607
$c_0 = 0.75$													
P_5	0.3092	0.0493	1.1588	0.1734	1.6555	0.6788							
P_{25}	2.7618	0.5732	5.2001	1.3942	9.3076	3.8424							
P_{50}	6.4965	1.2834	7.9472	2.4565	9.6049	5.0446							
P_{75}	7.9681	1.5341	5.8823	2.0707	5.1776	2.9492							
P_{95}	3.3351	0.5666	1.4379	0.4562	0.6825	0.4211							

Tabla 3.20: Valores del error cuadrático medio, bajo diseño fijo, modelo esférico, $M = 1000$, $n = 100$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo esférico del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
$c_2 = 0.2$							$c_2 = 0.4$						
P_5	0.5045	0.0668	1.4791	0.2243	2.0990	0.7906		0.5483	0.1211	1.7587	0.3400	2.5502	0.9575
P_{25}	4.0603	0.6434	7.5722	1.7478	12.4516	4.7100		4.9001	1.2835	9.0053	2.7040	13.2621	5.8201
P_{50}	8.4811	1.5035	11.0422	3.1542	13.9209	6.3458		8.9821	2.9783	12.3192	4.6424	14.0630	7.6400
P_{75}	9.9274	1.7597	8.1759	2.6046	7.1161	3.8608		10.3835	3.3696	8.7350	3.8100	6.7300	4.7406
P_{95}	4.0190	0.6898	1.9969	0.5664	0.8528	0.5538		3.6908	1.2321	1.9995	0.7991	0.5781	0.6623
$c_2 = 0.6$							$c_2 = 0.8$						
P_5	0.5083	0.1816	1.9466	0.5076	3.5151	1.2411		0.7316	0.1916	1.9399	0.5881	3.0819	1.3754
P_{25}	3.6594	1.6007	9.6015	3.4393	14.2735	6.7587		4.4384	1.9084	8.2999	3.7566	12.3798	6.9306
P_{50}	9.0029	3.6823	14.0228	5.8276	14.9355	8.9112		10.4606	4.5474	12.0223	6.2860	12.2550	9.0009
P_{75}	10.6476	4.3884	9.5146	4.9239	7.4811	5.6071		12.2138	5.4023	7.9388	5.3138	5.9016	5.7786
P_{95}	3.7177	1.6208	2.4312	1.0970	1.0841	0.9004		5.4209	2.0662	1.4961	1.1793	0.7686	0.8692
$c_2 = 1$													
P_5	0.5501	0.2329	1.8478	0.6031	3.2588	1.3810							
P_{25}	4.8070	2.4049	10.0670	4.3724	14.5030	7.7150							
P_{50}	9.8493	5.1259	13.6650	7.2757	16.7572	10.2662							
P_{75}	10.8403	5.8189	9.7503	5.9997	7.9815	6.5883							
P_{95}	4.9604	2.3573	1.8464	1.4270	1.0222	1.0893							

Tabla 3.21: Valores del error cuadrático medio, bajo diseño fijo, modelo exponencial, $M = 1000$, $n = 100$, $\mu = x + y + 2$, $\sigma^2 = 1$, $c_0 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Modelo esférico del variograma

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$			$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP		KI	NP	KI	NP	KI	NP
σ^2 =1							σ^2 =1.5						
P_5	0.6643	0.1243	1.3259	0.3657	2.4650	1.0557		0.2586	0.0819	1.5982	0.3260	2.1551	0.9347
P_{25}	4.2182	1.2101	7.5222	2.7603	12.4559	6.1946		4.1395	1.3442	9.9278	3.1199	13.7994	5.9886
P_{50}	9.4410	2.9579	11.6942	5.0423	14.7349	8.5913		9.9104	3.4345	14.1132	5.3842	15.2095	7.7974
P_{75}	10.6925	3.5294	8.5126	4.4056	7.6820	5.6642		10.9482	3.8331	9.7071	3.8965	7.0708	4.3119
P_{95}	4.3568	1.4494	1.7005	1.0474	0.9149	0.9206		3.3979	0.9891	1.6731	0.5560	0.5607	0.4226
σ^2 =2							σ^2 =2.5						
P_5	0.3920	0.0748	0.8334	0.2663	1.4792	0.7504		0.1292	0.0681	0.5752	0.2516	1.1332	0.6540
P_{25}	5.7944	1.5588	7.8359	3.4692	11.8357	6.2746		5.7896	1.5123	7.8694	3.4281	10.8984	6.0761
P_{50}	13.2348	3.9823	12.3893	6.0247	15.5769	8.4960		13.3772	3.9628	12.7491	5.7990	14.1320	8.0757
P_{75}	13.3737	4.0211	7.7147	4.1006	7.4756	4.6070		12.6888	3.8374	7.4792	3.5092	5.6067	3.5894
P_{95}	2.9749	0.7611	1.0781	0.4417	0.3183	0.3696		2.5685	0.6601	0.6534	0.3409	0.1538	0.2144
σ^2 =10													
P_5	0.1251	0.1236	0.2644	0.2420	0.4637	0.4070							
P_{25}	4.7721	2.6208	7.4122	4.1360	9.5272	6.0154							
P_{50}	14.3196	5.1282	15.0764	6.8486	17.0325	8.7889							
P_{75}	9.3373	4.0724	7.1381	3.6985	5.7741	3.4760							
P_{95}	0.6524	0.4509	0.2764	0.2508	0.1406	0.1534							

Tabla 3.22: Valores del error cuadrático medio, bajo diseño fijo, modelo esférico, M = 1000, n = 100, $\mu = x + y + 2$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

Finalmente, se ha comparado el funcionamiento del estimador no paramétrico indicador de la función de distribución propuesto (2.6), frente al funcionamiento del estimador obtenido mediante kriging indicador (2.1), en presencia de tendencia espacial no lineal y para un proceso no gaussiano.

Para las primeras simulaciones, a partir de las localizaciones espaciales seleccionadas s_i, $1 \leq i \leq n$, se supondrá que el proceso espacial $\{Z(s)/s \in D \subset \mathbb{R}^d\}$ presenta tendencia no lineal, es decir, $\mu(s) = \mu(x,y) = af(x) + bg(y) + c$, donde algunas de las funciones $f(\cdot)$ o $g(\cdot)$ es no lineal,

$$Z(s) = \mu(s) + Y(s).$$

Para el escenario con tendencia no lineal se generarán n datos gaussianos $Z(s_i)$ en las localizaciones s_i consideradas, a partir de un proceso gaussiano de media dependiente de la localización, en particular se tomará $\mu(x,y) = 3x^2 + 3y + 2$ y $\mu(x,y) = \exp(x)$, obteniendo la correspondiente estructura de dependencia a partir del modelo de variograma seleccionado, en este caso el modelo isotrópico exponencial de parámetros $c_0 = 0.4$, $c_2 = 0.4$ y $\sigma^2 = 2.5$.

De nuevo para obtener las estimaciones del kriging indicador es necesario estimar la tendencia del proceso y efectuar los pasos indicados en la sección anterior: obtener los valores sin tendencia, construir un estimador no paramétrico del semivariograma (de nuevo se recurrirá al de Matheron) y obtener un variograma paramétrico válido (mediante el ajuste por mínimos cuadrados ponderados).

Por consiguiente, para estimar la tendencia del proceso $\mu(x,y) = 3x^2 + 3y + 2$ se ha ajustado un modelo de regresión polinómica con errores homocedásticos, usando como variables explicativas las localizaciones:

$$\mu(x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_p x_i^p + \beta_1' y_i + \beta_2' y_i^2 + + ... + \beta_q' y_i^q + \epsilon_i,$$

donde (x_i, y_i) es el vector de localizaciones, p y q los grados del polinomio (mayor grado implica mayor flexibilidad en el ajuste) y los errores ϵ_i son independientes tales que $\mathbb{E}(\epsilon_i) = 0 \ \forall i$.

Por otro lado, para estimar la tendencia del proceso $\mu(x,y) = \exp(x)$ se ha ajustado un modelo aditivo con smoothing splines, usando como variables explicativas las localizaciones:

$$\mu(x_i, y_i) = X^*\Theta + f_1(x_i) + f_2(y_i),$$

donde $X^*\Theta$ corresponde con la parte estrictamente paramétrica del modelo y $f_1(\cdot)$ y $f_2(\cdot)$ es el efecto parcial de x_i y y_i en el predictor.

Para la comparación de los estimadores en caso de proceso no gaussiano, se han simulado n datos $Z(s_i) = Z'(s_i)^2$ en las localizaciones s_i consideradas, a partir de un proceso gaussiano $Z'(s_i)$ de media dependiente de la localización, $\mu(x,y) = x + y + 2$ con variograma isotrópico exponencial de parámetros $c_0 = 0.4$, $c_2 = 0.4$ y $\sigma^2 = 2.5$.

$$\gamma(t) = \begin{cases} 0, \text{ si } t = 0 \\ \\ 0.4 + 1.18\left(1 - \exp\left(-\frac{3t}{0.4}\right)\right), \text{ si } t > 0. \end{cases}$$

Para todas las simulaciones se ha tomado como parámetro ventana h la matriz de suavizado obtenida mediante validación cruzada, $H = diag(h_1, h_2)$. Además, puesto que se conoce la distribución real del proceso estocástico se han obtenido los correspondientes valores reales de la función de distribución para los valores de corte dados en las tres localizaciones $s^{(k)}$ y se ha obtenido el error cuadrático medio para las $M = 1000$ simulaciones, en los tres escenarios.

La Tabla 3.23 recoge los resultados correspondientes, donde se puede observar que el error cuadrático medio asociado al estimador NP ha sido menor para todos los valores de corte y localizaciones objetivo que el correspondiente al estimador KI. Además, para ambos estimadores el error ha aumentado notablemente en el caso de proceso no gaussiano.

	$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$		$100 \times MSE_{s^{(1)}}$		$100 \times MSE_{s^{(2)}}$		$100 \times MSE_{s^{(3)}}$	
	KI	NP	KI	NP	KI	NP	KI	NP	KI	NP	KI	NP
	$\mu(x,y) = 3x^2 + 3y + 2$						$\mu(x,y) = \exp(x)$					
P_5	0.4399	0.0040	0.3805	0.0797	1.5930	0.5621	5.9238	1.2295	3.7624	2.3249	4.8364	3.7644
P_{25}	1.9232	0.1911	7.8151	1.8207	12.4914	6.3310	13.0788	4.4031	12.2354	5.5362	14.1793	6.9529
P_{50}	6.1838	1.1018	14.2575	5.2353	17.4364	10.7360	10.2389	3.1419	8.5253	2.8795	8.1944	3.1667
P_{75}	12.5692	2.7691	13.3836	6.1376	9.6843	7.1622	4.3831	0.8566	3.0023	0.5362	2.9492	0.4846
P_{95}	11.5987	2.9039	3.5479	1.7868	1.1806	0.9025	0.8691	0.0264	0.2838	0.0092	0.1938	0.0058
	$Z(\cdot)$ no gaussiano											
P_5	0.0622	0.0622	0.2500	0.2500	0.6721	0.6721						
P_{25}	3.2863	1.4878	7.8687	3.3907	11.7046	6.2809						
P_{50}	10.4237	7.2056	16.9235	11.9028	21.7442	17.1376						
P_{75}	26.3531	21.7994	30.8513	26.6443	33.4326	30.5454						
P_{95}	48.0696	41.8838	40.0964	37.7500	35.1168	35.4592						

Tabla 3.23: Valores del error cuadrático medio para distintos escenarios con tendencia no lineal, bajo diseño fijo, modelo exponencial, $M = 1000$, $\sigma^2 = 2.5$, $c_0 = 0.4$, $c_2 = 0.4$, $s^{(1)} = (0.5, 0.5)$, $s^{(2)} = (0.25, 0.25)$ y $s^{(3)} = (0.05, 0.05)$.

3.2. Aplicación a datos reales

En esta sección se ha realizado un estudio numérico con datos reales para ilustrar el comportamiento en la práctica del estimador NP propuesto y compararlo con el método kriging indicador, tradicionalmente utilizado en este contexto. En la realización de este estudio, se tendrán en cuenta las conclusiones empíricas obtenidas en las simulaciones anteriores. Para ello, se ha recurrido a la base de datos de carbón del libro de Cressie (1993) obtenida originalmente por Gómez y Hazen (1970) en el Condado de Green (Pennsylvania), para estimar en qué localizaciones existe mayor probabilidad de encontrar más cantidad de mineral. En total se dispone de 208 localizaciones con distancias de separación entre ellas de 2500 pies (1 pie $\approx$ 0.762 km). Como se indica en Cressie (1993), el proceso estocástico correspondiente presenta tendencia lineal, es decir $\mu(s) = \mu(x, y) = ax + by + c$; a, b, $c \in \mathbb{R}$,

$$Z(s) = \mu(s) + Y(s).$$

En la Figura 3.10 se pueden observar las localizaciones muestreadas. A su vez, en la Figura 3.11 se recoge el nivel de carbón por colores a partir de la discretización de sus valores en sus cuartiles muestrales. Se ha incluido también el histograma y el diagrama de cajas de los porcentajes de carbón en las distintas localizaciones (Figura 3.12), en donde se puede observar que la función de densidad estimada parece acercarse a la campana correspondiente a una distribución normal. Para contrastar esta hipótesis se ha efectuado un test de normalidad a los datos, habiendo previamente eliminado tres valores atípicos correspondientes a las localizaciones $(5, 6)$, $(6, 8)$ y $(3, 17)$ (Figura 3.13). La realización del contraste de Shapiro-Wilks, con un p-valor de 0.5813 y un estadístico $W = 0.994$, no implica rechazar la hipótesis de normalidad para los niveles de carbón. Por lo tanto se supondrá que los datos son gaussianos.

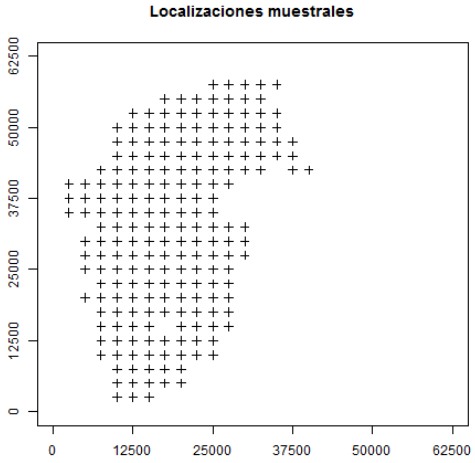

Figura 3.10: Localizaciones muestreadas de la base de datos de carbón. Unidades de localización de los ejes en pies (ft).

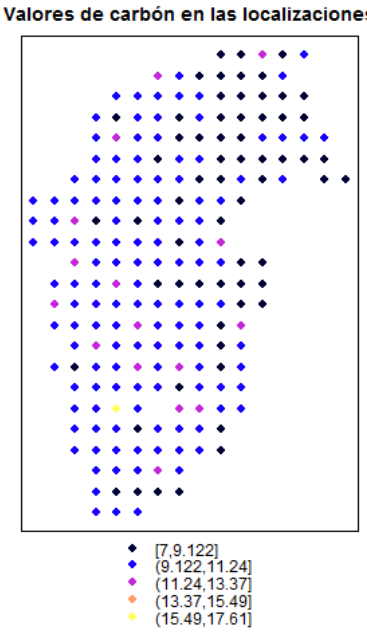

Figura 3.11: Discretización en cuartiles muestrales del porcentaje de carbón en las localizaciones muestreadas.

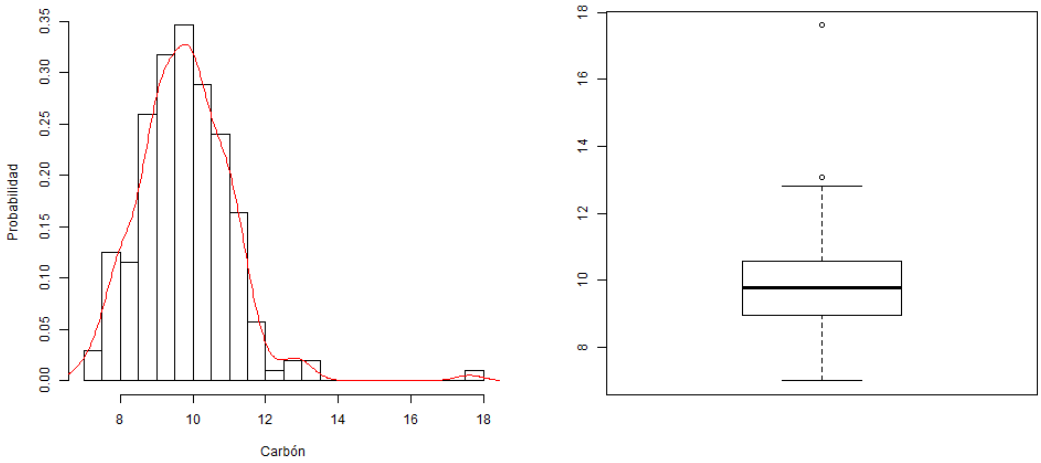

Figura 3.12: Histograma y diagrama de cajas de los porcentajes de carbón.

En Cressie (1993) se puede encontrar el variograma esférico isotrópico ajustado a los datos de carbón utilizados en este caso práctico:

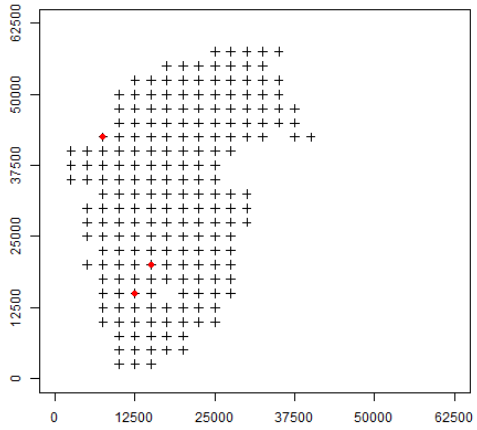

Figura 3.13: Localizaciones de los valores atípicos eliminados para el test de normalidad. Unidades de localización de los ejes en pies (ft).

$$\gamma(t) = \begin{cases} 0, \text{ si } t = 0; \\ 1.78 + 0.28\left(1.5\frac{t}{4.31} - 0.5\left(\frac{t}{4.31}\right)^3\right), \text{ si } 0 < t < 4.31; \\ 2.06, \text{ si } t \geq 4.31. \end{cases}$$

La tendencia del proceso $\mu(s)$ se ha estimado ajustando un modelo de regresión lineal múltiple con errores homocedásticos, usando como variables explicativas las localizaciones:

$$\mu(x_i, y_i) = \beta_0 + \beta_1 x_i + \beta_2 y_i + \epsilon_i,$$

donde (x_i, y_i) es el vector de localizaciones y los errores ϵ_i son independientes tales que $\mathbb{E}(\epsilon_i) = 0 \ \forall i$, $1 \leq i \leq n$.

Sólo ha resultado significativo el parámetro β_1 y el intercepto, con valores -0.184 y 11.162 respectivamente. De esta forma la tendencia estimada es,

$$\hat{\mu}(x_i, y_i) = 11.162 - 0.184 x_i, \tag{3.1}$$

donde (x_i, y_i) es el vector de localizaciones, $1 \leq i \leq n$. Se puede comprobar el ajuste de la tendencia en la Figura 3.14.

Para comprobar la estructura de dependencia propuesta por Cressie (1993), teniendo en cuenta la estimación de la tendencia y el variograma propuesto se ha simulado un proceso estocástico en las localizaciones muestreadas. La media de las diferencias estandarizadas entre los valores reales de carbón y los valores simulados ha sido de 0.037, muy próxima a cero. De manera análoga, la desviación típica obtenida ha sido de 1.14, próxima a 1. Por consiguiente, se considerará que las estimaciones son correctas. En la Figura 3.15 se puede observar el histograma relativo a las diferencias estandarizadas junto con la función de densidad empírica y la función de densidad de la distribución normal estándar.

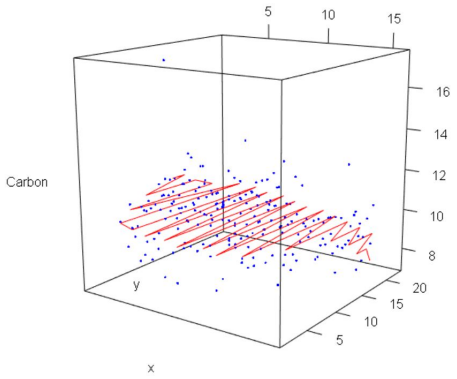

Figura 3.14: Ajuste de la tendencia del proceso.

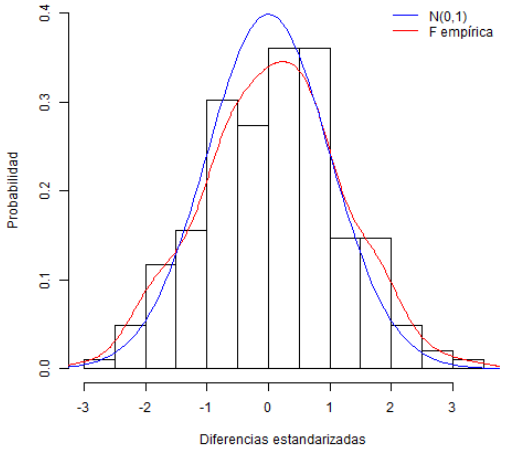

Figura 3.15: Diferencias estandarizadas entre los valores reales de carbón y los valores simulados.

Para la predicción no paramétrica propuesta, $\bar{F}_s(x)$, se ha calculado la matriz de suavizado $H = diag(h_1, h_2)$ mediante el método de validación cruzada presentado en capítulos anteriores. Para la aplicación de la predicción kriging indicadora, $\hat{F}_s(x)$, como ya se ha mencionado en capítulos anteriores, ha sido necesario seguir los siguientes pasos:

1. Estimar la tendencia, con el modelo de regresión lineal ya presentado.
2. Obtener los datos sin tendencia $Y(s) = Z(s) - \mu(s)$ y los valores de corte $x_s = x - \mu(s)$.
3. Construir un estimador no paramétrico del semivariograma para lo cual se ha recurrido al estimador de Matheron.
4. Ajustarlo a un modelo paramétrico, para lo cual se ha empleado el ajuste por mínimos cuadrados ponderados.

En una primera fase de estudio, para comparar el funcionamiento del estimador no paramétrico propuesto frente al estimador clásico kriging indicador se ha llevado a cabo un procedimiento de

validación cruzada. De esta forma, se calculará el valor del error cuadrático medio dado por:

$$MSE_{\bar{F}(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\left(\bar{F}_{s_i}^{(-i)}(x) - I_Z(s_i,x)\right)^2$$

donde $\bar{F}_{s_i}^{(-i)}(x)$ denota la estimación no paramétrica de tipo núcleo de la distribución que se obtendría en la localización s_i al trabajar con todos los datos observados salvo el i-ésimo; y de manera análoga:

$$MSE_{\hat{F}(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{F}_{s_i}^{(-i)}(x) - I_Z(s_i,x)\right)^2$$

donde $\hat{F}_{s_i}^{(-i)}(x)$ denota la estimación kriging indicadora de la distribución que se obtendría en la localización s_i al trabajar con todos los datos observados salvo el i-ésimo.

El estudio del error cuadrático medio para ambos estimadores se realizó sobre un grid de valores de corte x, en concreto se han tomado los cuantiles muestrales P_5, P_{25}, P_{50}, P_{75} y P_{95}, que se corresponden con los valores 7.843, 8.96, 9.785, 10.567 y 11.599 respectivamente. Los resultados se recogen en la Tabla 3.24.

	P_5	P_{25}	P_{50}	P_{75}	P_{95}
$100 \times MSE_{\bar{F}(\cdot)}$	4.5844	15.6062	19.4162	17.1960	5.0785
$100 \times MSE_{\hat{F}(\cdot)_{Matheron}}$	4.9687	17.3706	23.1178	18.1534	5.2662

Tabla 3.24: Errores cuadráticos medios obtenidos mediante validación cruzada, correspondientes al estimador no paramétrico propuesto y al estimador kriging indicador para los valores de corte indicados.

Si se compara el error asociado al estimador no paramétrico propuesto (NP) frente al del kriging indicador (KI) empleando la estimación del variograma de Matheron, el primero ha dado lugar a errores cuadráticos medios más bajos para todos los puntos de corte. Además cabe destacar, como ya se ha mencionado anteriormente, que el estimador propuesto no requiere la estimación de la tendencia del proceso, además de tener menor coste computacional comparado con el estimador KI.

En una segunda fase, haciendo uso del estimador no paramétrico propuesto y el estimador kriging indicador empleando el estimador del semivariograma de Matheron, se construirá un mapa de riesgo para detectar zonas con mayor probabilidad de encontrar más cantidad del mineral. Para construir un mapa de riesgo se ha seleccionado un grid de 839 localizaciones (Figura 3.16) dentro de la región muestreada, en las que estimar la probabilidad $\mathbb{P}(Z > x)$, $x \in \mathbb{R}$, empleando de nuevo la matriz de suavizado $H = diag(h_1, h_2)$ obtenida mediante validación cruzada.

En la Figura 3.17 y 3.18 se recogen los mapas de riesgo asociados a la estimación de la probabilidad $\mathbb{P}(Z > 10)$ y $\mathbb{P}(Z > 11)$ respectivamente, mediante el estimador no paramétrico propuesto y el estimador kriging indicador. Se pueden observar en los mapas de riesgo construidos para el valor de corte $x = 10$ (Figura 3.17) que las localizaciones con mayor probabilidad asociada (en color rosa) están concentradas en la zona oeste para ambos estimadores. Los puntos con probabilidad 1, que se observan dispersos a lo largo de la rejilla para el kriging indicador, se corresponden con las estimaciones realizadas sobre las localizaciones muestreadas con porcentajes de carbón superiores a 10, puesto que los estimadores kriging son interpoladores exactos.

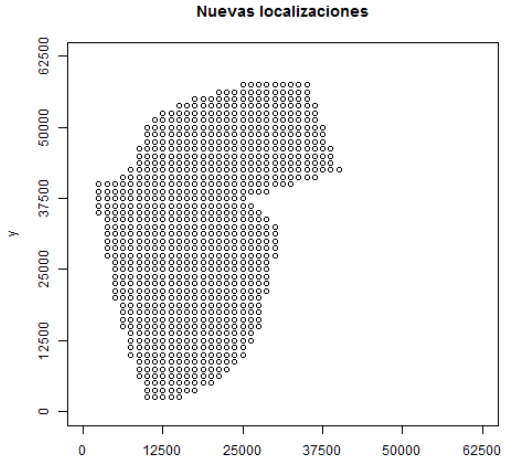

Figura 3.16: Rejilla con las nuevas localizaciones en las que efectuar la predicción. Unidades de localización de los ejes en pies (ft).

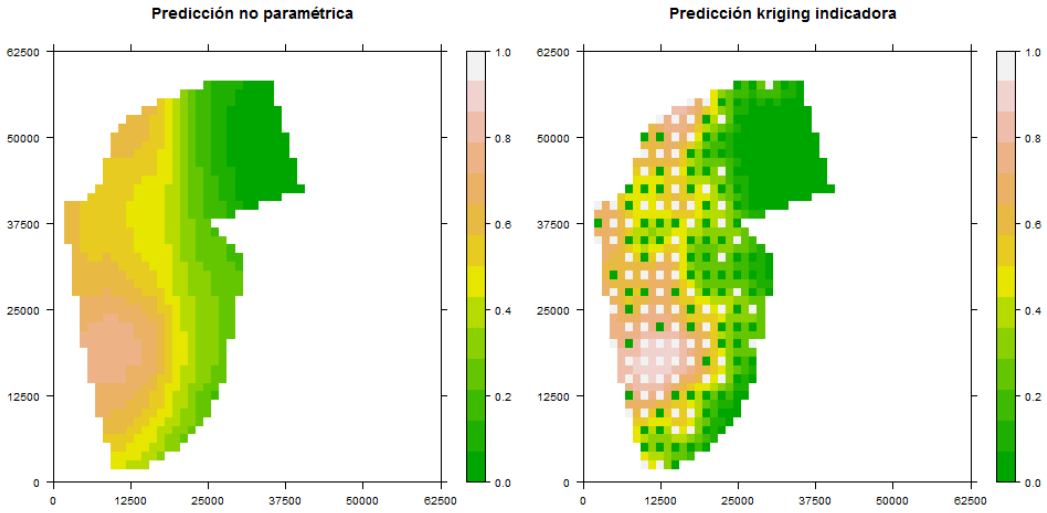

Figura 3.17: Predicciones para $\mathbb{P}(Z > 10)$ mediante el estimador no paramétrico propuesto (izquierda) y el estimador kriging indicador (derecha) sobre la nueva rejilla. Unidades de localización de los ejes en pies (ft).

En el caso de los mapas de riesgo construidos para el valor de corte $x = 11$ (Figura 3.18), muy próximo al cuantil muestral P_{95} de los porcentajes de carbón, se puede apreciar que los valores de la probabilidad han disminuido considerablemente. De nuevo la zona oeste acumula los valores más altos, pero estos no exceden de 0.27 para el caso NP. En el caso del estimador KI las probabilidades son todavía menores, con excepción de los puntos que coinciden con localizaciones muestreadas con porcentajes de carbón superiores a 11 donde, como ya se ha dicho, el estimador KI es interpolador exacto, obteniéndose una probabilidad asociada a la localización con valor 1.

Además, en las Figuras 3.17 y 3.18 se puede observar una ventaja adicional del estimador NP, y es que por la forma de construcción, proporciona una aproximación más suavizada de la función de

distribución que el estimador KI.

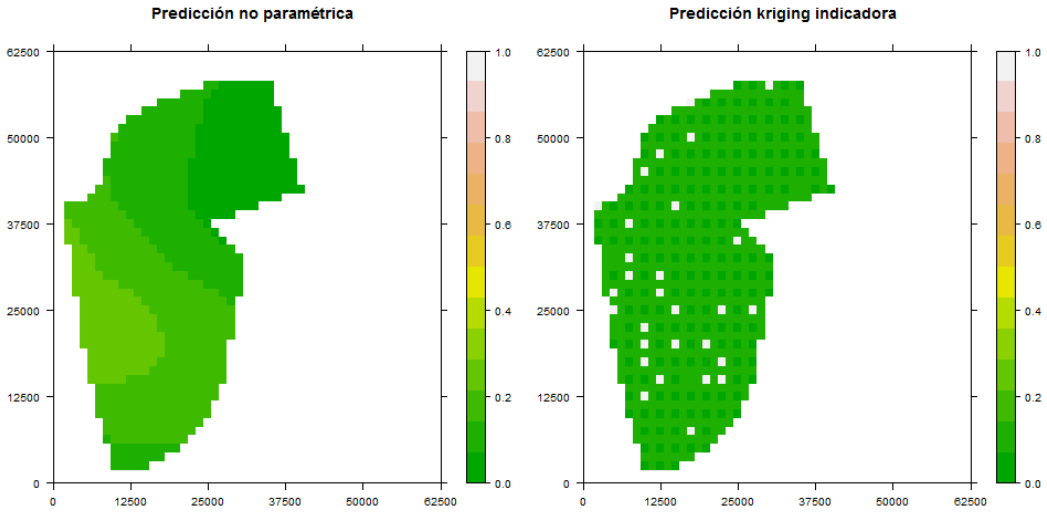

Figura 3.18: Predicciones para $\mathbb{P}(Z > 11)$ mediante el estimador no paramétrico propuesto (izquierda) y el estimador kriging indicador (derecha) sobre la nueva rejilla. Unidades de localización de los ejes en pies (ft).

Capítulo 4

Conclusiones

El presente trabajo fin de máster tenía por objeto analizar el comportamiento de un estimador no paramétrico (NP) de la función de distribución para datos espaciales, introducido como alternativa al mecanismo de aproximación basado en la aplicación del kriging indicador (KI), tradicionalmente utilizado en este contexto. Para el diseño de la propuesta no paramétrica se utilizó una media ponderada de las funciones indicadoras, evaluadas en las localizaciones muestreadas, a las que se asignaron distintos pesos en función de la separación entre dichas posiciones y la localización objeto de estudio.

Una ventaja importante del estimador NP frente a otros procedimientos, en contextos no estacionarios, es que no requiere la caracterización de la tendencia. Sin embargo, la utilización de la metodología de tipo núcleo para la definición de los pesos en la propuesta NP supone la selección de un parámetro ventana o de una matriz ventana, que deben ser adecuadamente elegidos en la práctica, para producir resultados fiables. En particular en este trabajo, se ha recurrido a la técnica de validación cruzada para la propuesta de selectores en ambos casos.

En los estudios numéricos llevados a cabo se observó, como era esperable, un mejor comportamiento de la matriz ventana, frente al parámetro de suavizado unidimensional, por lo que se adoptó este procedimiento para las restantes simulaciones. En este sentido, se diseñaron distintos escenarios, para generar datos de procesos estacionarios y no estacionarios, con distribución gaussiana o no gaussiana, con diferentes tendencias y estructuras de dependencia. En general, los resultados obtenidos ponen de manifiesto el buen funcionamiento en la práctica del estimador NP propuesto frente a la metodología KI, observándose distintos patrones de comportamiento para los diferentes contextos analizados. En este sentido, se ha confirmado que el aumento del tamaño muestral se traduce en una mejora de los resultados del estimador NP en todos los escenarios. Por otro lado, en el caso estacionario, el aumento de la varianza ha dado lugar a mayores errores para ambos estimadores, mientras que el aumento del parámetro pepita ha dado lugar al resultado contrario. La modificación del parámetro meseta no ha dado lugar a diferencias destacables en el caso del estimador KI, pero sí ha provocado un mayor error para el estimador NP conforme éste aumenta.

Si se comparan los resultados entre un escenario estacionario y uno con tendencia lineal, los errores han aumentado en el segundo para ambos estimadores. Sin embargo, en este caso los resultados correspondientes al estimador NP han sido notablemente mejores que los relativos al KI, además de no precisar la estimación de la tendencia $\mu(\cdot)$.

El empleo de estimadores no paramétricos (Nadaraya-Watson y lineal local) en la construcción del variograma necesario para el kriging indicador de nuevo ha dado lugar a peores MSE que la alternativa NP propuesta, empleando para ambos casos la matriz de suavizado bidimensional obtenida mediante validación cruzada. No obstante, los resultados asociados al KI han mejorado en comparación con los resultados del estimador empírico clásico del semivariograma.

Además, se incluyó un breve estudio de escenarios con tendencias no lineales que fueron estimadas mediante modelos lineales generalizados o modelos aditivos generalizados según fuese conveniente. En estos casos de nuevo el estimador NP dio mejores resultados que el estimador KI. Por último, se incluyó

la estimación de la función de distribución mediante ambas metodologías (no paramétrica y kriging indicadora) para un proceso no gaussiano, donde de nuevo el MSE ha sido menor para el estimador propuesto.

En la parte final de este trabajo, se ha querido ilustrar el comportamiento de las técnicas presentadas con datos reales. Concretamente se ha recurrido a un caso, muy conocido en este contexto y que se describe con detalle en Cressie (1993), donde el objetivo es la medición del nivel de carbón en la mina Robena (Condado de Green, Pennsylvania). El proceso espacial subyacente presentaba una tendencia lineal, por lo que era un buen candidato para la aplicación de la metodología NP. Para este estudio, la tendencia fue estimada mediante un modelo de regresión lineal múltiple. Una vez eliminados los datos atípicos, los niveles de carbón han pasado los contrastes de normalidad. A partir de ahí se aplicaron los estimadores NP y KI, empleando para este último la estimación del semivariograma de Matheron. Los correspondientes resultados confirman nuevamente la ventaja del estimador propuesto frente al que proporciona el kriging indicador. Para finalizar esta aplicación se construyeron mapas de riesgo asociados a los niveles carbón con ambas metodologías.

A modo de resumen, podemos concluir que los resultados aportados en este trabajo ponen de manifiesto las ventajas de la utilización del método no paramétrico propuesto para la estimación de la función de distribución de un proceso estocástico espacial. Como complemento a esta investigación, en un futuro podrían abordarse modificaciones sobre el estimador NP que previsiblemente mejorarían su comportamiento cuando el número de datos es reducido, particularmente en la frontera de la región de observación. En este sentido, cabría plantearse la utilización de un núcleo frontera, en vez de la función simétrica de tipo núcleo, o la aplicación de la estimación lineal local, que generalmente reducen el sesgo en las proximidades de la frontera. Asimismo, podrían explorarse otras alternativas de selección de la matriz de suavizado, basadas en la metodología Bootstrap o en la utilización de una estimación de la matriz de varianzas y covarianzas de los datos.

Apéndice A

Notación

$\|t\|$	=	$\sum_{i=1}^{d} \|t_i\|$, para $t = (t_1, ..., t_d) \in \mathbb{R}^d$.
$\|\|t\|\|$	=	$(\sum_{i=1}^{d} t_i^2)^{\frac{1}{2}}$, para $t = (t_1, ..., t_d) \in \mathbb{R}^d$
$\gamma(\cdot)$		Semivariograma.
$\gamma_I(\cdot)$		Semivariograma indicador.
$\hat{\gamma}(\cdot)$		Estimador empírico del semivariograma.
$\bar{\gamma}(\cdot)$		Semivariograma válido.
$\rho(\cdot)$		Correlograma.
$\mu(s)$		Tendencia del proceso estocástico espacial $Z(s)$.
σ^2		Varianza.
$C(\cdot)$		Covariograma o función de covarianza.
$\hat{C}(\cdot)$		Estimador empírico del covariograma o función de covarianza.
c_0		Efecto pepita del semivariograma.
c_1		Umbral o meseta del semivariograma.
c_2		Rango o alcance del semivariograma.
c_2'		Rango o alcance efectivo del semivariograma.
$F_s(x)$		Función de distribución del proceso estocástico espacial $\{Z(s)/s \in D\}$ en la localización s para el valor de corte x.
$\bar{F}_s(x)$		Estimador indicador no paramétrico (NP) de la función de distribución F en la localización s para el valor de corte x.

$\hat{F}_s(x)$	Estimador kriging indicador (KI) de la función de distribución F en la localización s para el valor de corte x.
h	Parámetro de suavizado o parámetro ventana.
h_{CV}	Parámetro de suavizado o parámetro ventana obtenido por validación cruzada.
H	Matriz de suavizado.
H_{CV}	Matriz de suavizado obtenida por validación cruzada.
MSE	Error cuadrático medio.
MSE^I	Error cuadrático medio empleando la función indicadora.
$K(\cdot)$	Función de tipo núcleo.
$K_\nu(\cdot)$	Función de Bessel de segunda clase y orden ν modificada.
s_i	Localizaciones muestreadas.
$s^{(k)}$	Nuevas localizaciones en las que predecir.
x	Valor de corte.
$Y(s)$	Componente de variación aleatoria del proceso estocástico espacial $Z(s)$.
$Z(s)$	Variable aleatoria asociada a la localización $s \in D$, $\{Z(s)/s \in D\}$ un proceso estocástico espacial.
$\hat{Z}(s)$	Predictor de $Z(s)$.

Bibliografía

[1] Cressie N (1993) Statistics for Spatial Data. John Wiley and Sons, New York.

[2] Cressie N y Hawkins DM (1980) Robust estimation of the variogram. Mathematical Geology 12: 115-125.

[3] Díaz Viera M (2002) Geoestadística Aplicada. Instituto de Geofísica y Astronomía, CITMA, Cuba.

[4] Evans M y Rosenthal J (2005) Probabilidad y estadística. Reverté, Barcelona.

[5] Fernández Casal R (2003) Geoestadística espacio-temporal: Modelos flexibles de variogramas anisotrópicos no separables. Tesis, Universidad de Santiago de Compostela.

[6] Fernández Casal R (2015) Nonparametric spatial (geo)statistics. R package version 0.3-6. https://cran.r-project.org/web/packages/npsp/npsp.pdf. Accedido 20 junio de 2016.

[7] García-Soidán P, González-Manteiga W y Febrero-Bande M (2003) Local linear regression estimation of the variogram. Statistics and Probability Letters 64: 169-179.

[8] García-Soidán P, González-Manteiga W y Febrero-Bande M (2004) Nonparametric kernel estimation of an isotropic variogram. Journal of Statistical Planning and Inference 121: 65-92.

[9] García-Soidán P (2007) Asymptotic normality of the Nadaraya-Watson semivariogram estimators. TEST 16: 479-503.

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

[11] Gómez M y Hazen K (1970) Evaluating sulfur and ash distribution in coal seems by statistical response surface regression analysis. U.S. Bureau of Mines Report RI 7377.

[12] Goovaerts P (1997) Geostatistics for natural resources evaluation. Oxford University Press, Oxford.

[13] Hall P y Patil P (1994) On the nonparametric estimation of covariance functions. Annals of Statistics 22: 2115-2134.

[14] Journel AG (1983) Nonparametric estimation of spatial distribution. Mathematical Geology 15(3): 445-468.

[15] Journel AG y Huijbregts CJ (2003) Mining geostatistics. Blackburn Press, Caldwell (New Jersey).

[16] Leeuw J, Hornik K y Mair P (2009) Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. Journal of Statistical Software 32 (5): 1-24.

[17] Leeuw J, Hornik K y Mair P (2015) Active Set and Generalized PAVA for Isotone Optimization. R package version 1.1-0. https://cran.r-project.org/web/packages/isotone/isotone.pdf. Accedido 20 de junio de 2016.

[18] Mardia KV y Watkins AJ (1989) On multimodality of the likelihood in the spatial linear model. Biometrika 76 (2): 289-295.

[19] Martínez Ruiz F (2008) Modelización de la función de covarianza en modelos espacio-temporales: análisis y aplicaciones. Tesis, Universidad de Valencia.

[20] Matheron G (1963) Principles of geostatistics. Economic Geology 58: 1246-1266.

[21] Matheron G (1971) The theory of regionalized variables and its applications. École national supérieure des mines, París.

[22] Montero Lorenzo JM y Larraz Iribas B (2008) Introducción a la geoestadística lineal. Netbiblo, A Coruña.

[23] Pebesma E y Bivand R (2016) Classes and Methods for Spatial Data. R package version 1.2-3. https://cran.r-project.org/web/packages/sp/sp.pdf. Accedido 26 de junio de 2016.

[24] Pebesma E y Graeler B (2016) Spatial and Spatio-Temporal Geostatistical Modelling, Prediction and Simulation. R package version 1.1-3. https://cran.r-project.org/web/packages/gstat/gstat.pdf. Accedido 20 de junio de 2016.

[25] Ribeiro P y Diggle R (2016) Analysis of Geostatistical Data. R package version 1.7-5.2. https://cran.r-project.org/web/packages/geoR/geoR.pdf. Accedido 20 de junio de 2016.

[26] Yu K, Mateu J y Porcu E (2006) A kernel-based method for nonparametric estimation of variograms. Statistica Neerlandica 60: 1-25.