Trabajo Fin de Máster

Aplicación de técnicas *binning* en geoestadística no paramétrica

Carlos García Muñoz

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Aplicación de técnicas de *binning* en xeoestatística non paramétrica
Título en español: Aplicación de técnicas *binning* en geoestadística no paramétrica
English title: Application of *binning* techniques in nonparametric geostatistics
Modalidad: Modalidad A
Autor: Carlos García Muñoz, Universidad de A Coruña
Director: Rubén Fernández Casal, Universidad de A Coruña; Mario Francisco Fernández, Universidad de A Coruña
Tutor/a: Tutor/a 1, Empresa 1; Tutor/a 2, Empresa 2
Breve resumen del trabajo: Las técnicas *binning* se han utilizado para el cálculo rápido de estimadores tipo núcleo (Wand, 1994), así como para el cálculo de cantidades auxiliares relacionadas (Turlach y Wand, 1996). En este TFM se estudiarán extensiones de estos enfoques para datos geoestadísticos. En concreto, para el caso de un proceso geoestadístico, se estudiará el empleo de *binning* para la estimación no paramétrica de la tendencia y del variograma, además de cantidades auxiliares tales como medidas de grados de libertad, errores de validación cruzada, estimaciones de la varianza y medidas de error. Una parte importante del trabajo consistirá en la realización de simulaciones bajo distintos escenarios utilizando el software R y el paquete npsp. Referencias: Ruben Fernandez-Casal (2019). npsp: Nonparametric Spatial Statistics. R package version 0.8. http://github.com/rubenfcasal/npsp Turlach, B.A. and Wand, M.P. (1996). Fast Computation of Auxiliary Quantities in Local Polynomial Regression. Journal of Computational and Graphical Statistics. 5 (4), 337-350. Wand M.P. (1994). Fast Computation of Multivariate Kernel Estimators. Journal of Computational and Graphical Statistics, 3, 433-445.

Recomendaciones: Es especialmente recomendable haber cursado las materia del MTE "Estadística Espacial" , así como la materia "Estadística No Paramétrica", del antiguo plan de estudios, o la materia "Regresión No Paramétrica y Semiparamétrica", del actual plan de estudios.

Otras observaciones: Este TFM se podría considerar también como una fase previa a la realización de una posible tesis doctoral en el campo de la geoestadística no paramétrica (entre otros temas abiertos que podrían ser de interés estarían la selección de la ventana bajo heterocedasticidad, métodos bootstrap o modelado espaciotemporal). Por tanto, podría ser adecuado para aquellos estudiantes que deseen encauzarse hacia la investigación en este campo.

Don Rubén Fernández Casal, profesor contratado doctor de la Universidad de A Coruña, don Mario Francisco Fernández, Catedrático de la Universidad de A Coruña, informan que el Trabajo Fin de Máster titulado

Aplicación de técnicas *binning* en geoestadística no paramétrica

fue realizado bajo su dirección por don Carlos García Muñoz para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En A Coruña, a 2 de Febrero de 2022.

El director:

FERNANDEZ CASAL RUBEN - 33317724Q

Firmado digitalmente por FERNANDEZ CASAL RUBEN - 33317724Q
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-33317724Q, givenName=RUBEN, sn=FERNANDEZ CASAL, cn=FERNANDEZ CASAL RUBEN - 33317724Q
Fecha: 2022.02.02 11:02:44 +01'00'

Don Rubén Fernández Casal

El director:

FRANCISCO FERNANDEZ MARIO - 32660323R

Firmado digitalmente por FRANCISCO FERNANDEZ MARIO - 32660323R
Fecha: 2022.02.02 13:20:02 +01'00'

Don Mario Francisco Fernández

El/la tutor/a:

Don/doña Tutor/a 1

El/la tutor/a:

Don/doña Tutor/a 2

El autor:

Don Carlos García Muñoz

Agradecimientos

En primer lugar, me gustaría agradecer a mis tutores Rubén y Mario, ya que sin ellos este trabajo no hubiera sido posible. Gracias por ayudarme a desarrollar mis conocimientos en Geoestadística y en simulaciones.

También quiero agradecer el esfuerzo que han hecho todos los profesores del máster en técnicas estadísticas, que en estos años tan especiales debido al Covid, han conseguido que se notase lo menos posible en el desarrollo del máster, y gracias a ellos he podido continuar tanto mi desarrollo personal como profesional.

Por último y más importante, gracias a mi familia, que llevan apoyándome durante toda mi vida, y que, en los momentos más complicados, han estado ahí para darme la fuerza que necesitaba. Os quiero mucho.

Índice general

Resumen

Resumen en español

Este trabajo se centra en el estudio del comportamiento del *binning*, y más en concreto, en comprobar si la reducción de los tiempos computacionales que se obtiene al implementarlo, ya que se reduce el número de cálculos a realizar, se traduce en una pérdida de eficiencia estadística de los métodos de estimación empleados sobre este conjunto de datos modificado. Además, es necesario destacar que este estudio se desarrollará para el caso de datos espaciales, y se considerará tanto un conjunto de datos reales como datos que han sido creados mediante simulaciones.

Para llevar a cabo este análisis, el trabajo se centrará en la estimación no paramétrica de dos funciones de interés en este contexto: la tendencia espacial y el variograma. Estas comparaciones se llevarán a cabo en base a simulaciones y al estudio de datos reales, por lo que el software R tendrá un papel fundamental en el desarrollo del presente trabajo. Por último, basándonos en todos los estudios realizados a lo largo de este trabajo se tratará de concluir si se recomienda el empleo del *binning* en comparación con el procedimiento habitual.

English abstract

This thesis focuses on the study of the behavior of *binning*, and more specifically, on checking whether the reduction in computational times obtained by implementing it, where the number of calculations required is less, translates into a loss of statistical efficiency of the estimation methods used on this set of modified data. In addition, it is necessary to emphasize that this study will be developed for the case of spatial data, and both a real data set and data that have been created through simulations will be considered.

To carry out this analysis, the work will focus on the non-parametric estimation of two functions of interest in this context: the spatial trend and the variogram. These comparisons will be carried out based on simulations and the study of real data, so the R software will have a fundamental role in the development of this work. Finally, based on all the studies developed throughout this work, we will try to conclude if the use of *binning* is recommended compared to the usual procedure.

Capítulo 1

1. Introducción a la geoestadística

El primer capítulo del trabajo se va a centrar en la introducción a los datos espaciales y la metodología que se emplea para su tratamiento, así como los conceptos teóricos de interés que atañen a este tipo de procesos. En la sección 1.1. se presentarán los procesos geoestadísticos, su modelización y, además, se estudiará su estacionariedad. La sección 1.2. se estudiará el variograma, que es una función de gran importancia dentro de la estadística espacial y que se empleará para realizar estudios sobre el *binning* en capítulos posteriores. Por último, la sección 1.3. se centrará en las predicciones espaciales, y más concretamente en el *kriging*, aunque este concepto solo se tratará teóricamente en este trabajo, puede servir de introducción para futuros estudios.

1.1. Los procesos geoestadísticos

La geoestadística es una rama de la estadística, y más en concreto de la estadística espacial, que surgió en la segunda mitad del siglo XX para afrontar un tipo de problemas existentes que los métodos tradicionales no eran capaces de tratar de una manera adecuada, en concreto, problemas en los que los datos dependen de la posición espacial en la que se encuentran. Esta nueva metodología se distingue del resto de aproximaciones en que, no solo tiene en cuenta la tendencia espacial, sino que también le da importancia a la correlación espacial. Este último término se basa en la idea de que, las observaciones que se encuentren a una distancia pequeña van a ser muy parecidas entre sí, pero a medida que se aumenta la separación espacial entre ellas, el grado de similitud se reduce. Esta asunción resulta de gran utilidad para múltiples campos de estudio como, por ejemplo: minería, geografía o agricultura.

Por otra parte, los datos espaciales pueden dividirse en tres categorías en función del dominio espacial sobre el que se encuentran definidos: procesos espaciales continuos, procesos espaciales discretos y procesos o patrones puntuales. De entre ellos, la geoestadística es la rama que se encarga del análisis de datos espaciales que cumplen las características:

- Las observaciones de los valores de la variable son realizadas sobre un conjunto de localizaciones muestrales discreto que se encuentran dentro de una región espacial.
- Cada una de las observaciones mide un fenómeno espacial continuo dentro de las localizaciones muestrales consideradas o de una región.

A la hora de modelizar los procesos geoestadísticos, generalmente se suele emplear un modelo que distinga entre la variación a gran escala y la variación a pequeña escala. Por ello, el modelo

general que normalmente se emplea es:

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \varepsilon(\boldsymbol{x}) \quad (1.1)$$

Donde $\mu(\boldsymbol{x})$ la media condicional pues $\mu(\boldsymbol{x}) = E(Y|_{\boldsymbol{X}=\boldsymbol{x}})$, y que también es comúnmente conocida como función de regresión o tendencia (y que es una función determinística), que es la que representa la variación a gran escala.

- Y siendo $\varepsilon(\boldsymbol{x})$ un error aleatorio de media 0 y de varianza constante, en el que se incluye la dependencia espacial y además es la componente que representa a la variación a pequeña escala.

La dependencia espacial se puede incorporar a través de la covarianza:

$$Cov\left[Y(\boldsymbol{x}_i), Y(\boldsymbol{x}_j)\right] = E\left[\left(Y(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)\right)\ \left(Y(\boldsymbol{x}_j) - \mu(\boldsymbol{x}_j)\right)\right]$$

Sin embargo, para poder realizar inferencia, en necesario asumir ciertas hipótesis de estacionariedad para $\varepsilon(\boldsymbol{x})$. Por ello, se van a estudiar a continuación los tipos de estacionariedad que aparecen de manera más habitual en la geoestadística.

1.1.1. Estacionariedad de los procesos geoestadísticos

Por lo general, cuando se emplean procesos geoestadísticos, se asumen algunas condiciones de estacionariedad sobre $\varepsilon(\boldsymbol{x})$. Esto es debido a que normalmente no se dispone del proceso completo, sino que solo se tiene una realización parcial. Para ello, en esta sección consideraremos un proceso geoestadístico sin tendencia (que equivaldría al proceso de error en el modelo general anterior), de la forma:

$$\{Y(\boldsymbol{x}) : \boldsymbol{x} \in D \subset R^d\}$$

En el que cada una de las $\boldsymbol{x}$ representa una localización espacial, estando todas ellas contenidas en una región de observación, que es denotada por D. Una vez mostrada esta expresión, se puede presentar la de la distribución conjunta de este tipo de procesos:

$$F_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(y_1,\ldots,y_n) = P(Y(\boldsymbol{x}_1) \leq y_1,\ldots,Y(\boldsymbol{x}_n) \leq y_n)$$

Y, partiendo de la distribución conjunta de los procesos geoestadísticos, se puede estudiar su estacionariedad. En primer lugar, podríamos considerar:

- Estacionariedad estricta: es el tipo de estacionariedad más restrictiva, y que consiste en que la distribución conjunta se mantiene invariante ante una traslación (siendo indistinta la dirección) de una configuración de posiciones espaciales. Esto se puede expresar como, considerando un salto $\boldsymbol{u}$:

$$F_{\boldsymbol{x}_1+\boldsymbol{u},\ldots,\boldsymbol{x}_n+\boldsymbol{u}}(y_1,\ldots,y_n) = F_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}(y_1,\ldots,y_n)\ \forall \boldsymbol{x}_i \text{ con } i = 1,\ldots n; \forall \boldsymbol{u} \in D; \forall n \in N$$

- Estacionariedad de segundo orden: este tipo es menos restrictivo que el anterior y se puede obtener a partir de los momentos de orden uno y dos del proceso geoestadístico. Se considera que un proceso es estacionario de segundo orden si se cumplen las dos siguientes condiciones:

$$E[Y(\boldsymbol{x})] = \mu, \forall \boldsymbol{x} \in D$$

$$Cov[Y(\boldsymbol{x}), Y(\boldsymbol{x}+\boldsymbol{u})] = C(\boldsymbol{u}), \forall \boldsymbol{u} \in D$$

La segunda de las hipótesis contiene a la función $C(\cdot)$ que es el covariograma (también llamado función de covarianza) y que, como puede verse en la expresión, no depende $\boldsymbol{x}$, solo de $\boldsymbol{u}$. Además, dentro de los procesos estacionarios de segundo orden hay dos tipos, los isotrópicos y los anisotrópicos. Los procesos de segundo orden isotrópicos son aquellos en los que el covariograma no depende de la dirección del salto, sino que solo se ve afectado por su magnitud. Por el contrario, en los anisotrópicos el covariograma depende de ambos factores. Además, es necesario destacar que todo proceso estrictamente estacionario cuyos momentos de primer y segundo orden son finitos es también un proceso estacionario de segundo orden, por lo que estas propiedades son equivalentes para los procesos gaussianos.

Otro concepto de interés en los procesos geoestadísticos es el correlograma, que en ocasiones puede emplearse en lugar del covariograma. Partiendo de esto, y considerando $C(0) = Var(Y(\boldsymbol{x})) > 0$, se puede obtener la expresión que define al correlograma:

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

Por otro lado, hay un tipo de procesos aleatorios que no tienen definida la varianza, pero en los que sus diferencias o incrementos son estacionarios de segundo orden. Este tipo se clasificaría dentro de un tipo de procesos más generales, que comúnmente son conocidos como procesos intrínsecamente estacionarios (o procesos intrínsecos). Las dos principales características que tienen este tipo de procesos son:

$$E[Y(\boldsymbol{x} + \boldsymbol{u}) - Y(\boldsymbol{x})] = 0, \forall \boldsymbol{x} \in D$$

$$Var[Y(\boldsymbol{x}) - Y(\boldsymbol{x} + \boldsymbol{u})] = 2\gamma(\boldsymbol{u}), \forall \boldsymbol{u} \in D$$

En la segunda de ellas aparece la función $2\gamma(\cdot)$ que es el variograma (se estudiará más en profundidad en la sección 1.2.), mientras que al semivariograma se le denotaría como $\gamma(\cdot)$. Además, del mismo modo que sucedía con el covariograma, tanto el semivariograma como el variograma son independientes de $\boldsymbol{x}$, es decir, solo se ven afectados por $\boldsymbol{u}$, y este tipo de procesos pueden dividirse también en intrínsecamente estacionarios anisotrópicos (dependen de la dirección y magnitud del salto) e isotrópicos (solo les influye la magnitud del salto).

Por último, destacar que esta clase de procesos son más generales con respecto a los estacionarios de segundo orden. Esto puede demostrarse de la siguiente manera, si el covariograma de un proceso estacionario de segundo orden es $C(\cdot)$, entonces:

$$Var[Y(\boldsymbol{x}) - Y(\boldsymbol{x} + \boldsymbol{u})] = Var[Y(\boldsymbol{x})] + Var[Y(\boldsymbol{x} + \boldsymbol{u})] - 2Cov[Y(\boldsymbol{x}), Y(\boldsymbol{x} + \boldsymbol{u})]$$

$$= 2C(0) - 2C(\boldsymbol{u})$$

De tal forma que el semivariograma es:

$$\gamma(\boldsymbol{u}) = C(0) - C(\boldsymbol{u})$$

Esto significa que es un proceso intrínsecamente estacionario, mientras el recíproco no tiene por qué cumplirse (una demostración de esto puede verse para un movimiento browniano en Cressie, 1993), aunque sí que ocurre en una gran parte de los casos. Esta ausencia de reciprocidad generalmente suele deberse a que la media del proceso no sea constante. En esa situación, habría que utilizar la expresión:

$$E[(Y(\boldsymbol{x}) - Y(\boldsymbol{x} - \boldsymbol{u}))]^2 = 2\gamma(x, x + \boldsymbol{u}) + (\mu(\boldsymbol{x}) - \mu(\boldsymbol{x} + \boldsymbol{u}))^2$$

Aunque en este apartado se han presentado múltiples conceptos de gran relevancia en la geoestadística, la siguiente sección se va a centrar en un estudio más profundo del variograma.

1.2. El variograma

Este concepto, que ya apareció en la sección anterior, va a ser de gran importancia a lo largo de este trabajo, ya que en capítulos posteriores se realizarán estudios sobre el *binning* que requerirán de su estimación. El variograma o semivariograma es un instrumento analítico que se emplea para el estudio del comportamiento espacial sobre el área definida de una variable. Además, también se puede representar la expresión teórica del variograma. Para llevar a cabo la estimación del variograma, de manera general suele emplearse el estimador empírico (también llamado estimador clásico) del semivariograma, cuya expresión es:

$$\hat{\gamma}(\boldsymbol{u}) = \frac{1}{2|N(\boldsymbol{u})|} \sum_{i,j \in N(\boldsymbol{u})} \left[Y(\boldsymbol{x}_i) - Y(\boldsymbol{x}_j)\right]^2$$

En esta expresión las funciones $Y(\boldsymbol{x})$ representan la variación espacial y $|N(\boldsymbol{u})|$ es la cantidad de pares distintos en $N(\boldsymbol{u})$, siendo $N(\boldsymbol{u}) = \{(i,j): \boldsymbol{x}_i - \boldsymbol{x}_j \in Tol(\boldsymbol{u})\}$, en la que $Tol(\boldsymbol{u}) \subset R$ hace referencia a una región de tolerancia que debe ser fijada previamente. Además, puede resultar de interés presentar las propiedades teóricas del semivariograma y que son extensibles al variograma:

- $\gamma(0) = 0$
- $\gamma(\boldsymbol{u}) \geq 0, \forall \boldsymbol{u} \in R^d$
- $\gamma(\boldsymbol{u}) = \gamma(-\boldsymbol{u}), \forall \boldsymbol{u} \in R^d$

Estas expresiones muestran que tanto el semivariograma como el variograma son funciones no negativas y simétricas. Otro factor a tener en cuenta sobre el variograma son sus características geométricas. Para poder estudiarlas, se puede emplear (ver por ejemplo Fernández-Casal, 2003):

- Efecto *nugget* o efecto pepita: aunque teóricamente el variograma es nulo en el origen, hay ocasiones en las que, por errores de medida o debido a la escala espacial del muestreo (entre otras cosas), este valor no es 0, y es este valor lo que se conoce como efecto *nugget*, que se definiría formalmente como: $c_0 = \lim_{\|\boldsymbol{u}\| \to 0} \gamma(\boldsymbol{u})$. Por otro lado, hay un tipo especial de efecto *nugget* conocido como efecto *nugget* puro, que es aquel en el que la dependencia permanece constante en todos los saltos considerados.

- Umbral: cuando el variograma está acotado y tiene un límite, el valor de este límite es lo que denominamos umbral, que se puede expresar matemáticamente como:

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

En el caso de que el proceso sea estacionario de segundo orden, y si asumimos además que: $\lim_{\|u\|\to\infty} C(\boldsymbol{u}) = 0$, esto implicaría que $\sigma^2 = C(0)$. Además, puede resultar de interés estudiar cómo actúa el variograma en los saltos grandes, pues esto permitiría detectar la presencia de "deriva" o dependencia espacial.

- Umbral parcial: sirve para medir el grado de dependencia espacial que hay en los datos. Si hay efecto *nugget* en un variograma, el umbral parcial sería la diferencia: $c_1 = \sigma^2 - c_0$.
- Rango: partiendo de que hay un umbral, la expresión del rango del semivariograma en dirección $\boldsymbol{e}_0 = \boldsymbol{u}_0/\|\boldsymbol{u}_0\|$ sería:

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

Destacar que es posible que haya variogramas no acotados, esto se produciría cuando no existe ningún umbral. Además, como se estudia en Armstrong (1998) hay casos, como por ejemplo el de los variogramas anisotrópicos, en los que el rango no coincide para todas las direcciones. En el caso en el que el umbral se alcanza de manera asintótica, lo que comúnmente se hace es redefinir a como el rango práctico a la distancia a la que el variograma llega al 95% del umbral parcial.

- Anisotropía: esta característica se produce cuando la forma del variograma $\gamma(\boldsymbol{u})$ se ve afectada por la dirección del salto $\boldsymbol{u}$. Para ciertos casos es posible solucionar este problema mediante una transformación lineal de las coordenadas del salto $\boldsymbol{u}$, situación que se conoce como anisotropía geométrica y los variogramas direccionales en este caso tendrían el mismo umbral pero distintos rangos. Otra posibilidad es que el variograma únicamente dependa de alguna dirección o componente de salto, situación que se conoce como anisotropía zonal o estratificada. Para solucionarlo, lo que generalmente se hace es descomponer el variograma en una parte isotrópica más otro variograma que actúe solo en esa dirección.

Uno de los principales inconvenientes que tiene la estimación del variograma es que puede tener unos costes computacionales muy elevados, en especial cuando el conjunto de datos es muy grande. Una de las formas de solucionar este problema es mediante la transformada rápida de Fourier (FFT, de aquí en adelante). Este algoritmo, que es tratado por ejemplo por Marcotte (1996), permite reducir la cantidad de operaciones a realizar, siendo la complejidad computacional cuando los datos se encuentran en una rejilla regular de dimensiones $N \times M$, de:

$$\sum_{k}\sum_{l}(N-k)(M-l)$$

en la que k y l denotan a los diferentes desplazamientos que se pueden hacer en las direcciones x e y. Además, el dominio en el que se definen estos parámetros depende del número de *lags* y de direcciones para las que debe calcularse el variograma.

La implementación del algoritmo de FFT, a partir de la matriz $N \times M$, que es alargada con ceros hasta tener unas dimensiones $(2N-1) \times (2M-1)$ para que los resultados obtenidos por el FFT sean exactamente los mismos que se conseguirían a través del procedimiento habitual. Por tanto, la complejidad cuando se emplea el FFT, y asumiendo que $M < N$, es de:

$$(2M-1)(2N-1)\log_2(2N-1)$$

Otra gran desventaja que tienen los estimadores del variograma es que no se pueden utilizar para realizar predicción espacial (*kriging*), lo que es debido a que no son condicionalmente semidefinidos negativos. Para resolver este problema, la solución a la que generalmente se recurre es la de seleccionar un modelo paramétrico que resulte adecuado para la dependencia espacial de los datos en estudio.

Existen múltiples criterios para escoger qué modelo utilizar, algunos ejemplos son: mínimos cuadrados ordinarios, mínimos cuadrados generalizados o máxima verosimilitud. En este trabajo se va a optar por los mínimos cuadrados ponderados (WLS, de aquí en adelante). Partiendo de un variograma teórico $2\gamma(\boldsymbol{u}, \boldsymbol{\theta}_0)$ y denotando a las estimaciones del semivariograma que se obtuvieron a través de algún tipo de estimador piloto como: $\hat{\gamma}_i = \hat{\gamma}(\boldsymbol{u_i}), i = 1, \dots, K$. El parámetro $\boldsymbol{\theta}_0$ se estimaría por WLS minimizando:

$$\big(\hat{\gamma} - \gamma(\boldsymbol{\theta})\big)^T \boldsymbol{V}(\boldsymbol{\theta})\big(\hat{\gamma} - \gamma(\boldsymbol{\theta})\big)$$

Expresión en la que $\hat{\gamma} = \big(\hat{\gamma}(\boldsymbol{u_1}), \dots, \hat{\gamma}(\boldsymbol{u}_K)\big)^T$, siendo $\gamma(\boldsymbol{\theta}) = \big(\gamma(\boldsymbol{u_1}; \boldsymbol{\theta}), \dots, \gamma(\boldsymbol{u}_K; \boldsymbol{\theta})\big)^T$ y en la que $\boldsymbol{V}(\boldsymbol{\theta})$ representa a una matriz de dimensiones $K \times K$, que es semidefinida positiva y que para el caso de WLS es $\boldsymbol{V}(\boldsymbol{\theta}) = diag\big(w_1(\boldsymbol{\theta}), \dots, w_K(\boldsymbol{\theta})\big)$, siendo $w_i(\boldsymbol{\theta}) \geq 0$, con, $i = 1, \dots, K$. Los pesos se suelen establecer generalmente como:

$$\boldsymbol{W}_i(\boldsymbol{\theta}) = \frac{|\boldsymbol{N}(\boldsymbol{u}_i)|}{\gamma(\boldsymbol{u_i}, \boldsymbol{\theta})^2}$$

que son aproximadamente inversamente proporcionales a $Var\left(\hat{\gamma}(\boldsymbol{u_i})\right)$ (ver por ejemplo Cressie, 1993). Una de las principales propiedades que tiene este ajuste es que el residuo en el salto $\boldsymbol{u}_i$ recibe un mayor peso cuanto mayor es $|N(\boldsymbol{u}_i)|$ y, además, a menor valor del variograma teórico, mayor es el peso que recibe el correspondiente residuo. Esta propiedad suele traducirse en que los saltos cercanos al origen tienen unos pesos elevados, lo que produce que el variograma se ajuste bien en el origen. A diferencia de los estimadores piloto, a partir de estos variogramas sí que sería posible realizar correctamente *kriging*.

1.3. El *kriging*

A pesar de que este trabajo está orientado hacia la estimación de la tendencia (concepto presentado en el Capítulo 2) y en menor medida a la estimación del variograma (concepto presentado en la sección 1.2.), a partir de estos modelos también se pueden realizar predicciones, que en geoestadística reciben el nombre de *kriging*. Este método es de gran utilidad cuando hay presencia de dependencia espacial, pues en esta situación el estimador de la tendencia no sería el predictor óptimo (ver por ejemplo Cressie, 1990).

Los métodos *kriging* dependen del modelo que se tome para la función aleatoria $Y(\boldsymbol{x})$ que se presentó en (1.1), y está compuesta por dos elementos: la tendencia determinística espacial y el proceso de error, que es estacionario de media cero y en el que se incluye la dependencia espacial. En caso de que no haya un modelo ajustado, no puede utilizarse el *kriging*, por lo que va a tener una gran importancia conseguir un modelo válido.

Por tanto, a partir de una muestra de n observaciones de un proceso espacial $\mathrm{Y}(\cdot)$, de la forma: $\mathrm{Y} = (Y(\boldsymbol{x}_1), \dots, Y(\boldsymbol{x}_n))^t$, que están situadas en unas posiciones espaciales conocidas: $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$, los algoritmos *kriging* permiten obtener a partir de la estructura de segundo orden de este proceso espacial, los predictores lineales óptimos, siendo la predicción para una localización no observada $\boldsymbol{x}_0$, denotada por $\hat{Y}(\boldsymbol{x}_0)$, que puede expresarse formalmente como:

$$\hat{Y}(\boldsymbol{x}) = \lambda_0 + \sum_{\alpha=1}^{n(x)} \lambda_\alpha(\boldsymbol{x}) Y(\boldsymbol{x}_\alpha)$$

Expresión en la que $\lambda_\alpha(\boldsymbol{x})$ es un determinado peso que se le asigna a la observación $y(\boldsymbol{x}_\alpha)$, la cual representa a una realización de $Y(\boldsymbol{x}_\alpha)$, y donde $\mu(\boldsymbol{x})$ y $\mu(\boldsymbol{x}_\alpha)$ hacen referencia a los respectivos valores esperados para $Y(\boldsymbol{x})$ y $Y(\boldsymbol{x_\alpha})$. También hay que destacar que no siempre se emplean todos los datos de los que se dispone, sino que simplemente se tienen en cuenta los $n(\boldsymbol{x})$ que se encuentren más cercanos al punto en el que se quiera realizar la estimación de $\boldsymbol{x}$, es decir, se tienen en cuenta los vecinos más cercanos, estando este vecindario centrado en $\boldsymbol{x}$ y que se representa como $n(\boldsymbol{x})$ (ver por ejemplo Reyes, 2010) Por tanto, el método *kriging* pretende minimizar $\sigma_E^2(\boldsymbol{x})$, es decir, la varianza del error de predicción, cuya expresión es:

$$\sigma_E^2(\boldsymbol{x}) = Var\{\, \hat{Y}(\boldsymbol{x}) - Y(\boldsymbol{x}) \,\}$$

Teniendo como requisito que el estimador sea insesgado, es decir:

$$E\big[\, \hat{Y}(\boldsymbol{x}) - Y(\boldsymbol{x})\big] = 0$$

Además, dentro de las técnicas *kriging* existen tres tipos (ver por ejemplo Castillo, 2017), que están basados en la suposición que se asume sobre la media del proceso:

- *Kriging* simple (KS): se basa en que $\mu(x)$ es conocida. Partiendo de un vector de tendencias conocido $\boldsymbol{\mu} = \big(\mu(\boldsymbol{x}_1), \dots, \mu(\boldsymbol{x}_n)\big)^t$, se obtiene la expresión:

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

- *Kriging* ordinario (KO): en este método se supone $\mu(\boldsymbol{x})$ desconocida pero constante.
- *Kriging* universal (KU): las suposiciones que se hacen sobre $\mu(\boldsymbol{x})$ en este método son que es desconocida, no constante pero que se puede expresar como una combinación lineal de funciones conocidas.

Por último, es necesario destacar que los distintos tipos de *kriging* están relacionados entre sí. Un ejemplo de estas relaciones es que utilizar el método KU es equivalente al KS, pero usando la media estimada. Un método que se podría emplear en futuros trabajos sería el *kriging* residual con tendencia no paramétrica (o *kriging* con tendencia externa no paramétrica).

Capítulo 2

2. Estimación no paramétrica de la tendencia espacial

Este capítulo se va a emplear para introducir el concepto de regresión, así como su utilidad y su aplicación práctica, pues serán temas recurrentes a lo largo de este trabajo. Para ello, en la sección 2.1. se presenta el modelo general de la función de regresión y se explican sus componentes. La sección 2.2. consiste en una introducción a la regresión no paramétrica, y se profundizará en la selección de la ventana, debido a su gran relevancia dentro de este tipo de métodos y también se estudiarán dos de los principales estimadores no paramétricos, el estimador de Nadaraya-Watson y el local lineal (centrándonos especialmente en el segundo). En la sección 2.3. se tratará el concepto del *binning*, que es el punto central de este trabajo. Por último, en la sección 2.4. se analizarán algunos de los paquetes empleados para la estimación de la tendencia y el variograma en el lenguaje informático R, para así decidir cuál es el adecuado para los estudios que se van a realizar en este trabajo.

2.1. El modelo general de regresión

Una de las cuestiones más comunes a la hora de analizar un conjunto de datos consiste en estimar si existe alguna relación entre una variable respuesta y una o varias variables explicativas, y son estas relaciones entre variables lo que comúnmente recibe el nombre de regresión. El análisis de la regresión es un método que permite modelizar esta relación, en el que se emplean como variables explicativas a $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$, y donde la variable respuesta, explicada o dependiente es Y . Además de las modelización de estas relaciones, otro de los principales objetivos que tiene la construcción de estos modelos es, a partir de ellos, realizar predicciones del valor de la variable respuesta Y cuando se conoce el valor de $\boldsymbol{X}$. En este caso, el modelo que se va a considerar es el modelo espacial general que se presentó en (1.1).

Por otra parte, los modelos de regresión pueden dividirse en dos categorías, en función de si se asume o no una forma predeterminada para $\mu(\boldsymbol{x})$: los modelos paramétricos y los no paramétricos. Los primeros de ellos sí que consideran una forma predeterminada para la función de regresión, mientras que en la regresión no paramétrica solo se asumen ciertas condiciones o hipótesis generales sobre $\mu(\boldsymbol{x})$, las cuales están relacionadas con la suavidad de la función (continuidad y diferenciabilidad), y van a ser este último tipo de métodos los que se empleen a lo largo del presente trabajo.

2.2. La regresión no paramétrica

El objetivo principal que se busca a través del análisis de la regresión, es la estimación de la función de regresión, concepto que ya se definió en la sección 2.1. Además, la modelización del conjunto de datos necesita ser adecuada, ya que en caso contrario las predicciones que se realicen a partir de él van a ser poco precisas. Esta estimación puede realizarse tanto a través de métodos paramétricos, que son aquellos en los que la forma funcional es completamente conocida una vez que se conoce el valor de todos los parámetros que deben de ser estimados (ver por ejemplo Härdle, 1990 o Eubank, 1999), como a través de métodos no paramétricos, que son aquellos que estima $\mu(\boldsymbol{x})$ sin considerar previamente ninguna forma específica, y va a ser en este segundo tipo de métodos en los que se centra este trabajo.

Los métodos no paramétricos tienen una serie de ventajas con respecto a los paramétricos, como por ejemplo, que son más flexibles y que se ahorra la necesidad de especificar una forma para μ. Por contra, esta metodología también tiene varias desventajas, como que los resultados que se obtienen por estos procedimientos tienen por lo general una interpretación más complicada o que en situaciones en las que se dispone de pocos datos, estos métodos pueden disponer de una potencia muy inferior a la de los enfoques tradicionales.

Por otro lado, existen múltiples métodos no paramétricos para la estimación de la tendencia, por lo que va a ser necesario escoger uno que sea adecuado para este trabajo. Uno generalmente empleado es el estimador de Nadaraya-Watson (NW, de aquí en adelante). El funcionamiento de este estimador consiste en que para un $\boldsymbol{x}$ determinado, el vecindario correspondiente sería $(x - h, x + h)$ y, la estimación de la función de regresión en este punto se calcularía como el promedio de los valores $\boldsymbol{Y}_i$ cuyo $\boldsymbol{X}_i$ correspondiente pertenece a este intervalo, en otras palabras, se obtiene por medio de un promedio ponderado local. De lo que se acaba de comentar, se puede deducir que el ancho de ventana h (que para el caso multivariante es una matriz), va a tener una gran relevancia a la hora de emplear este estimador, por tanto, resultará imprescindible su correcta selección. El estimador NW en el punto x se obtiene como la minimización de la expresión:

$$\min_{\boldsymbol{\beta}_0} \sum_{i=1}^{n} \{Y(\boldsymbol{x}_i) - \beta_0(\boldsymbol{x})\}^2 K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})$$

Pero este estimador tiene un gran inconveniente, no estima correctamente en los puntos frontera. En este caso, se considerará el estimador local lineal (LL, de aquí en adelante), y posteriormente se realizará una breve comparación de ambos para ver si en efecto mejora el comportamiento del estimador de NW.

2.2.1. Estimador local lineal

Este estimador no paramétrico no es más que una generalización del estimador de NW. Para este caso, se presenta la fórmula del estimador LL para la tendencia espacial, que es el tipo de tendencia que se va a estimar en este trabajo. Por tanto, si partimos de un modelo similar al que se presentó en (1.1), se podría obtener el estimador LL multivariante: $\hat{\mu}_{\boldsymbol{H}}(\boldsymbol{x}) = \hat{\beta}_0$, realizando la siguiente minimización:

$$\min_{\beta_0, \boldsymbol{\beta}_1} \sum_{i=1}^{n} \{Y(\boldsymbol{x}_i) - \beta_0 - \boldsymbol{\beta}_1^t(\boldsymbol{x}_i - \boldsymbol{x})\}^2 K_{\boldsymbol{H}}(\boldsymbol{x}_i - \boldsymbol{x})$$

Expresión en la que $Y(\boldsymbol{x}_i)$ representa a la componente i-ésima del vector $\boldsymbol{Y}$, que ha sido observado en las posiciones $\boldsymbol{\chi} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, de tal forma que $\boldsymbol{x}_i \in D \subset R^d$. Por su parte, $K_{\boldsymbol{H}}(u) = |\boldsymbol{H}|^{-1} K(\boldsymbol{H}^{-1}u)$, en la que K es una función tipo núcleo d-dimensional y, por último, $\boldsymbol{H}$ representa a la matriz de ventanas que tiene estructura $d \times d$, simétrica no singular.

En lo referente a las propiedades de este estimador, hay múltiples trabajos que las tratan, como por ejemplo Rupert y Wand (1994), donde se presentan las propiedades asintóticas para el caso del estimador LL con datos incorrelados bajo diseño aleatorio, mientras que en Francisco-Fernández y Vilar-Fernández (2001) se exponen las expresiones del sesgo y varianza para el caso univariante. En Opsomer et al. (2001) se estudiaron las propiedades asintóticas para el caso bidimensional con datos correlados. Posteriormente, Liu (2001) presentó las propiedades asintóticas que mayor interés van a tener para este trabajo, las del caso multidimensional con datos correlacionados. Bajo ciertas condiciones, que se encuentran disponibles en esta última referencia, se pueden obtener las expresiones del sesgo y varianza asintóticos en el caso multidimensional y con errores correlados. Previamente destacar que bajo la suposición de que $\rho_n(\cdot)$ es el correlograma, entonces: $\rho_I = \lim_{n\to\infty} n \int \rho_n(\boldsymbol{u}) d\boldsymbol{u}$. Una vez comentado esto, se muestran las expresiones asintóticas del sesgo y la varianza:

$$E[\hat{\mu}_{\boldsymbol{H}}(\boldsymbol{x}) - \mu(\boldsymbol{x})|\boldsymbol{\chi}] = \frac{1}{2} m_2(K) tr\left(\boldsymbol{H}^2 \boldsymbol{H}_\mu(\boldsymbol{x})\right) + o_p(tr(\boldsymbol{H}^2)) \tag{2.1}$$

$$Var[\hat{\mu}_{\boldsymbol{H}}(\boldsymbol{x})|\boldsymbol{\chi}] = \frac{m_2(K)\sigma^2(1 + f_x(\boldsymbol{x})\rho_I)}{n|\boldsymbol{H}| f_x(\boldsymbol{x})} + o_p\left(\frac{1}{n|\boldsymbol{H}|}\right) \tag{2.2}$$

En estas expresiones, $tr(\boldsymbol{A})$ se utiliza para referirse a la traza de la matriz $\boldsymbol{A}$, mientras que $\boldsymbol{H}_\mu(\boldsymbol{x})$ es la matriz Hessiana de $\mu(\cdot)$ cuando se evalúa en $\boldsymbol{x}$ y, por último, $m_2(K)\boldsymbol{I} = \int \boldsymbol{u}\boldsymbol{u}^t K(\boldsymbol{u}) d\boldsymbol{u}$ siendo $m_2(K) \neq 0$.

Por otro lado, en Seifert y Gasser (1996) se presenta el mayor inconveniente que tiene este estimador, que consiste en que cuando se emplea una función *kernel* de soporte compacto la varianza condicional para el caso de muestras finitas no está acotada, esto provocaría que, en muchos casos, la función de regresión estimada por medio del LL tendría una mala representación gráfica. Es necesario destacar que este problema no le ocurre al otro estimador presentado en este trabajo, el de NW (Nadaraya, 1964; Watson, 1964).

Si se realiza una comparación entre el estimador LL y el estimador de NW, el sesgo asintótico del LL es menor (aunque del mismo orden) pero ambos tienen la misma varianza asintótica (ver por ejemplo Fan y Gijbels, 1996). Y, como ya se comentó anteriormente, el estimador LL tiene un mejor comportamiento en la frontera, pero no es capaz de solucionar los problemas de dimensionalidad que ya se presentaban en el de NW. Este problema consiste en que cuando se aplica el estimador en casos multidimensionales, es necesario que el tamaño de las muestras sea mucho más grande que para el caso unidimensional debido al problema comúnmente conocido como "la maldición de la dimensionalidad". Este asunto, que ha sido tratado por una gran cantidad de autores (ver por ejemplo Kuo y Sloan, 2005), consiste en que, al aumentar el número de dimensiones consideradas, también se incrementa el volumen del espacio considerado provocando que los datos disponibles se dispersen y afectando a cualquier método que requiera de la significación estadística, necesitando un incremento exponencial de datos para obtener los mismos resultados a medida que se aumentan las dimensiones. En este trabajo, debido a sus características, el estimador que se va a utilizar es el estimador LL.

Basándonos en lo previamente comentado, puede deducirse como el ancho de ventanas va a ser uno de los factores más importantes para que los estimadores funcionen correctamente, por lo que va a ser necesario seleccionar un valor adecuado (o una matriz en el caso multidimensional). Por ello, se van a estudiar a continuación distintos criterios que se pueden emplear para determinar qué ancho de ventanas escoger.

2.2.2. La matriz de ventanas

Uno de los parámetros de mayor importancia a la hora de realizar las estimaciones es el ancho de la ventana $\boldsymbol{H}$, ya que va a tener una importante repercusión en el grado de precisión del estimador. Recordemos que si el ancho de ventana elegido es demasiado grande el estimador será infrasuavizado y si es demasiado pequeño será sobresuavizado. Para el caso multidimensional (en concreto en este trabajo va a ser bidimensional) la ventana se expresa como una matriz $d \times d$ simétrica y de la forma:

$$\boldsymbol{H} = \begin{pmatrix} h_{1,1} & & h_{1,d} \\ & \ddots & \\ h_{d,1} & & h_{d,d} \end{pmatrix}$$

Además, por simplicidad consideraremos únicamente matrices ventanas diagonales (muchos paquetes de R implementan este tipo de ventanas), en la que cada $h_{l,l}$ (valores diagonales de la matriz $\boldsymbol{H}$) es seleccionado en función de la variabilidad de $\boldsymbol{X}_l$. Para un caso más general, se podría considerar que la matriz $\boldsymbol{H}$ guarda relación con la matriz de covarianzas de las variables explicativas.

Para el caso en el que existe dependencia (ver por ejemplo Castillo, 2017), que es el que atañe al presente trabajo (de tipo espacial), una de las maneras de seleccionar la ventana sería a través de la minimización del error cuadrático medio (MSE, de aquí en adelante), que depende del sesgo y la varianza y que puede calcularse como:

$$MSE(\hat{\mu}(x), \boldsymbol{H}) = E[\hat{\mu}_{\boldsymbol{H}}(\boldsymbol{x}) - \mu(\boldsymbol{x})]^2 + Var[\hat{\mu}_{\boldsymbol{H}}(\boldsymbol{x})]^2$$

Aunque, para el caso de selección de la ventana podría ser más recomendable emplear su aproximación asintótica, denotado por $AMSE(\hat{\mu}(x), \boldsymbol{H})$, que permite obtener las expresiones de una manera más sencilla. El AMSE se obtendría como la suma del cuadrado del sesgo y la varianza asintóticos, es decir: (2.1) y (2.2). Por tanto, la ventana óptima local según este método será aquella que minimice el AMSE, y se obtendría a partir de la expresión:

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

siendo $m(K^2) = \int K^2(\boldsymbol{u})d\boldsymbol{u}$ y $\widetilde{\boldsymbol{H}}_{\mu}(\boldsymbol{x}) = \boldsymbol{H}_{\mu}(\boldsymbol{x})$, en el caso en el que $\boldsymbol{H}_{\mu}(\boldsymbol{x})$ es definida positiva. Cuando sea definida negativa se iguala a $-\boldsymbol{H}_{\mu}(\boldsymbol{x})$. Por tanto, el término $\left(\widetilde{\boldsymbol{H}}_{\mu}(\boldsymbol{x})\right)^{-1/2}$ sirve para establecer la forma y orientación del vecindario local que se emplea para la estimación en un punto dado de la tendencia mientras que el primer término es el que determina el tamaño de ventana.

El método que se acaba de presentar sirve para obtener la ventana óptima local, sin embargo, en muchas ocasiones nos va a interesar en lugar de emplear un óptimo local, hacer uso de un óptimo global. En esa situación, el criterio que se suele utilizar para obtener la ventana es el del Error Cuadrático Medio Integrado (MISE, de aquí en adelante), que tiene la siguiente expresión:

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

en la que $w(\cdot)$ es una función de pesos que tiene como función principal eliminar el efecto frontera. Destacar que para el caso de diseño aleatorio, es la función de densidad $f_x(\boldsymbol{x})$ la que se comúnmente se utiliza como función de pesos. Por otro lado, al igual que ocurría para el caso del MSE, podría resultar más interesante minimizar la aproximación asintótica del MISE, que es el AMISE, y cuya expresión es:

$$AMISE(\boldsymbol{H}) = \int AMISE(\hat{\mu}(\boldsymbol{x}, \boldsymbol{H}))f_x(\boldsymbol{x})d\boldsymbol{x}$$

el principal problema que tiene este criterio es que a día de hoy no hay disponible una expresión de la ventana obtenida por esta medida de error. Por ello, se propone obtener aquella ventana que minimice la aproximación del MISE, que comúnmente es conocida como MASE, y que consiste en minimizar el promedio de los errores cuadráticos ponderados de los datos observados, es decir:

$$MASE(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n} E\left[\left(\hat{\mu}_{\boldsymbol{H}}(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)\right)^2\right] w(\boldsymbol{x}_i)$$

Por otro lado, como el estimador lineal local de la tendencia es un suavizador lineal, podemos expresar las estimaciones en los puntos de observación como: $\widehat{\boldsymbol{\mu}} = \boldsymbol{SY}$, siendo $\boldsymbol{S}$ la matriz de suavizado cuya i-ésima fila viene dada por $s_{X_i}^t$, que es el vector de suavización para $\boldsymbol{x} = \boldsymbol{X}_i$, la ventana óptima se obtendría minimizando:

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

expresión en la que el segundo término hace referencia a la esperanza matemática de una forma cuadrática con respecto al vector $(\boldsymbol{SY} - \boldsymbol{\mu})$. Sería posible reescribir la expresión anterior como:

$$MASE(\boldsymbol{H}) = \frac{1}{n}(\boldsymbol{S\mu} - \boldsymbol{\mu})^t(\boldsymbol{S\mu} - \boldsymbol{\mu}) + \frac{1}{n}tr(\boldsymbol{S\Sigma S}^t)$$

Aunque en la práctica no es aplicable, puesto que $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ (la matriz de varianzas y covarianzas respectivamente) son desconocidas. Debido a esto, en la práctica lo métodos de regresión tipo núcleo suelen emplear el método de Validación Cruzada (CV, de aquí en adelante). Este procedimiento consiste en escoger aquella ventana o matriz de ventanas que minimice la expresión:

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

siendo $\hat{\mu}_{-i}$ el estimador de la tendencia obtenido al omitir el i-ésimo par $(\boldsymbol{x_i}, Y(\boldsymbol{x}_i))$ y s_{ii} el i-ésimo término de la diagonal de la matriz $\boldsymbol{S}$. Además, siendo la última igualdad de la expresión cierta debido a que la suma de los elementos de cada fila de $\boldsymbol{S}$ suman 1.

Otra manera de seleccionar la ventana consistiría en reemplazar el término s_{ii} de la expresión de CV por su promedio, esto matemáticamente se haría de la forma: $\sum_{i=1}^{n}\frac{s_{ii}}{n} = tr(\boldsymbol{S})/n$. Este método recibe el nombre de Validación Cruzada Generalizada (GCV, de aquí en adelante) y cuya función a minimizar para obtener la matriz de ventanas $\boldsymbol{H}$ óptima es:

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

sin embargo, y como ocurre para los métodos de validación cruzada en el caso de datos correlacionados, este criterio de selección no resulta adecuado debido a que esto afecta a su esperanza por el hecho, que ya se nombró anteriormente, de que las ventanas óptimas dependen de la matriz de varianzas y covarianzas de los datos. Aunque es necesario resaltar que esto no solo le ocurre a este criterio, sino que la mayor parte de los métodos de selección de $\boldsymbol{H}$ no funcionan adecuadamente cuando hay correlación espacial. Por tanto, y para evitar caer en este posible error, un criterio de selección que se podría emplear sería la corrección del GCV que propusieron Francisco-Fernández y Opsomer (2005):

$$GCV_c(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\boldsymbol{x}_i) - \hat{\mu}(\boldsymbol{x}_i)}{1 - \frac{1}{n}tr(\boldsymbol{SR})}\right)^2$$

siendo $\boldsymbol{R}$ la matriz de correlación de las observaciones, que en la práctica suele ser desconocida. Para solucionarlo, estos mismos autores sugieren asumir un modelo paramétrico de covarianza y estimar $\boldsymbol{R}$, de tal modo que la expresión para el cálculo de la matriz de ventana quedaría como:

$$GCV_{ce}(\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\boldsymbol{x}_i) - \hat{\mu}(\boldsymbol{x}_i)}{1 - \frac{1}{n}tr(\boldsymbol{S}\widehat{\boldsymbol{R}})}\right)^2$$

2.3. Introducción al *binning*

Esta metodología de agrupación de datos va a ser el punto central de estudio del presente trabajo, por lo que es necesario que tanto el concepto como su funcionamiento se entiendan correctamente. El *binning* es un procedimiento que como su propio nombre sugiere (*bin* es contendor en inglés) consiste en agrupar en grupos los valores que forman parte de un conjunto de datos. Este agrupamiento de algunos componentes del conjunto de datos puede tener como principal beneficio una reducción de los costes computacionales ya que, por ejemplo, para el cálculo de la función de regresión no paramétrica, se disminuiría el número de evaluaciones *kernel* necesarias (que son la parte que más tiempo consume de las implementaciones directas) y además resulta mucho más sencillo calcular distancias (lo que es de especial interés en geoestadística). En este trabajo, en concreto se va a estudiar cómo funciona en el caso de la estimación de la tendencia y en el de la estimación del variograma, conceptos que ya han sido presentados previamente. Además de la posible ganancia en términos computacionales, también resultará adecuado comprobar si las estimaciones que se obtienen empleando el conjunto con datos agrupados tienen la misma precisión que las que se consiguen con el conjunto original.

Dentro del concepto general del *binning*, existen varios tipos en función de cómo se agrupan los datos. Si nos basamos en el trabajo de Wand (1994) hay dos reglas que son las comúnmente utilizadas:

- *Binning* simple: una aproximación al funcionamiento de esta metodología *binning* se puede encontrar en Fan y Gijbels (1996). Se parte de un conjunto original que está formado por datos denotados por $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ y siendo los puntos de la rejilla de la forma $\left(\widetilde{\boldsymbol{x}}_j, \widetilde{\boldsymbol{Y}}_i\right)$ y siendo $\boldsymbol{I}_j$ un conjunto de índices de la forma $\boldsymbol{I}_j = \left\{i: \boldsymbol{X}_i \to \widetilde{\boldsymbol{x}}_j\right\}$, en el que se registran las observaciones $\boldsymbol{X}_i$ que se han desplazado al punto de cuadrícula $\widetilde{\boldsymbol{x}}_j$. Este tipo de *binning* consiste en que dato es movido al punto de cuadrícula más cercano, el conjunto de datos agrupados que se obtiene al realizar esto y que está compuesto por una cantidad de datos n, puede ser expresado como:

$$\left\{\left(\widetilde{\boldsymbol{x}}_j, \overline{\boldsymbol{Y}}_j, \boldsymbol{c}_j\right) : \text{ j} = 1, \dots, n_{grid}\right\}$$

Expresión en la que $\overline{\boldsymbol{Y}}_j = average\left\{\widetilde{\boldsymbol{Y}}_i : i \in \boldsymbol{I}_j\right\}$ que está formado por las medias de los *bins*, y

siendo $\boldsymbol{c}_j$ un recuento de las observaciones que hay dentro de cada uno de los *bins*.

- *Binning* lineal: este tipo de *binning* es simplemente un refinamiento del *binning* simple, que consiste en que, en lugar de que cada dato le asigne todo su peso al punto de rejilla más cercano (como hace el *binning* simple), cada dato $(\boldsymbol{X}_i, \boldsymbol{Y}_i)$ se "asigna" a los dos puntos de la rejilla más cercanos de forma proporcional a su distancia. La repartición de los pesos se puede expresar para el caso unidimensional (para el multidimensional se consideran pesos multiplicativos) como:

$$\boldsymbol{w}_{i,j} = \left(1 - \frac{|X_i - \tilde{x}_j|}{\Delta}\right) \qquad \text{con } i = 1, \dots, n \text{ y j} = 1, \dots, n_{grid}$$

Siendo $\Delta = \tilde{\boldsymbol{x}}_j - \tilde{\boldsymbol{x}}_{j-1}$, es decir, el ancho de cada uno de los *bins*. Por tanto, los pesos se asignarían de la forma: $\boldsymbol{w}_{i,j}$ es el peso asignado al centro de *bin* $\tilde{\boldsymbol{x}}_j$ y $\boldsymbol{w}_{i,j+1}$ se le da al centro de *bin* $\tilde{\boldsymbol{x}}_{j+1}$. La representación gráfica de esta metodología mostrada en la Figura 1 puede ayudar a comprender su funcionamiento.

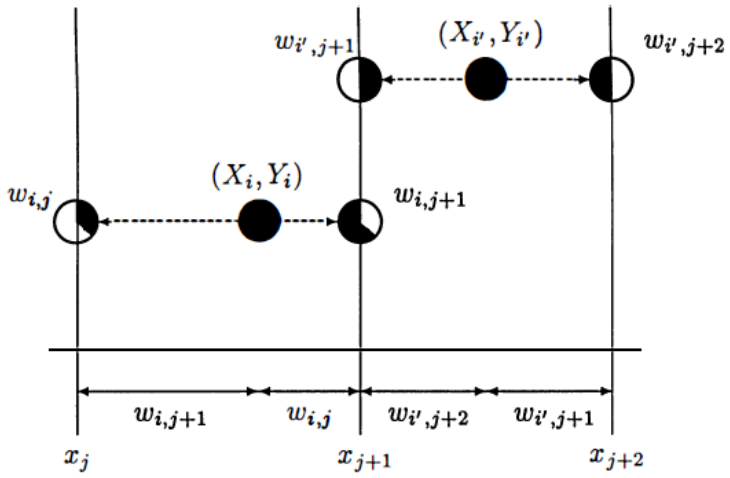

Figura 1. Representación del binning lineal. Fuente: Fan y Gijbels (1996).

Basándonos en todo lo anteriormente comentado sobre el *binning*, se puede obtener la expresión del estimador LL de la regresión cuando se emplea sobre un conjunto de datos *binned*, que es el que se va a emplear a lo largo de este trabajo, su expresión se muestra a continuación:

$$\min_{\beta_0,\beta} \sum_{j=1}^{n_{grid}} \left\{\tilde{Y}(\tilde{\boldsymbol{x}}_j) - \beta_0 - \beta^T(\tilde{\boldsymbol{x}}_j - \boldsymbol{x})\right\}^2 \boldsymbol{c}_j K_{\boldsymbol{H}}(\tilde{\boldsymbol{x}}_j - \boldsymbol{x})$$

Además, basándonos en estas mismas explicaciones, también se puede obtener la función del estimador el variograma cuando se realiza sobre un conjunto de datos agrupados. Esta expresión es una modificación de la presentada en la Sección 1.2., y sería:

$$\hat{\gamma}(\boldsymbol{u}) = \frac{1}{2\left|\sum_{i,j\in\widetilde{N}(\boldsymbol{u})} \boldsymbol{c}_i \boldsymbol{c}_j\right|} \sum\nolimits_{i,j\in\widetilde{N}(\boldsymbol{u}))} \left[\tilde{Y}(\widetilde{\boldsymbol{x}}_i) - \tilde{Y}(\widetilde{\boldsymbol{x}}_j)\right]^2 \boldsymbol{c}_i \boldsymbol{c}_j \tag{2.3}$$

siendo $\widetilde{N}(\boldsymbol{u}) = \{(i,j): \widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_j \in Tol(\boldsymbol{u})\}$. De forma que el estimador del variograma con *binning* es una media ponderada, en la que en lugar de dividir la suma de las semivarianzas por la cantidad de pares (como ocurría para el estimador sin *binning*), se divide por la suma de los pesos. Una vez explicado el *binning*, en los capítulos posteriores de este trabajo se estudiará tanto para el caso de datos reales (Capítulo 3) como para simulaciones (Capítulo 4), si esta metodología permite reducir los costes computacionales de las estimaciones de la tendencia y del variograma y también resultará de interés comprobar si son igualmente precisas las estimaciones obtenidas a partir de datos "crudos" y agrupados.

2.4. Estudio de paquetes de R

Para poder estudiar el *binning* como primer paso se buscaron funciones en R que lo implementen (pues es el lenguaje informático que se va a emplear para realizar las simulaciones) que permitan emplearlo tanto para la estimación de la tendencia como para la del variograma. Además, también resultaría de interés que ese mismo paquete fuera capaz de realizar los cálculos cuando se emplea el conjunto de datos sin modificar. Con este objetivo, a lo largo de esta sección se van a presentar algunos de los paquetes que permiten realizar estas estimaciones y se estudiará más profundamente aquellos que se utilizarán de manera activa a lo largo de este trabajo.

El primero de los paquetes a analizar va a ser *KernSmooth*, sobre este paquete lo más destacado es que permite realizar la estimación de la función de regresión a través de la función *locpoly()* únicamente para el caso con *binning*, aunque a pesar de eso este paquete no va a ser adecuado para este trabajo debido a que solo es capaz de estimar la función de regresión en el caso unidimensional, y en este trabajo se van a utilizar datos bidimensional. Destacar que sin embargo este paquete sí que es capaz de realizar estimaciones de la densidad para el caso bidimensional.

Por su parte, el paquete *ks* permite realizar *binning* lineal sobre conjuntos de datos tanto unidimensionales como multidimensionales hasta un máximo de 4 dimensiones mediante la función *binning()* aunque este paquete tiene como principal inconveniente que no permite estimar la tendencia, por lo que tampoco va a resultar adecuado para el presente trabajo. Otro paquete que también es conocido por el empleo del *binning* es el *npsp.*

Este paquete permite hacer *binning* lineal utilizando la función *binning()* sobre el conjunto de datos y también permite estimar la función de regresión mediante la función *locpol()* pero solo implementa la estimación con *binning*, ya sea introduciéndole los datos ya agrupados o un conjunto "crudo" que la propia función agrupa internamente. Este paquete además de permitir estimar la tendencia, también tiene funciones para estimar el variograma, la densidad, así como la selección de ventanas y la realización de predicción *kriging*. Destacar que este paquete tiene una característica especial, y es que el ajuste necesario para calcular el estimador se realiza en Fortran a través de la librería *lapack*, por lo que los tiempos computacionales que va a necesitar para llevar a cabo las estimaciones a priori van a ser inferiores y, por tanto, difícilmente comparables con los de otros paquetes que no fueron diseñados de la misma manera. Por todo esto, este paquete puede resultar adecuado en este trabajo para estudiar el caso con *binning* de manera individual, pues por lo que se acaba de explicar, comparar los resultados obtenidos con los que se obtendrían de estimar sin *binning* con otro paquete no reflejaría únicamente la ganancia en eficiencia por el empleo de *binning*, sino que también estaría reflejada la mejora por la implementación.

Debido a esto, sigue siendo necesario encontrar un paquete para el caso sin *binning* o idealmente que permita ambos, pues la comparación entre los dos métodos utilizados por un mismo paquete sería a priori lo más justo. Por ello, se presenta el paquete *sm*. Este paquete emplea métodos de suavizado no paramétricos basándose en Bowman y Azzalini (1997) y, además, los aspectos algorítmicos de este mismo son tratados en Bowman y Azzalini (2003). Por otro lado, la estimación de la regresión se lleva a cabo a través de la función *sm.regression()* que emplea *binning* de manera automática para conjuntos de datos grandes aunque este paquete también permite realizarla sin *binning*, como se mostrará más adelante.

Algunas de las propiedades que tiene este paquete son: el núcleo que emplea es Gaussiano, que es el que más comúnmente se utiliza debido a sus propiedades; las funciones de este paquete son aplicables tanto para una como para varias dimensiones (hasta un máximo de 3), se puede utilizar para conjuntos de datos de distintos tipos como datos de supervivencia, series de tiempo, recuentos y datos binomiales, además de que admiten datos faltantes. Por todo esto, este paquete va a ser el que principalmente se utilice en este trabajo. Para analizar esta función, va a mostrar a continuación con las principales opciones que se emplearon en este trabajo:

$$sm.regression(x, y, h\ , nbins, ngrid, eval.points, display, \dots)$$

- x: es la componente en la que se introducen las coordenadas espaciales de los datos, que deben ingresarse como una matriz de dos columnas.
- y: es un vector en el que se incluyen los valores obtenidos de la variable respuesta.
- h: permite seleccionar la matriz de ventanas que se va a emplear en la estimación de la tendencia. En caso de no ser especificada, se selecciona de manera predeterminada por la función. Destacar que solo admite ventanas diagonales

- $nbins$: es la cantidad de *bins* que se van a emplear al realizar *binning* sobre los datos. Este valor debe ser introducido como un vector, por ejemplo: $nbins = c(20,20)$, pues como estamos considerando un caso de estudio bidimensional, cada componente hace referencia a la cantidad de *bins* a emplear en una de las dimensiones. Además, si este parámetro se establece como 0, la función va a realizar la estimación de la función de regresión sin *binning*.
- $ngrid$: establece las dimensiones de la rejilla sobre la que se realizan las estimaciones de la tendencia. Este parámetro es necesario especificarlo, ya que esta rejilla tiene un valor predeterminado (100, 50 y 20 para una, dos y tres dimensiones respectivamente), lo que puede generar problemas si el número de *bins* seleccionados no es de exactamente ese valor. Destacar que, la cantidad establecida para el caso unidimensional se refiere al número de puntos sobre los que realizar las estimaciones mientras que para los casos de dos y tres dimensiones el valor determina el número de puntos a lo largo de cada eje en cada una de las dimensiones.
- $eval.points$: permite escoger en qué puntos se lleva a cabo la estimación de la tendencia. Para el caso bidimensional habría que pasar en esta componente una matriz de dos columnas.
- $display$: si no se especifica nada, la función de manera predeterminada realiza una representación gráfica de la tendencia espacial estimada. Para que no se realice esto, hecho que va a ser de gran importancia a la hora de valorar los tiempos computacionales (pues la representación gráfica supone un coste computacional extra e innecesario), es necesario establecer este factor como igual a "none".

Por todo esto, se escogió a esta función para realizar la estimación de la tendencia tanto para el caso en el que se emplea *binning* como en el que no. Pero todavía queda por seleccionar una función que se pueda utilizar para la estimación del variograma. Por ello, se presenta el paquete *geoR*.

Este paquete está orientado al análisis de datos geoestadísticos y dispone de la función *variog()* que permite estimar el variograma empírico a partir de una muestra de datos. Aunque destacar que para el caso con *binning* se utilizará una modificación de esta función, en la que se tienen en cuenta los pesos (expresión mostrada en 2.3.). Para explicar el funcionamiento de esta función, se va a seguir el mismo procedimiento que para *sm.regression()*, por lo que se muestra a continuación la función con sus principales argumentos:

$$variog(geodata, coords, data, uvec, option, estimator.type, max.dist, \dots)$$

- $geodata$: este parámetro se utiliza para introducir los datos sobre los que estimar el variograma. Para ello, habría que introducir una lista de clase "geodata" que estuviera compuesta por las coordenadas y los valores de los datos. Esta es una de las dos opciones de introducir los datos en la función, la otra consistiría en incluir en la función las

coordenadas y los valores por separado, como se muestra a continuación a través de $coords$ y $data$.

- $coords$: en este argumento se aportarían las coordenadas espaciales en las que se encuentran los datos. Para ello, habría que pasarle una matriz de dos columnas, en la que cada fila es la posición de una observación.

- $data$: se emplea para introducir un vector compuesto por los valores de los datos.

- $uvec$: en esta componente se introduce un vector en el que se incluyen los puntos en los que se van a estimar valores.

- $option$: permite escoger el tipo de salida de la función. Hay tres opciones: "*bin*", que es la que utiliza por defecto, y devuelve los valores de un variograma *binned*, "*cloud*" tiene como salida la nube del variograma, que es una gráfica en la que se exponen todas las semivarianzas, es decir, cada valor de $[Y(\boldsymbol{x}_i + \boldsymbol{u}) - Y(\boldsymbol{x}_i)]^2$. Por último, la opción "*smooth*" devuelve el variograma *kernel* suavizado.

- $estimator.type$: se puede elegir entre emplear el estimador clásico utilizando "*classical*" (es el que toma la función por defecto), o utilizar el estimador propuesto en Cressie (1993) si se pone "*modulus*".

- $max.dist$: en esta componente se puede determinar la distancia máxima que emplea el variograma, en otras palabras, para el cálculo del variograma se ignoran los pares de observaciones que se encuentren a una distancia superior a la marcada como límite. En caso de no ser fijado ningún valor, la función toma por defecto la mitad de la máxima distancia entre dos pares de observaciones, de entre todas las que componen el conjunto de datos aportado previamente.

Capíulo 3

3. Efecto del *binning* en la estimación de la tendencia de un conjunto de datos reales

En este capítulo se va a presentar un análisis sobre la influencia del *binning* en el tiempo computacional de la estimación de la tendencia para el caso de un conjunto de datos reales, haciendo uso del lenguaje informático R. En la sección 3.1. se presenta el conjunto de datos *precipitation*, que es el que se va a emplear en este capítulo. En la sección 3.2. se analizarán los tiempos medios obtenidos para los métodos con y sin *binning* bajo las mismas condiciones. La sección 3.3. consistirá en un estudio sobre cómo afecta la selección de la ventana a los costes computacionales y, por último, en la sección 3.4. se comprobará si variar el número de *bins* a la hora de realizar agrupamientos de datos afecta a los tiempos requeridos para estimar la tendencia, para lo que se empleará tanto la función del paquete *npsp* como la del *sm*.

3.1. Estudio del conjunto de datos

Para el análisis del comportamiento del *binning* que se va a realizar en el presente capítulo, se va a optar por emplear el conjunto de datos espaciales *precipitation*, que se encuentra disponible en el paquete *npsp* y se obtienen en R simplemente llamando a la librería a través de la función *library(npsp)*. Este conjunto de datos espaciales está formado por 1053 observaciones y por 6 variables. La variable que se mide, es decir, la variable explicada de este conjunto de datos aparece en el conjunto de datos con el propio nombre de *y*, y lo que representa es la cantidad de precipitaciones totales (en raíz cuadrada de la precipitación en pulgadas) que se registraron durante el mes de marzo del año 2016 en 1053 (número de filas) ubicaciones de la parte continental de los Estados Unidos.

Al ser un conjunto de datos espaciales, en él se proporcionan las coordenadas, en el paquete aparecen como *coordinates*, en las que se produjo cada una de las observaciones y que en R se muestran como una matriz de dos columnas, siendo llamadas *x1* y *x2* representando la longitud y la latitud respectivamente. Otra variable que se muestra es *WBAN* que representa al número de identificación de 5 dígitos de la estación meteorológica en la que se produjo cada observación. La última de las columnas de este conjunto de datos es *state*, que es una variable categórica que en este caso es *USA* para todas las observaciones. Lo importante de esta última variable no es solo ese nivel que muestra, sino que además presenta una serie de atributos: *labels*, *border* e *interior*. El primero de ellos es una lista formada por los datos y las etiquetas de las variables, el segundo

contiene polígonos espaciales que establecen los límites de la parte continental de los Estados Unidos y el tercero contiene polígonos espaciales que determinan los límites de cada uno de los estados (que pueden resultar muy útiles a nivel gráfico).

Una vez presentado el conjunto de datos, puede resultar de utilidad mostrar una representación gráfica de sus valores, esto se muestra en la Figura 2.

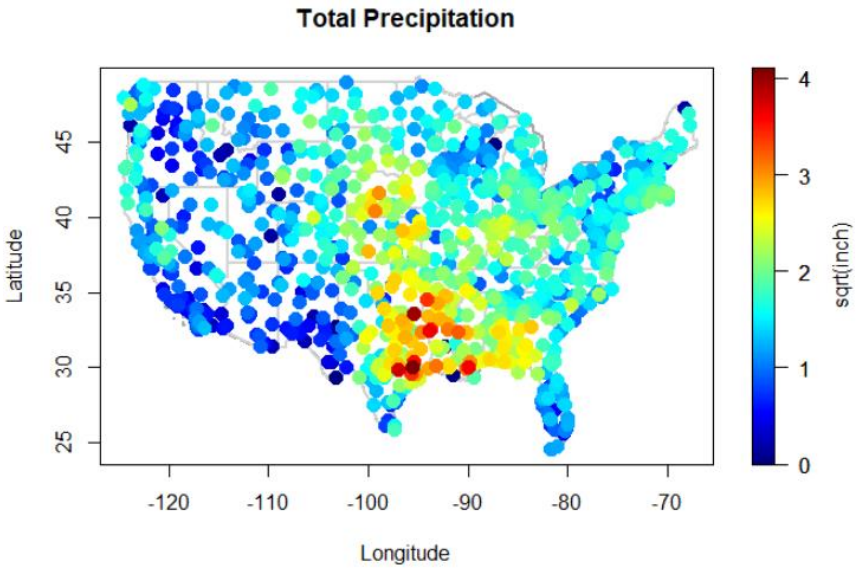

Figura 2. Representación de observaciones del conjunto de datos precipitation.

En esta representación puede comprobarse como en efecto aparecen los límites tanto de la parte continental de los Estados Unidos como la de los estados que la forman. Además, si se observa a la derecha del gráfico, se muestra una escala en la que se exponen los colores que toman los puntos según la cantidad de cantidad de precipitaciones totales registradas durante marzo de 2016. Basándonos en esto, puede verse como la zona occidental es en la que menor cantidad se han producido, mientras que es en el sur del país donde se han medido unas mayores cantidades, en concreto en los estados de Arkansas, Texas y Luisiana. Además, en la representación gráfica se ven indicios de dependencia espacial entre los datos, puesto que los puntos que se encuentran geográficamente cercanos tienen un color similar.

Una vez presentado el conjunto de datos que se va utilizar a lo largo del presente capítulo, puede comenzar a realizarse el estudio del *binning* para la estimación tanto de la tendencia como del variograma.

3.2. Comparación de los tiempos computacionales

Dentro del estudio del *binning*, uno de los principales temas de interés es el ahorro en tiempos computacionales que ofrece con respecto a emplear los datos “en crudo”. Para analizar el

comportamiento de este procedimiento, en esta sección se van a evaluar los tiempos computacionales medios requeridos para estimar la tendencia de un conjunto de datos existente utilizando el estimador LL. Previamente a presentar el estudio, se presentan las circunstancias bajo las que se lleva a cabo. Uno de los más importantes, como se comentó en la sección 2.2.2. de este trabajo, es la matriz de ventanas $\boldsymbol{H}$. En este caso, se optó por seleccionar la que toma por defecto la función *sm.regression()* del paquete *sm* (redondeada):

$$\boldsymbol{H} = \begin{pmatrix} 9 & 0 \\ 0 & 3 \end{pmatrix}$$

El siguiente parámetro a considerar es el número de *bins*, es decir, la cantidad de grupos en los que se van a agrupar los datos. En este caso, se consideran $(20{,}20)$, es decir 400 *bins*, pues los valores entre paréntesis representan la cantidad de puntos que conforman el *grid* que se toman sobre el eje de cada dimensión (bidimensional en este caso).

Por otro lado, para llevar a cabo las estimaciones de la tendencia, se emplean dos paquetes, el *sm* y el *npsp*. El primero de ellos, dispone de la función *sm.regression()* que permite realizar esta estimación tanto para el caso con *binning* como para el que no lo emplea. El siguiente paquete que se va a emplear va a ser el *npsp*, con la función *locpol()* para realizar la estimación para el caso de datos agrupados. De este modo, el estudio se realizará para tres estimaciones distintas, dos con *binning* y una sin él, para así poder analizar tanto si agrupar los datos permite reducir los costes computacionales, como si ambos paquetes funcionan de la misma manera en términos de computación. Para ello, se presenta la Tabla 1, que muestra los costes computacionales medios (en segundos) de 100 repeticiones de las estimaciones llevadas a cabo por cada uno de los procedimientos.

	Sin Binning (sm)	*Con Binning (npsp)*	*Con Binning (sm)*
Tiempo (seg)	5.3228	0.0016	0.1109

Tabla 1. Comparativa tiempos computacionales (en seg.) de las estimaciones con binning y sin binning.

Como se puede apreciar en la tabla, los tiempos computacionales medios requeridos para estimar la tendencia del conjunto de datos *precipitation* de los métodos que emplean *binning* son mucho más pequeños que para el caso que no lo emplea. Por otro lado, también puede resultar adecuado comparar los costes computacionales requeridos para los dos métodos que emplearon *binning*. En este caso, ambos procedimientos requirieron menos de un segundo para llevar a cabo la estimación, pero fue el paquete *npsp* el más rápido realizó los cálculos. Este mejor funcionamiento en términos de computación del paquete *npsp* era de esperar, pues como se

comentó en el capítulo 2.4., está diseñado para, empleando la librería *lapack* de Fortran, llevar a cabo el ajuste necesario para el cálculo del estimador, dándole de este modo una mayor eficiencia. Por último, destacar que los resultados presentados en esta sección parecen indicar que la agrupación de los datos va a permitir acelerar los cálculos, pero es necesario corroborar este hecho. Esto puede realizarse, por ejemplo, llevando a cabo el estudio de los procedimientos bajo distintas circunstancias.

3.3. Variación tiempos computacionales según la matriz de ventanas

Una vez presentados los costes computacionales de cada método para unos valores concretos de parámetros, se puede empezar a estudiar cómo afecta la modificación de los distintos parámetros a la hora de estimar la tendencia. En primer lugar, se va a analizar para el caso en el que varía la matriz de ventanas $\boldsymbol{H}$. Para ello, se va a construir el conjunto de ventanas a analizar. En este caso, se va a tomar que los dos valores de la diagonal son iguales, y estos valores van a ser generados a partir de una secuencia desde 5 hasta 200 tomando valores de 5 en 5. Además, a esta secuencia se le van a añadir a mayores algunas ventanas más pequeñas $(0.05, 0.5 \text{ y } 1)$, por si acaso en ventanas de tamaño muy pequeño los tiempos de computación se elevasen. Por tanto, la cantidad total de anchos de ventana analizados es de 43. En cuanto al número de *bins* empleados, se mantienen los que fueron fijados para la comparación de los tiempos de computación, es decir: $(20{,}20)$. Destacar que estos estudios se van a realizar únicamente para las estimaciones con y sin *binning* que se obtienen a través del paquete *sm*, pues como ya se comentó en el apartado anterior, la función del paquete *npsp* al haber sido creada ayudándose de Fortran hace que la comparación no sea justa. Una vez fijados los valores que va a tomar $\boldsymbol{H}$ se pueden llevar a cabo las estimaciones y a partir de ellas se obtienen las representaciones gráficas de los tiempos de computación (en segundos) según el tamaño de ventana, que se muestran en la Figura 3.

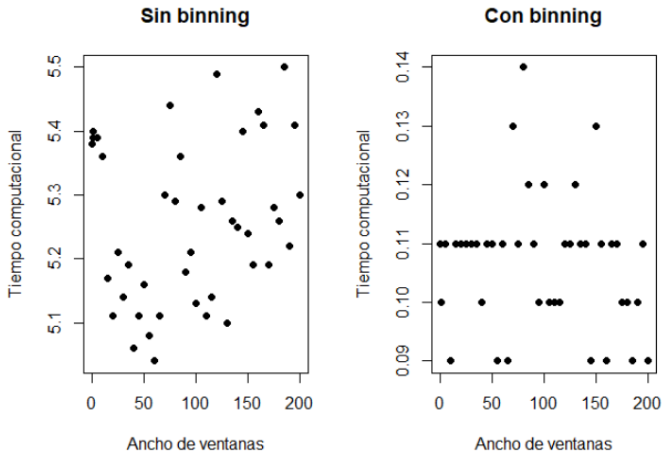

Figura 3. Tiempos computacionales (en seg.) para distintas matrices de ventanas.

Observando los gráficos, puede verse como el tiempo que se requiere para estimar la tendencia tanto con ambos métodos, no parecen verse muy afectados por el tamaño de ventana empleado, sino que más bien da la impresión de que varía de una manera aleatoria. Además, destaca que para el caso sin *binning*, la diferencia entre la estimación que tardó más tiempo en llevarse a cabo y la que menos tiempo necesitó es muy pequeña, en concreto inferior a un segundo. Para el caso sin *binning*, esta diferencia es todavía menor, de apenas cinco centésimas de segundo. Por último, destacar que los resultados obtenidos en este estudio presentan las mismas diferencias en términos de tiempos computacionales que las obtenidas en el apartado anterior, donde se estudió su comportamiento en función de si se emplea o no *binning* sobre el conjunto de datos, antes de hacer operaciones sobre él.

3.4. Variación tiempos computacionales según número de *bins*

Otro de los parámetros que puede provocar que varíe el tiempo computacional necesario para estimar la tendencia es la cantidad de *bins* considerados. Obviamente este parámetro solo afecta al método con *binning* y en este caso se van a estudiar los tiempos computacionales para valores de *bins* entre $(10,10)$ y $(250,250)$, es decir, el rango considerado es desde 100 *bins* hasta 62500, y que se crean a través de una secuencia que toma valores de 5 en 5. Además, se vuelve a tomar de nuevo a

$$\boldsymbol{H} = \begin{pmatrix} 9 & 0 \\ 0 & 3 \end{pmatrix}$$

como la matriz de ventanas que se va a emplear para cada una de las estimaciones de la función de regresión. Una vez fijados los parámetros que van a ser empleados para este estudio se puede comenzar el análisis. En este caso se va a estudiar cómo afecta al tiempo computacional tanto para la función del paquete *npsp* como para la función del paquete *sm*. Esto es debido a que, aunque ya se comentó anteriormente que de la manera en la que estaba diseñado el paquete *npsp* tenía ventaja computacional a la hora de llevar a cabo las estimaciones, como en este apartado lo que interesa es analizar de qué manera afecta al tiempo computacional en cada caso en vez de compararlos, pues resulta adecuado ver si en ambos casos la variación de la cantidad de *bins* tiene algún efecto, y si lo tiene, si es el mismo para ambos. Los resultados obtenidos se muestran a continuación en la Figura 4.

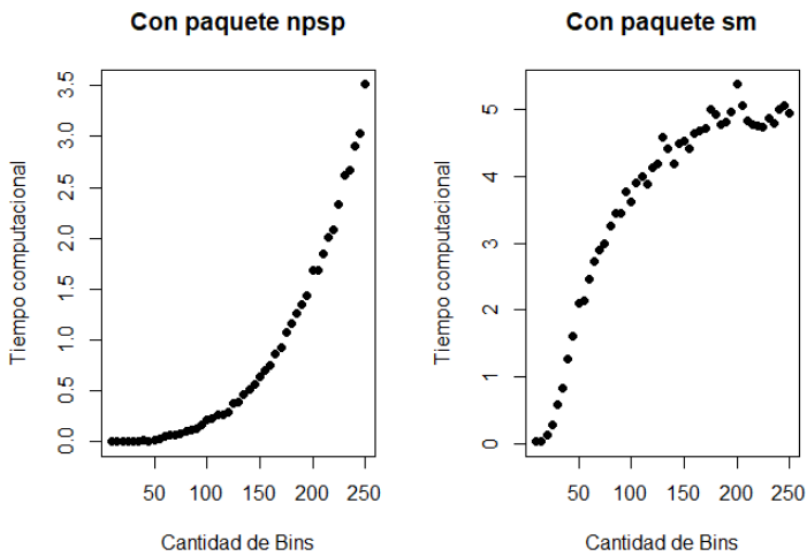

Figura 4. Tiempos computacionales (en seg.) para distintas cantidades de bins.

Observando el gráfico se puede ver claramente como el tiempo computacional requerido va aumentando a un ritmo muy elevado a medida que aumenta el número de *bins* considerados para ambos paquetes. En el gráfico que representa los tiempos del paquete *npsp* el aumento de los costes computacionales parece actuar de una manera exponencial, siendo bajo para pocos *bins* y muy elevado cuando la cantidad de *bins* considerada es grande. Por otro lado, en el caso del *sm* el incremento de los tiempos computacionales es diferente al del paquete *npsp*, en este caso el mayor crecimiento se produce al comienzo, cuando hay pocos *bins* y este crecimiento se va reduciendo a medida que se consideran más y más *bins*. La evolución obtenida de los costes computacionales de la estimación en función del número de *bins* es fácilmente comprensible, pues a mayor cantidad de ellos mayor esfuerzo debe realizar el estimador a la hora de agrupar los valores y, por tanto, mayor es el tiempo requerido para estimar la función de regresión. A diferencia de lo que se obtuvo para los anchos de ventana, la cantidad de *bins* empleados sí que va a tener una influencia importante con respecto al tiempo que tarda R en estimar la tendencia.

Capítulo 4

4. Efecto del *binning* en la estimación de la tendencia espacial con datos simulados

En esta parte de trabajo se va a presentar un estudio del *binning* sobre un conjunto de puntos irregularmente espaciados, que fueron generados de manera aleatoria en R. En primer lugar, en la sección 4.1. se va a presentar la tendencia teórica, que es, la que se empleará en estas simulaciones y la selección del ancho de ventana para llevar a cabo su estimación. En la sección 4.2. se analizarán distintas medidas de error de las estimaciones de la tendencia. En concreto, se centrarán en los errores medios y los errores cuadráticos tanto para el caso que emplea *binning* como el que no. Además, en la 4.3. se estudiarán los tiempos computacionales de ambos procedimientos para comprobar si los resultados obtenidos en el capítulo anterior sobre un conjunto de datos ya existente, también se dan para el caso de datos generados aleatoriamente. En 4.4. se analizará la relación existente tanto en términos de costes computacionales como en precisión de las estimaciones entre el tamaño de un conjunto de datos y el número de *bins* que se escogen para realizar *binning* sobre él. Por último, la sección 4.6. se va a centrar en el estudio de la influencia que tiene la dependencia espacial sobre la calidad de las estimaciones, comparando para ello los procedimientos con y sin *binning*. Por último, comentar que el estudio completo de la tendencia en este capítulo se va a realizar con la función *sm.regression()* del paquete *sm*. Mientras que para el análisis del variograma se utilizará la función *variog()* del paquete *geoR*. Por último, destacar que ambas utilizan al estimador LL.

4.1. El modelo teórico

Una de las principales características de interés cuando se realiza estimación es que los valores obtenidos se ajusten de la mejor manera posible a los valores teóricos, pero por lo general suelen ser desconocidos. En este caso, al ser un estudio de simulaciones, la tendencia teórica es conocida. En este caso, se optó por la siguiente expresión:

$$\mu(\boldsymbol{x}) = sin(\pi \boldsymbol{x}_1) + 4(\boldsymbol{x}_2 - 0.5)^2$$

Una vez presentada la expresión de la tendencia teórica, puede resultar de interés su representación gráfica. Las observaciones se generaron con distribución uniforme en el cuadrado unidad y las estimaciones se calcularon en una rejilla regular. En la Figura 5 se muestra la tendencia

teórica.

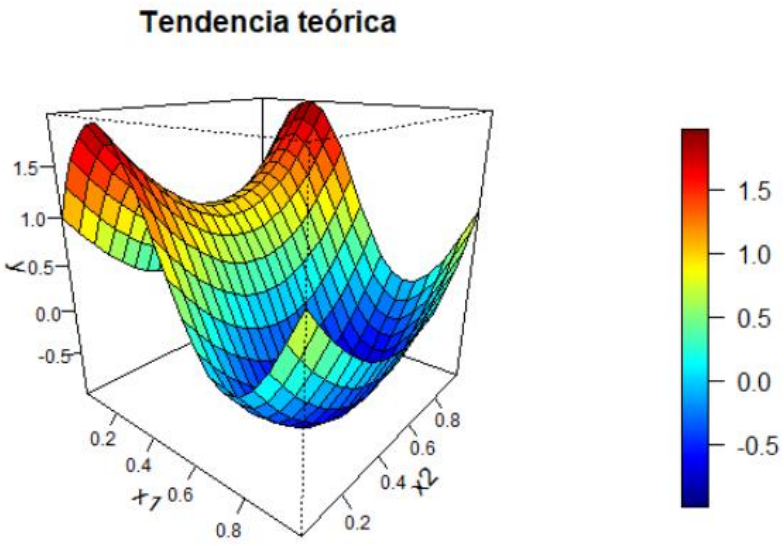

Figura 5. La tendencia teórica.

Estos valores teóricos son los que se van a emplear como referencia la hora de medir la precisión de estimación de los distintos procedimientos planteados. Existen múltiples formas de medir el grado de error de las estimaciones, pero este trabajo se va a basar en dos: el error medio de estimación y el error cuadrático medio.

Además, como se mostró en la sección 1.1., el modelo tiene otra componente: la dependencia espacial. En este caso, se considera un modelo exponencial (que es un caso particular del modelo de Matern con parámetro de suavizado 0.5). Considerando que $\boldsymbol{u} \neq 0$, la expresión del modelo es:

$$\gamma(\boldsymbol{u}|\boldsymbol{\theta}) = c_0 + c_1\left(1 - exp\left(-\frac{3\|\boldsymbol{u}\|}{a}\right)\right) \tag{4.1}$$

en la que $c_0 = 0.2$ denota al efecto *nugget*, $c_1 = 1$ al umbral parcial, $a = 0.6$ es el rango práctico (concepto definido en 1.2.). Destacar que para cada combinación de parámetros se generaron 1000 simulaciones.

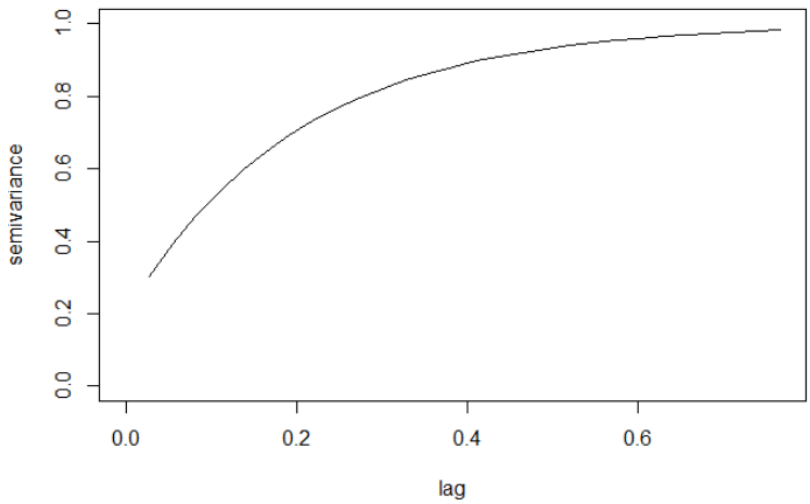

Figura 6. El variograma teórico.

En lo referente a la matriz de ventanas, las herramientas del paquete dificultan la obtención de la ventana MASE. Por ello, la opción por la que se va a optar va a consistir en considerar tres matrices de ventanas distintas, representar sus estimaciones de la tendencia media, a partir de 1000 simulaciones, y escoger a la que mejor actúe gráficamente, es decir, aquella que ni infrasuavice ni sobresuavice. Las matrices de ventas H consideradas son:

$$\begin{pmatrix} 0.05 & 0 \\ 0 & 0.0.5 \end{pmatrix} \qquad \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix} \qquad \begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}$$

Una vez presentadas las opciones se muestran a continuación las representaciones gráficas de la media de las tendencias estimadas para un conjunto de 200 datos y empleando cada una de estas matrices tanto para el caso con *binning*, utilizando $(20{,}20)$ *bins*, como para el que no lo emplea.

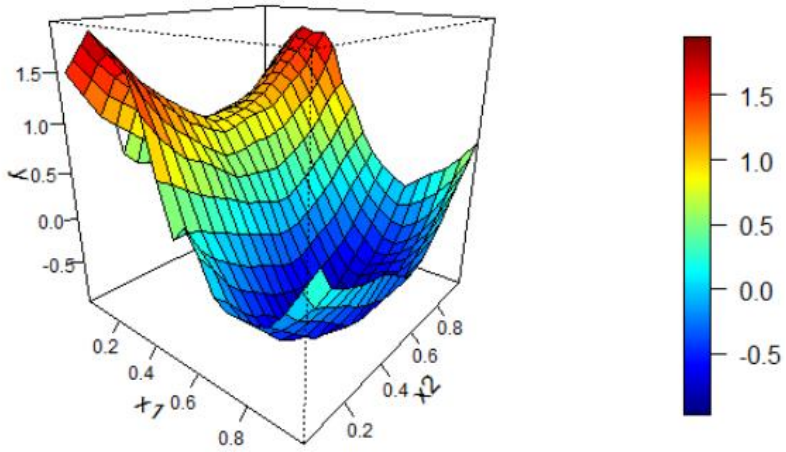

Figura 7. Tendencia estimada con datos binned para ventanas 0.05.

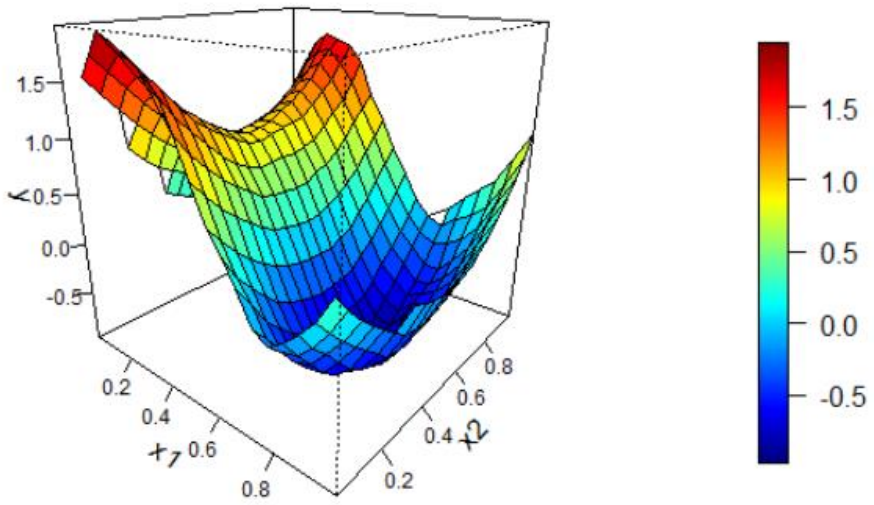

Figura 8. Tendencia estimada con datos binned para ventanas 0.05.

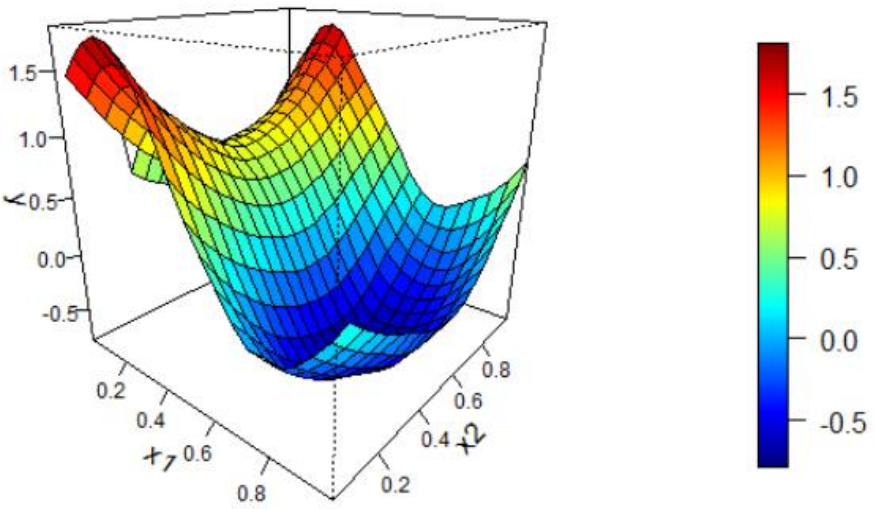

Figura 9. Tendencia estimada con datos binned para ventanas 0.1.

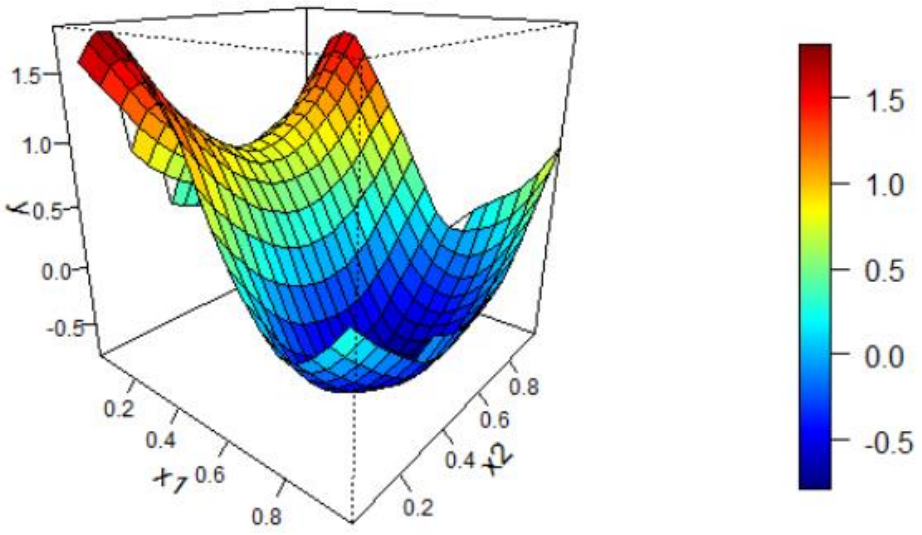

Figura 10. Tendencia estimada con datos originales para ventanas 0.1.

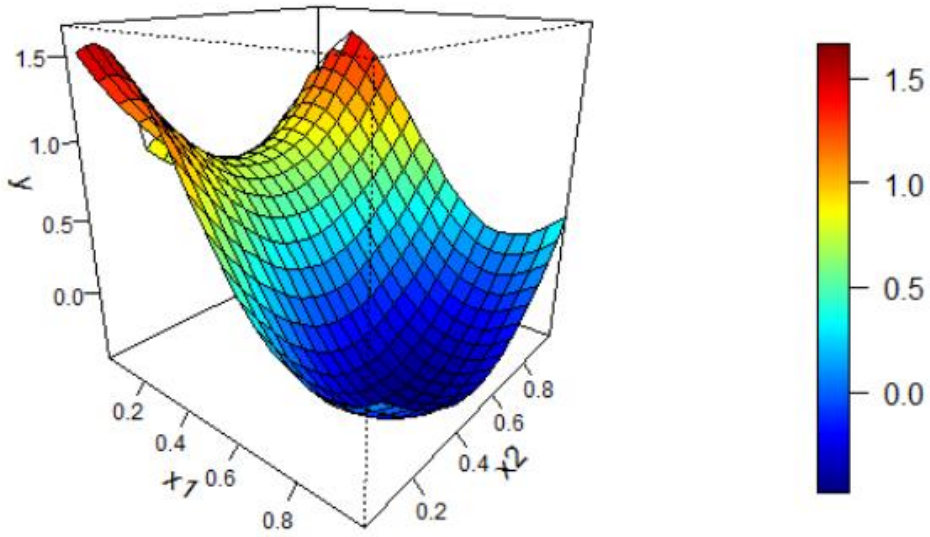

Figura 11. Tendencia estimada con datos binned para ventanas 0.2.

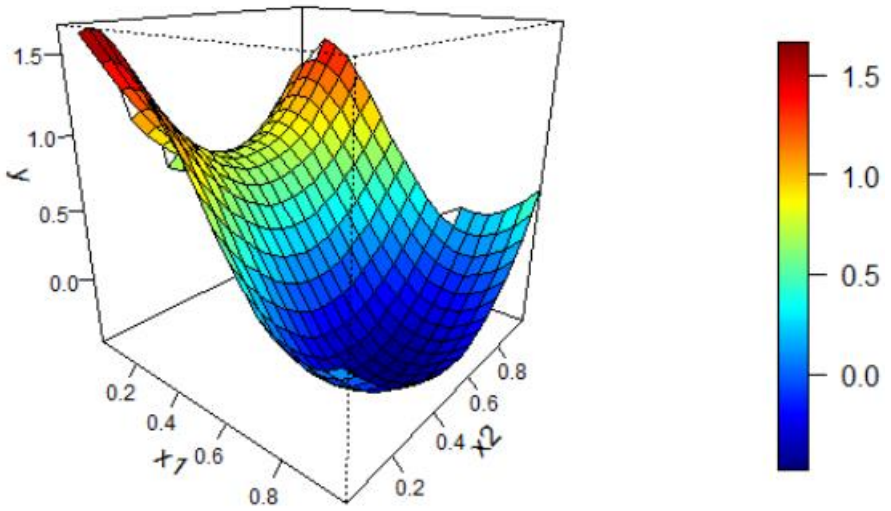

Figura 12. Tendencia estimada con datos originales para ventanas 0.2.

La primera matriz de ventanas considerada, es decir, la que tiene como elementos diagonales 0.05, obtuvo una estimación demasiado infrasuavizada en ambos casos, lo que sugiere que puede ser conveniente emplear un ancho de ventana mayor. Basándonos en esto, se analizan las representaciones obtenidas por la matriz con 0.2 en su diagonal. En este caso, la función de regresión estimada está claramente sobresuavizada, por lo que sería adecuado escoger un ancho de ventana menor. Por ello, se analizan las estimaciones de la tendencia obtenidas empleando la matriz de ventanas con elementos diagonales 0.1. En este caso, las representaciones gráficas muestran que la tendencia estimada empleando este ancho de banda parecen ser adecuadas (en especial para el caso que empleó *binning*), ni sobresuaviza ni infrasuaviza, por lo que va a ser esta la matriz de ventanas que se utilice en este estudio de simulación.

4.2. Medidas de error de las estimaciones

4.2.1. Error medio de las estimaciones

Esta sección del trabajo se centra en medir el error de las estimaciones de la tendencia que cometen por término medio tanto en el caso de emplear *binning* como cuando se utilizan los datos "crudos", bajo las condiciones presentadas en la sección anterior, es decir, utilizando un conjunto de 200 datos, con $(20{,}20)$ *bins* en el caso con *binning* y con la matriz de ventanas que tiene como elementos diagonales 0.1 (estas condiciones también se emplearán en el apartado 4.2.2.). El valor teórico se aproxima mediante simulación promediando las diferencias entre el valor real y la estimación:

$$ME(\boldsymbol{x}) = E\big(\mu(\boldsymbol{x}) - \hat{\mu}(\boldsymbol{x})\big)$$

Siendo $\hat{\mu}_j(\boldsymbol{x})$ la estimación de la simulación j-ésima, entonces:

$$\widehat{ME}(\boldsymbol{x}) = \frac{1}{nsim} \sum_{j=1}^{nsim} \left(\mu(\boldsymbol{x}) - \hat{\mu}_j(\boldsymbol{x})\right)$$

Expresión en la que se tomaron $nsim = 1000$. En las Figuras 13 y 14 pueden verse representados gráficamente los resultados obtenidos.

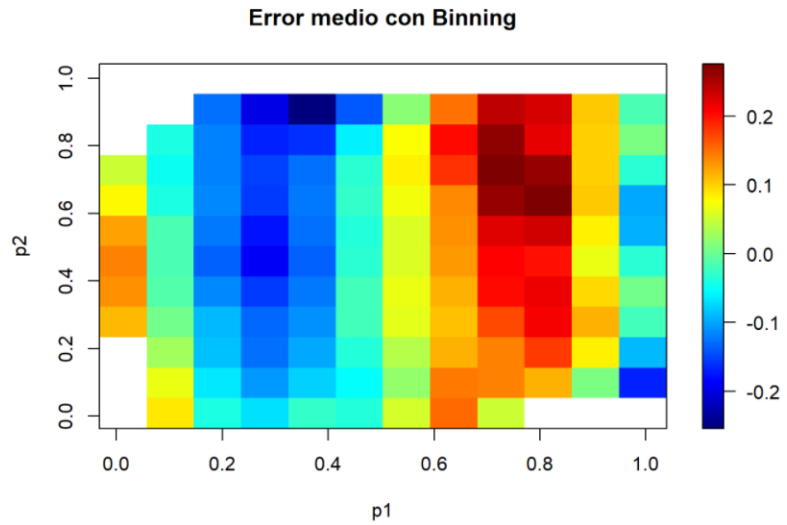

Figura 13. Error medio de estimación de la tendencia cuando se emplea binning.

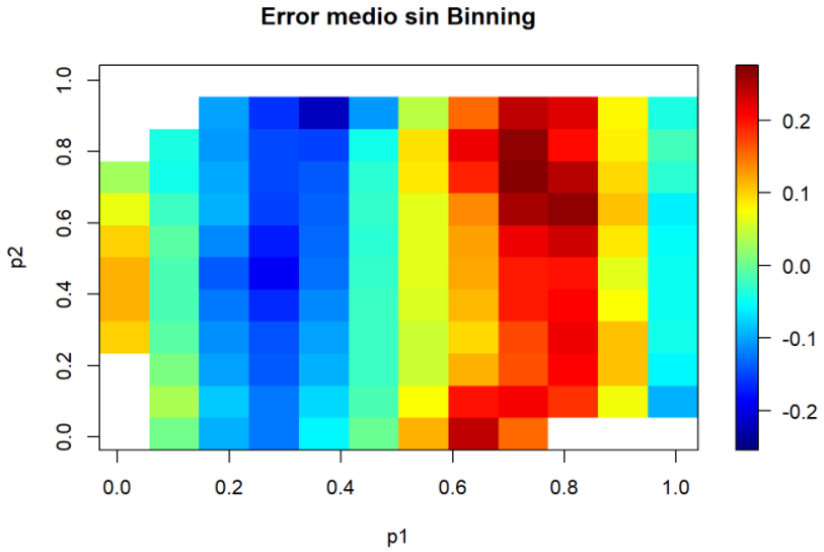

Figura 14. Error medio de estimación de la tendencia cuando no se emplea binning.

Ambos métodos dan la impresión de que actúan de una manera muy parecida, aunque parece que el caso con *binning* es propenso a alejarse ligeramente más del valor teórico de la tendencia en las coordenadas extremas que el caso que no emplea *binning*, mientras que en los puntos centrales tiene un mejor comportamiento que el caso con *binning*.

Una vez analizados de manera visual, se pueden analizar los valores numéricos del error medio. En la Tabla 2 se presentan la media, mediana y desviación típica de estos errores.

	Con Binning	*Sin Binning*
Media	0.0229366	0.0263075
Mediana	0.0199018	0.0230314
Desviación típica	0.6895349	0.6895269

Tabla 2. Comparativa de los errores de estimación de la tendencia.

Como se puede observar en la tabla, los errores medios de ambos métodos son muy pequeños, siendo ligeramente inferior para el caso que emplea *binning*, y teniendo prácticamente la misma variabilidad. Debido a esto, se puede considerar que según esta medida ambos procedimientos tienen un comportamiento muy similar a la hora de estimar la tendencia espacial. Por otro lado, esta medida de error tiene un gran inconveniente, el signo de los errores tiene una gran importancia, ya que podrían darse situaciones en las que, teniendo las estimaciones individuales un error asociado muy elevado, que el error medio fuese muy bajo, lo que sería debido a que errores de símbolo opuesto se anulen. Para prevenir esta posible incorrecta interpretación de la calidad de las estimaciones, resulta de gran interés el empleo de otra medida de error que no sufra este problema, en este trabajo se opta por estudiar los errores cuadráticos de estimación que, al considerar el valor de los errores al cuadrado, el símbolo que tengan esos errores carece de relevancia.

4.2.2. Error cuadrático medio de las estimaciones

Otra medida del error de estimación que se puede analizar es el MSE, que se obtiene promediando las diferencias al cuadrado entre el estimador y el valor teórico, lo que se realizaría de la siguiente forma:

$$MSE(\boldsymbol{x}) = E\big(\mu(\boldsymbol{x}) - \hat{\mu}(\boldsymbol{x})\big)^2$$

que se aproxima mediante simulación. Una vez explicado cómo se va a realizar el cálculo del MSE, se muestran representaciones gráficas del MSE que se obtuvieron tanto para el caso en el que al conjunto de datos se le aplicó *binning*, como al que empleó el conjunto de datos original.

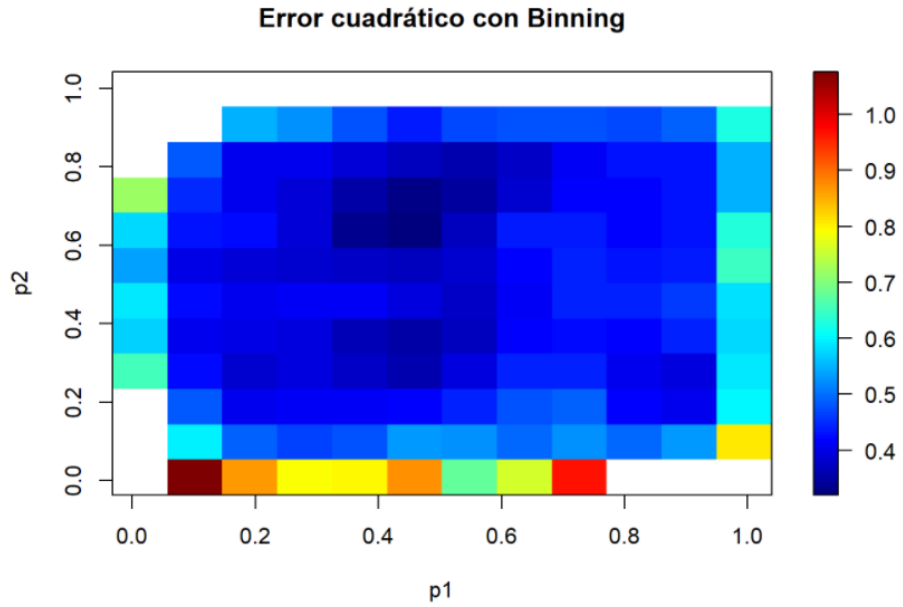

Figura 15. Error cuadrático medio de estimación de la tendencia cuando se emplea binning.

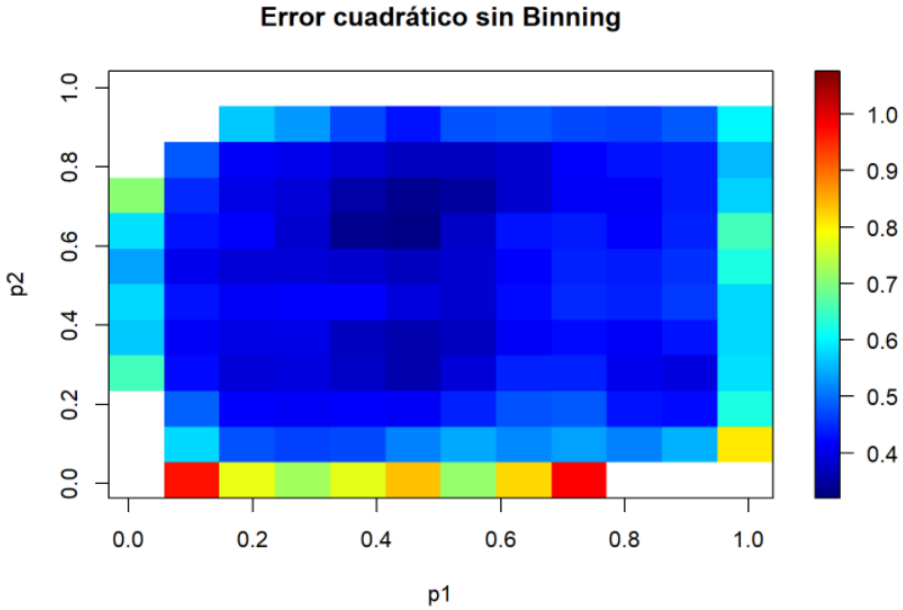

Figura 16. Error cuadrático medio de estimación de la tendencia cuando no se emplea binning.

Según estos gráficos, ambos procedimientos parecen estimar de una manera muy parecida. Esta valoración visual, puede verse reforzada mediante el análisis de los resultados numéricos, que se muestran en la Tabla 3.

	Con Binning	*Sin Binning*
Media	0.4759807	0.4761356
Mediana	0.2093701	0.2093685
Desviación típica	0.7095927	0.7052075

Tabla 3. Comparativa de los errores cuadráticos de estimación de la tendencia.

Los valores que se muestran en la tabla ratifican el análisis realizado anteriormente sobre las representaciones gráficas, ambos métodos cometen MSE muy similares, siendo ligeramente inferior el que emplea *binning*, pero con una mayor variabilidad. Sin embargo, es necesario destacar que las diferencias entre los MSE de ambos son muy pequeñas, en este caso de en torno a 0.0001, por lo que puede afirmarse que en este caso ambos métodos se comportan de una manera muy similar. Por último, destacar que los resultados MSE obtenidos muestran que en el estudio de los errores medios realizado en 4.2. no se dio la posible situación previamente planteada en la que los errores de signo negativo y positivo se anulan, dando lugar a un error medio muy pequeño, lo que hubiera provocado una interpretación errónea. Sin embargo, debido a la posible aparición de este problema, la medida de error que se empleará para realizar los estudios de precisión de las estimaciones a lo largo de este trabajo será el MSE.

4.3. Estudio de los tiempos computacionales al implementar *binning*

Esta sección se va a centrar en un estudio comparativo de los tiempos computacionales requeridos para la estimación de la tendencia entre el método que estima a partir de datos agrupados y el que utiliza el conjunto de datos "crudo" para distintos números de datos, mediante el empleo de simulaciones. Antes de comenzar con este estudio, resulta adecuado presentar las condiciones bajo las que se lleva a cabo. En este caso, el número de *bins* considerados es de $(20,20)$, es decir, 400. Mientras que la matriz de ventanas $\mathbf{H}$ empleada es la seleccionada en la sección 4.1.

$$\boldsymbol{H} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$$

Para este análisis, el parámetro que nos va a interesar variar es el tamaño muestral. Los que se van a considerar son: $100, 200, 500$ y 1000. Una vez presentadas las condiciones bajo las que se llevan a cabo las simulaciones, se puede comenzar con el estudio de los tiempos computacionales para ambos procedimientos. Para ello, se muestra a continuación la Tabla 4, en la que se presentan los valores obtenidos de tiempos de computación medios para las estimaciones de la tendencia utilizando o no *binning*.

Tamaño muestral	*Con Binning*	*Sin Binning*
100	0.018	0.021
200	0.038	0.043
500	0.129	0.674
1000	0.289	5.642

Tabla 4. Comparación tiempos computacionales (en seg.) de estimación de la tendencia para distintos tamaños muestrales.

Como era de esperar, el tamaño muestral tiene una gran influencia en el tiempo necesario para estimar la tendencia, ya que esto implica que aumente el número de operaciones a realizar. En lo que respecta a los términos comparativos entre la metodología con *binning* y la que no lo ha utilizado, se puede apreciar en la tabla como ambos procedimientos tienen unos costes computacionales bastante similares para conjuntos de datos pequeños, sin embargo, a medida que aumenta el tamaño muestral, se puede apreciar como el que empleó *binning* sobre los datos reduce significativamente los tiempos requeridos para llevar a cabo la estimación. Estos resultados obtenidos para simulaciones, no hacen sino refutar los que se mostraron en el capítulo anterior de este trabajo para datos reales, en el que se había empleado para el estudio el conjunto de datos *precipitation*, formado por 1053 observaciones, y donde se obtuvo una relación similar a la de este apartado entre el *binning* y la reducción de los costes computacionales.

Además, también es posible representar gráficamente la evolución de los tiempos computacionales para cada número de *bins* considerados, a medida que se aumenta el tamaño muestral sobre el que se aplican. Esto puede verse en la Figura 16.

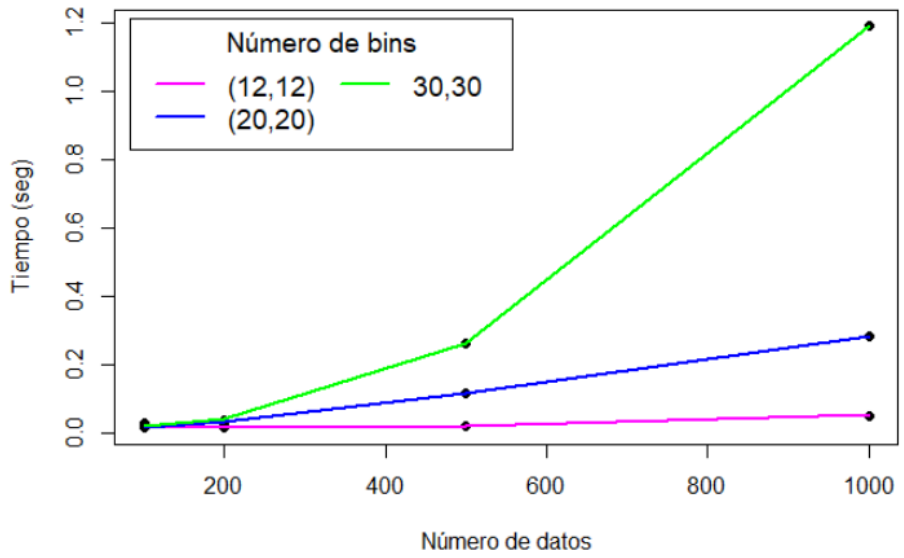

Figura 16. Evolución de los tiempos computacionales (en seg.) para estimar la tendencia según tamaño muestral y el número de bins.

Este gráfico permite apreciar de una manera visual como claramente la selección del número de *bins* tiene una gran importancia en términos computacionales y, en especial, cuando el número de observaciones disponibles es elevado, ya que en esta situación los tiempos computacionales aumentan de una manera exponencial. Por lo que, como ya se comentó anteriormente, en esta situación se recomendaría emplear un número moderado de *bins*.

4.4. Estudio de la relación entre el tamaño muestral y el número de *bins*

Como se presentó en la sección 2.4., el *binning* es un procedimiento que consiste en agrupar los datos de una muestra en *bins* con el objetivo de reducir la cantidad de operaciones necesarias cuando se emplee a ese conjunto de datos. De esta definición se puede extraer que el *binning* tiene una relación muy estrecha con el número de datos disponibles, ya que en una situación donde hay pocos datos, pero se emplean muchos *bins*, cada uno de ellos estará formado por una muy pequeña cantidad de valores, pudiendo provocar incluso que en vez de reducir los tiempos computacionales los aumente (debido al tiempo requerido para realizar el *binning*). La situación opuesta tampoco sería adecuada, si hay muchos datos, pero se emplean pocos *bins* es posible que el estimador no funcione de una manera adecuada.

Por este motivo, esta sección se va a centrar en el estudio de esta relación, para de esa manera

ser capaces de averiguar el número de bins que podrían ser adecuados para distintos tamaños muestrales. Por ello, se analizarán tanto los tiempos computacionales requeridos en cada uno de los casos, como la precisión de las estimaciones de la tendencia que se producen empleando conjuntos de datos de distintos tamaños y con distintos números de *bins*.

4.4.1. Estudio de los tiempos computacionales

Como bien se acaba de introducir, en este apartado del trabajo se va a examinar cómo afectan los valores de los distintos parámetros sobre los tiempos requeridos para estimar la tendencia para el caso en el que se emplea un conjunto de datos sobre el que se ha realizado *binning*. Para poder realizar este estudio, es necesario presentar los distintos escenarios que se van a considerar. En este caso, uno de los parámetros que se va a variar es el tamaño muestral, y se considerarán cuatro distintos: $100, 200, 500$ y 1000. El otro parámetro que va a variar será el número de *bins*, con valores considerados: $(12,12)$, $(20,20)$ y $(30,30)$, es decir, $144, 400$ y 900. Por otro lado, también destacar que la matriz de ventana que se va a utilizar será.

$$\boldsymbol{H} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$$

Introducidas las condiciones bajo las que se llevan a cabo estas simulaciones, se puede comenzar el análisis de la relación entre el tamaño muestral y el número de bins. Para ello, se presenta en la Tabla 5 los resultados obtenidos.

Tamaño muestral/Nºbins	$(12,12)$	$(20,20)$	$(30,30)$	*Sin binning*
100	0.0156	0.0176	0.0223	0.0220
200	0.0229	0.0372	0.0505	0.0385
500	0.0288	0.1158	0.2831	0.6410
1000	0.0302	0.2640	1.1896	5.5831

Tabla 5. Tiempos computacionales medios (en seg.) requeridos estimando la tendencia según el tamaño muestral y el número de bins.

Si se observa la tabla, se puede apreciar que tanto el tamaño muestral como el número de *bins* tienen una relación directa con los tiempos requeridos para realizar la estimación de la tendencia. Esto es fácilmente explicable, pues a mayor cantidad de cálculos, más elevado será el coste computacional correspondiente. En concreto, se puede apreciar que si uno de los dos factores, ya sea el tamaño del conjunto de datos o la cantidad de *bins*, es pequeño, que el otro tenga un valor elevado no va a suponer un gran incremento en términos computacionales, esto quiere decir que los tiempos se van a ver realmente amplificados cuando ambos parámetros tomen valores elevados.

4.4.2. Precisión de las estimaciones

Una vez analizada la relación entre el tamaño del conjunto de datos y el número de bins, otro factor a tener en cuenta es la precisión de las estimaciones que se producen, ya que no sería adecuado seleccionar para un conjunto de datos grande un número de bins pequeños si en esa situación se produjesen malas estimaciones. Con el objetivo de resolver esta cuestión, este apartado se va a centrar en el estudio del nivel de precisión de las estimaciones de la tendencia en función de la cantidad de datos disponibles y del número de bins que se emplean. Para que este estudio sea justo, se va a realizar bajo las mismas condiciones que se consideraron en el apartado anterior, y de ese modo tener las referencias tanto de error como de costes computacionales para todas ellas y poder llegar a conclusiones sobre qué cantidades de bins son adecuadas según el tamaño del conjunto de datos. Basándonos en esto, los tamaños muestrales que se van a considerar son: $100, 200, 500$ y 1000, mientras que los números de bins serán: $(12,12)$, $(20,20)$ y $(30,30)$ y todo ello utilizando la matriz de ventanas que se seleccionó en la sección 4.1., que es la que toma 0.1 como sus valores diagonales. En las Tablas 6-9 muestran el MSE, la mediana y la desviación típica de cada uno de los tamaños muestrales considerados:

Con 100 datos	$(12,12)$	$(20,20)$	$(30,30)$	*Sin binning*
Media	0.5585794	0.5129347	0.5009376	0.5039101
Mediana	0.2348137	0.2238562	0.2201710	0.2206104
Desviación típica	0.8810715	0.7704733	0.7473838	0.7534578

Tabla 6. MSE estimando la tendencia para distintos números de bins y sin binning con conjuntos de 100 datos.

Con 200 datos	(12,12)	(20,20)	(30,30)	*Sin binning*
Media	0.4559508	0.4759807	0.4598075	0.4761356
Mediana	0.2022841	0.2093701	0.2045666	0.2093685
Desviación típica	0.6700544	0.7095927	0.6734679	0.7052075

Tabla 7. MSE estimando la tendencia para distintos números de bins y sin binning con conjuntos de 200 datos.

Con 500 datos	(12,12)	(20,20)	(30,30)	*Sin binning*
Media	0.4629312	0.4581291	0.4575439	0.4584019
Mediana	0.2041548	0.2040691	0.2033928	0.2038433
Desviación típica	0.6759449	0.6661541	0.6643016	0.6644704

Tabla 8. MSE estimando la tendencia para distintos números de bins y sin binning con conjuntos de 500 datos.

Con 1000 datos	$(12,12)$	$(20,20)$	$(30,30)$	*Sin binning*
Media	0.4540014	0.4531031	0.4518731	0.4519142
Mediana	0.2006206	0.2008029	0.2003313	0.2008404
Desviación típica	0.6604476	0.6547808	0.6526779	0.6513686

Tabla 9. MSE estimando la tendencia para distintos números de bins y sin binning con conjuntos de 1000 datos.

El primer análisis que puede realizarse sobre estas tablas es con respecto a los MSE en función del tamaño muestral En este caso, los resultados obtenidos muestran que, independientemente del número de bins considerado, al aumentar el tamaño muestral, el error cometido en la estimación de la tendencia se reduce, así como su variabilidad. Si se hace un análisis comparativo entre el procedimiento con *binning* y el que no lo utiliza, puede verse como el MSE cometido por el primero de ellos es ligeramente mayor que para el procedimiento sin *binning*, aunque cuando el número de *bins* empleados es elevado, el comportamiento de ambos métodos es muy parecido.

Otro tema de interés es estudiar la calidad de las estimaciones en función del número de *bins* empleado para conjuntos de datos de distintos tamaños. Los valores obtenidos muestran que a medida que aumenta la cantidad de *bins* en los que se agrupan los datos, el MSE se reduce, siendo los menores los obtenidos para $(30,30)$ *bins*. Sin embargo, este aumento en términos de precisión de las estimaciones se consigue a cambio de unos mayores costes computacionales (como se mostró en 4.4.1.), por lo que puede haber ocasiones en las que en lugar de escoger un número excesivamente grande de *bins*, sería más adecuado optar por un término intermedio (sobre todo si la mejora en términos de precisión es pequeña). Basándonos en todo esto, a partir de cantidades iguales o superiores a 1000 datos sería recomendable utilizar $(30,30)$ *bins*, mientras que para conjuntos de datos de menor tamaño la opción $(20,20)$ actuaría adecuadamente, ya que de esta manera se obtendrían estimaciones muy parecidas a las del caso sin *binning* y con unos tiempos asumibles.

4.5. Estudio del comportamiento del *binning* según el grado de dependencia

4.5.1. En función del efecto *nugget*

Este apartado del trabajo se va a centrar en el estudio que tiene el efecto *nugget* o pepita sobre el nivel de precisión de las estimaciones de la tendencia, para lo que se emplearán los MSE de las

distintas estimaciones que se llevarán a cabo. Previamente al estudio, se presentan los escenarios en los que se realizan las simulaciones. Para el método con *binning*, el número de *bins* seleccionados es de $(20{,}20)$, que se aplicarán sobre un conjunto de datos de tamaño 200. La matriz de ventanas $\boldsymbol{H}$ que se utilizará será la seleccionada en la sección 4.1., es decir, la que tiene 0.1 como valores diagonales. Y los parámetros de la dependencia espacial van a ser los presentados en 4.1., es decir: el umbral vale 1 y el rango es 0.6.

Por otro lado, para llevar a cabo este estudio se van a considerar cuatro valores distintos para el efecto pepita: 0, 0.2, 0.5 y 1. El umbral parcial se tomón de la forma: $c_1 = 1 - c_0$, de manera que la varianza fuese siempre 1. Una vez presentado en qué va a consistir el análisis, se muestran a continuación en la Tabla 10 los MSE obtenidos para cada caso.

Efecto nugget	*Con Binning*	*Sin Binning*
0	0.5657629	0.5660389
0.2	0.4759807	0.4761356
0.5	0.3423968	0.3422353
1	0.1201994	0.1193323

Tabla 10. MSE medio de la estimación de la tendencia para distintos efectos nugget.

Basándonos en que a menor efecto *nugget* mayor es la dependencia presente en los datos, de forma que cuando vale 0 estamos ante una situación de máxima dependencia y si es 1 se considera el caso de independencia, los resultados obtenidos muestran que a mayor dependencia mayores errores de estimación se producen, siendo el MSE en la situación de máxima dependencia cinco veces mayor que para el caso de independencia. Por otro lado, los MSE que se obtuvieron tanto para el método que emplea el conjunto de datos original como el que emplea el conjunto *binned* no muestran grandes diferencias. En concreto, para los primeros casos considerados, es decir, en los que el efecto *nugget* es 0 y 0.2, los errores cometidos son menores para el caso con *binning*, mientras que para los valores 0.5 y 1 el MSE es menor para el procedimiento que utiliza al conjunto de datos original. Debido a esto, y a que las diferencias en las cuatro circunstancias consideradas son muy pequeñas, no se puede considerar que la dependencia espacial tenga una influencia

significativa a la hora de generar diferencias de precisión entre las estimaciones de la tendencia utilizando o no *binning*.

4.5.2. En función del rango

El rango es una de las componentes junto con el efecto *nugget* que mayor importancia tienen a la hora de determinar la dependencia espacial. Por ello, puede resultar de interés estudiar cómo afecta su valor al comportamiento de la estimación de la tendencia con y sin *binning*. Con este objetivo, se van a realizar simulaciones para ambos procedimientos variando el valor del rango y obteniendo los MSE de las estimaciones. Para poder realizar esto, es necesario fijar las circunstancias bajo las que se va a realizar el análisis, que van a ser similares a las empleadas para el estudio del efecto *nugget*. La cantidad de *bins* que se va a utilizar para agrupar los datos es $(20,20)$, es decir: 400 *bins*. El tamaño muestral es 200. La matriz de ventanas que se va a utilizar es:

$$\boldsymbol{H} = \begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$$

En lo referente a los parámetros que definen a la dependencia espacial, el umbral considerado es 1, mientras que el efecto pepita es 0.2. El rango, que es el componente a estudiar, va a tomar los valores: 0.3, 0.6 y 0.9, siendo a mayor valor mayor la dependencia, de tal forma que 0 representaría el caso de independencia y 1 el de máxima dependencia. Los MSE obtenidos bajo estas condiciones que se acaban de especificar se muestran en la Tabla 11.

Rango	*Con Binning*	*Sin Binning*
0.3	0.3259143	0.3256064
0.6	0.4759807	0.4761356
0.9	0.5609910	0.5611605

Tabla 11. MSE medio de la estimación de la tendencia para distintos rangos.

Los MSE cometidos en la estimación de la tendencia espacial para los distintos valores del rango considerados muestran como cuanto menor es, es decir, a menor dependencia, mejor es la calidad de las estimaciones. Por otro lado, se puede hacer un análisis comparativo entre el estimador que emplea un conjunto de datos *binned* y el que utiliza el original. Según los MSE obtenidos, la dependencia espacial no provoca diferencias significativas en la calidad de las estimaciones entre estos dos procedimientos, a pesar de que en todas las circunstancias es mejor el MSE cometido por el procedimiento con *binning*, las diferencias entre ambos métodos en todos los casos inferiores a 0.001.

Por tanto, las conclusiones a las que se han llegado en este apartado sobre la influencia de la dependencia en la estimación de la tendencia reafirman a las obtenidas en el apartado anterior. En resumen, la dependencia afecta negativamente a la calidad de las estimaciones, y el procedimiento con *binning* actúa de una manera similar al método que no lo utiliza.

Capítulo 5

5. Efecto del *binning* en la estimación del variograma

Esta parte del trabajo se va a centrar en el análisis del comportamiento del *binning* a la hora de estimar el variograma. Para ello, la sección 5.1. se centra en una introducción al estudio del variograma. En la sección 5.2. se lleva a cabo un análisis de la estimación del variograma en función del número de datos y de *bins* tanto de manera separada como conjunta. Por último, en 5.3. se comprobará como afecta tanto el efecto *nugget* como el rango a la calidad de las estimaciones del variograma. Por último, destacar que para los estudios de este capítulo se van a emplear 1000 simulaciones.

5.1. Estudio del variograma

En este capítulo se va a estudiar mediante simulaciones la estimación del variograma (concepto presentado en la sección 1.2.). En concreto, se va a centrar en el análisis de la precisión de las estimaciones que se consiguen a partir de los estimadores piloto del variograma tanto para el caso en el que se emplea un conjunto de datos *binned* como en el que se utiliza uno sin agrupar. Pero antes de comenzar con este estudio, se va a mostrar gráficamente el variograma teórico junto con las estimaciones obtenidas utilizando los datos originales y los agrupados. Es necesario destacar que, en este capítulo del trabajo, partiendo del modelo general de los procesos geoestadísticos (1.1), se va a considerar un caso de tendencia constante y, en concreto con valor 0, es decir: $\mu(\boldsymbol{x}) = 0$. Por lo que este modelo podría reescribirse de la forma:

$$Y(\boldsymbol{x}) = \varepsilon(\boldsymbol{x})$$

Además, el variograma teórico del que se va a partir en este apartado va a ser similar al de la expresión (4.1), es decir, suponemos un efecto *nugget* de 0.2, un rango de 0.6 y que el umbral es 1. De esta forma, y partiendo de que las estimaciones se realizaron a partir de un conjunto de 200 datos y que en el caso con *binning* se utilizaron $(20{,}20)$, se muestra en la Figura 17 la representación gráfica del variograma teórico y sus estimaciones.

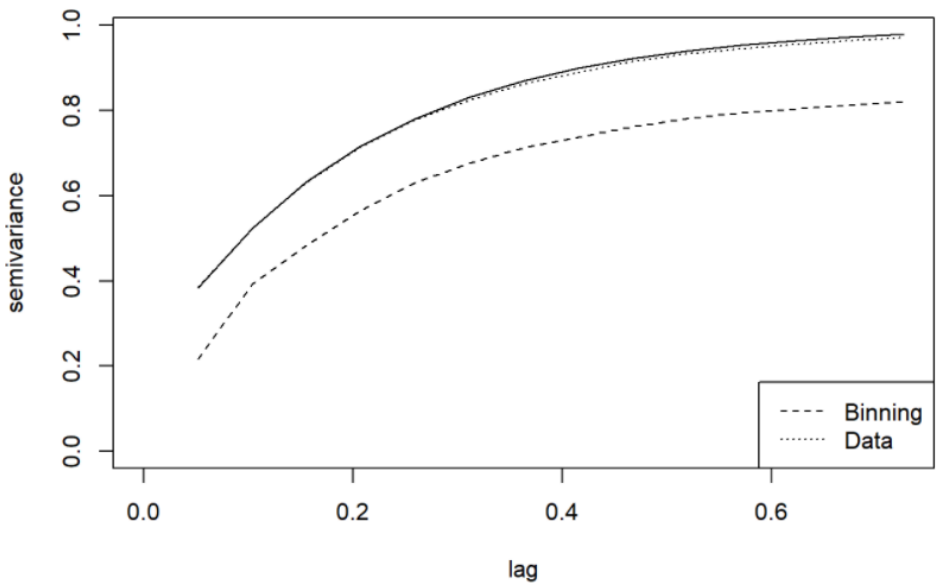

Figura 17. El variograma teórico y sus estimaciones.

En la Figura 17 puede verse como la estimación sin *binning* se encuentra más próxima a la teórica que la que se realizó con *binning* en la situación planteada, este efecto es el esperado, ya que la variabilidad de un promedio es inferior a la de los datos originales. Se podría tratar de corregir el sesgo, pero hay que tener en cuenta que las predicciones *kriging* no varían si se desplaza el variograma sumándole una constante. Para medir estas diferencias entre las estimaciones y el variograma teórico existen múltiples criterios de error, uno de ellos, que ya ha sido utilizado previamente en este trabajo, es el MSE. Para la situación planteada, puede representarse gráficamente también los MSE que cometen cada una de las estimaciones (ver Figura 18).

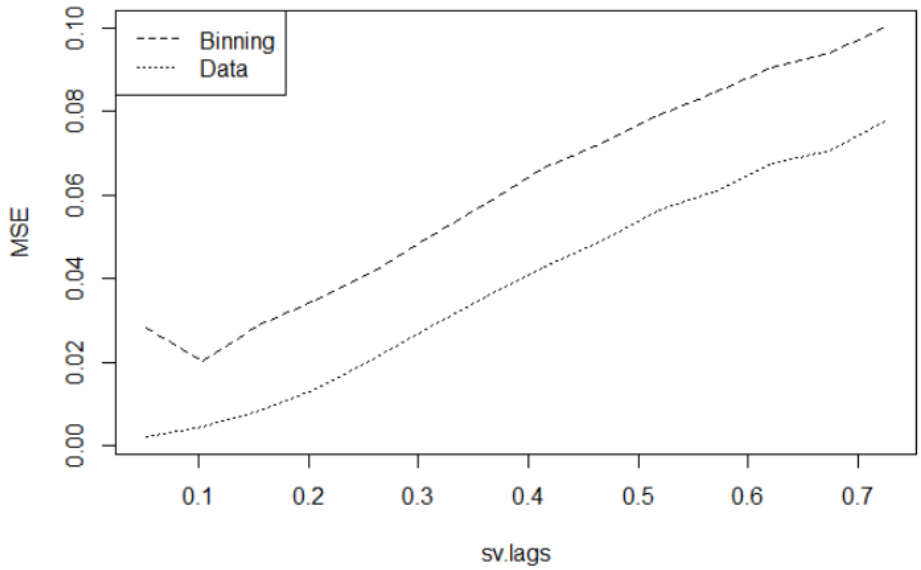

Figura 18. MSE medio de las estimaciones del variograma con y sin binning.

Antes de nada, destacar que en el eje horizontal se muestran las distancias consideradas. En la Figura 18 puede verse como, en efecto, es la estimación en la que se emplea *binning* donde se comete un mayor MSE. Para llegar a una conclusión sobre si es recomendable su implementación, se llevará a cabo en las siguientes secciones un análisis sobre la estimación del variograma bajo diversas circunstancias. Además, destacar que para estudiar el comportamiento del *binning* en la estimación del variograma, los tiempos computacionales no tienen una gran relevancia. Como primera aoriximación se implementó el estimador (2.2) de modo directo. Si se empleara un algoritmo basado en la FFT, los tiempos computacionales se reducirían drásticamente. A pesar de esto, a ejemplo ilustrativo, en ciertos casos sí que se mostrarán.

5.2. Estudio del *binning* en función del tamaño muestral y número de *bins*

5.2.1. Estimación del variograma para distintos tamaños muestrales

Esta sección se centra en el estudio de la precisión de las estimaciones del variograma considerando distintos tamaños muestrales y comparando los resultados obtenidos tanto para el caso con *binning* como para el que no lo emplea. Para ello, el criterio de error que se empleará es el MSE. Antes de analizar los resultados, se presentan las condiciones bajo las que se realiza el estudio. Los conjuntos de datos que se van a considerar son de: 100, 200, 500 y 1000 datos. El número de *bins* empleados para el caso con *binning* son $(20,20)$, es decir, 400. Y considerando, como ya se planteó en la introducción de este capítulo, que la tendencia es constante con valor 0. En la Tabla 12 se muestran los MSE obtenidos para ambos procedimientos.

Tamaño muestral	*Con Binning*	*Sin Binning*
100	0.056	0.052
200	0.061	0.039
500	0.093	0.036
1000	0.114	0.035

Tabla 12.. MSE medio de la estimación del variograma para distintos tamaños muestrales.

Los resultados obtenidos muestran dos tendencias contrarias, a medida que aumenta el tamaño muestral el MSE cometido en la estimación del variograma para el método que emplea *binning* aumenta, mientras que para el caso sin *binning* se reduce, aunque siendo casi todos ellos valores relativamente pequeños. Además, también puede apreciarse que el procedimiento sin *binning* tiene una mejor precisión de estimación en todas las situaciones planteadas. Además, también puede resultar interesante analizar la evolución de los tiempos computacionales a medida que aumenta el tamaño muestral, aunque como ya se comentó al comienzo de este capítulo, este factor no va a tener una gran relevancia en la estimación del variograma. Los tiempos (en segundos) obtenidos se presentan a continuación en la Tabla 13 tanto para el caso con *binning* como para el que no lo utiliza.

Tamaño muestral	*Con Binning*	*Sin Binning*
100	0.00691	0.0021
200	0.01053	0.00426
500	0.02091	0.02855
1000	0.01838	0.10861

Tabla 13. Tiempos computacionales estimando el variograma para distintos tamaños muestrales.

Los tiempos obtenidos muestran la evolución esperada, a medida que aumenta el número de datos disponibles mayor es el tiempo requerido para llevar a cabo la estimación del variograma, pero a cambio consiguiéndose una mejor precisión.

Todo lo anteriormente comentado tanto en términos de precisión como de tiempos computacionales, puede deberse a que la cantidad de *bins* escogida a la hora de hacer *binning* no fuera la adecuada. Para comprobar cómo afecta este factor a la estimación del variograma, en la siguiente sección se estudia el MSE cometido para distintas cantidades de *bins.*

5.2.2. Estimación del variograma para distintas cantidades de *bins*

Después de estudiar cómo afecta el tamaño muestral al comportamiento del *binning* con respecto al procedimiento que emplea el conjunto de datos original en la estimación del variograma, otro factor de gran relevancia es la cantidad de *bins* que se empleen. Por ello, en esta sección se van a obtener los MSE para tres situaciones: 144, 400 y 900 *bins*, es decir, $(12,12)$, $(20,20)$ y $(30,30)$, respectivamente, y también el caso sin *binning*. Además, todo este estudio se realizará bajo las condiciones: 200 datos y siendo la tendencia constante con valor 0. Una vez presentadas las condiciones, se muestran en la Tabla 14 los MSE obtenidos.

Con 200 datos	$(12,12)$	$(20,20)$	$(30,30)$	*Sin binning*
Media	0.133	0.061	0.042	0.039
Mediana	0.138	0.062	0.043	0.040
Desviación típica	0.030	0.027	0.025	0.026

Tabla 14. MSE medio cometido en la estimación del variograma para distintos bins.

Basándonos en la tabla, como cabría esperar, parece que a medida que aumenta la cantidad de *bins* que se aplican, se mejora la calidad de las estimaciones, siendo en el caso $(30,30)$, una estimación con un MSE muy parecido al del caso sin *binning*. Además, como puede observarse en los valores de la desviación típica, a medida que aumenta el número de *bins* menor es la variabilidad del error. Esta mejora en términos de estimación al aumentar el número de *bins* por lo general puede provocar unos mayores costes computacionales. En la Tabla 15 se presentan los requeridos en cada caso.

	$(12,12)$	$(20,20)$	$(30,30)$	*Sin binning*
Tiempo (seg)	0.00324	0.01137	0.0266	0.00469

Tabla 15. Tiempos computacionales (en seg.) para estimar el variograma con distintas cantidades de bins.

En efecto, la suposición previa sobre los tiempos computacionales se cumple, a mayor cantidad de *bins* mayores tiempos requeridos, siendo por ejemplo para el caso $(30{,}30)$ el tiempo de estimación medio obtenido más del doble que para $(20{,}20)$. Sin embargo, como ya se comentó al comienzo de este capítulo, los tiempos computacionales no tienen una gran importancia en la estimación del variograma, ya que la implementación del algoritmo FFT permitiría reducirlos enormemente, aunque en este caso ya son todos los tiempos presentados muy pequeños.

5.2.3. Estudio de la relación entre el tamaño muestral y el número de *bins*

Como se comprobó en las secciones 5.2.1. y 5.2.2., tanto el tamaño muestral como el número de *bins* tienen una gran importancia a la hora de determinar la precisión de las estimaciones. Además, estos factores no son independientes (como pudo verse en la sección 4.4. para el caso de la tendencia), por lo que puede ser de gran interés analizar de qué manera afectan a la estimación del variograma, teniendo la ventaja en este caso con respecto a la tendencia de que solo nos importa la precisión de las estimaciones. Con este objetivo, presenta a continuación en la Tabla 16 los MSE que se obtienen en la estimación del variograma, considerando conjuntos de datos de distintos tamaños: 100, 200, 500 y 1000 datos, y distintos número de *bins*: $(12{,}12)$, $(20{,}20)$ y $(30{,}30)$.

Tamaño muestral/Nºbins	$(12{,}12)$	$(20{,}20)$	$(30{,}30)$	*Sin binning*
100	0.093	0.056	0.051	0.052
200	0.133	0.061	0.042	0.038
500	0.164	0.093	0.055	0.036
1000	0.181	0.114	0.073	0.033

Tabla 16. MSE cometido en la estimación del variograma para distintos tamaños muestrales y número de bins.

Los resultados obtenidos muestran como a medida que se aumenta el número de *bins* empleados a la hora de hacer *binning* menor es el MSE cometido en la estimación del variograma, siendo muy cercano (e incluso algo menor en un caso) al cometido por el procedimiento sin *binning*.

Esto ocurre en especial en tamaños muestrales no muy elevados, como puede apreciarse, a medida que aumenta el número de datos también lo hacen las diferencias entre los procesos con y sin *binning*. De esta forma, y basándonos en lo comentado anteriormente sobre la poca relevancia de los tiempos computacionales en esta estimación, si se implementa el *binning* se recomendaría emplear una cantidad elevada de *bins*.

5.3. Estimación del variograma para distintos niveles de dependencia

5.3.1. Según el efecto *nugget*

Un factor de gran importancia a la hora de determinar el grado de dependencia espacial es el efecto *nugget* o pepita. En la sección 4.5. se pudo comprobar como tenía una gran influencia sobre la calidad de las estimaciones de la tendencia y, en este caso, interesa analizar si también la va a tener para la estimación del variograma. Para ello, se plantean los mismos cuatro valores para el efecto pepita que en la estimación de la tendencia: 0, 0.2, 0.5 y 1. Además, el estudio se realizará para el caso en el que se dispone de 200 datos, y que para el *binning* se emplearán: $(20{,}20)$ *bins.* Por otro lado, destacar que se considera un rango de 0.6 y un umbral de 1. Una vez planteadas las condiciones bajo las que se lleva a cabo el estudio, se exponen a continuación en la Tabla 18 los MSE obtenidos en cada caso.

Efecto nugget	*Con Binning*	*Sin Binning*
0	0.061	0.056
0.2	0.061	0.039
0.5	0.088	0.021
1	0.200	0.013

Tabla 17. MSE medio cometido en la estimación del variograma con distintos efectos nugget.

Partiendo de que la situación de máxima dependencia es en la que el efecto pepita es 0, y la de independencia cuando vale 1, el MSE cometido para los métodos con y sin *binning* evolucionan de una manera dispar. En el caso sin *binning*, a menor dependencia menor error se comete, lo que coincide con lo obtenido para la estimación de la tendencia en (4.5.1.), mientras que, por la contra, el procedimiento con *binning* estima peor el variograma a medida que disminuye la dependencia.

5.3.2. Según el rango

Este apartado se centra en el estudio de la estimación del variograma en función de su alcance. Con este objetivo, se plantean tres valores distintos: 0.3, 0.6 y 0.9, para los que se medirán los MSE en los que se incurre a la hora de llevar a cabo esta estimación. Las circunstancias consideradas para este análisis son similares a las del apartado anterior: 200 datos y $(20,20)$ *bins* para el caso con *binning*. Además, se considera un efecto pepita de 0.2 y un umbral de 1. En la Tabla 18 se exponen los MSE obtenidos en cada caso, para así poder comentar posteriormente su evolución.

Rango	*Con Binning*	*Sin Binning*
0.3	0.065	0.026
0.6	0.061	0.039
0.9	0.058	0.041

Tabla 18. MSE medio de la estimación de la tendencia para distintos rangos.

Siendo el escenario de máxima dependencia cuando el rango vale 1 y la de independencia cuando es 0, los MSE cometidos en cada una de las situaciones indican una evolución similar a la considerada en el apartado anterior, la precisión de las estimaciones de ambos procedimientos evolucionan de una manera contraria, a medida que aumenta la dependencia se mejora la calidad de las estimaciones del método con *binning* y empeora la del procedimiento que no lo emplea aunque en este caso la variación es mucho menor que para el efecto *nugget*.

Conclusiones y líneas futuras

Este trabajo se ha centrado en el estudio del efecto del *binning,* comparándolo con la aproximación tradicional, a la hora de realizar estimaciones tanto de la tendencia espacial como del variograma. Para ello, se utilizaron principalmente dos criterios de medición: los tiempos computacionales y la calidad de las estimaciones. Este estudio en primer lugar se llevó a cabo sobre la estimación de la tendencia para un conjunto de datos reales y posteriormente sobre datos generados mediante simulaciones. Por un lado, se demostró en 3.2. y en 4.3. lo que se planteó teóricamente sobre los tiempos computacionales, al agrupar los datos se reducen, lo que es debido a que la cantidad de cálculos necesario para realizar la estimación es menor, y esto se nota especialmente cuando se dispone de un tamaño muestral muy elevado, siendo de gran importancia la selección adecuada del número de *bins*, pues si es demasiado grande provocaría que no se produjese esa ganancia en términos computacionales.

Con respecto a la precisión de las estimaciones de la tendencia, en el estudio llevado a cabo en 4.4. se concluyó que el empleo de *binning* no afecta de una manera negativa siempre y cuando la cantidad de *bins* que se utilice no sea demasiado pequeña, ya que a medida que aumenta la cantidad empleada, también lo hace la calidad de las estimaciones. Como se observó en la sección 4.5., un factor que sí que tiene un efecto negativo en la precisión de las estimaciones es la dependencia, ya que cuanta más hay presente en los datos peor funcionan los estimadores. Sin embargo, es necesario destacar que perjudica de igual manera al procedimiento con *binning* como al que utiliza el conjunto de datos original, así que no es relevante a la hora de determinar si el empleo de *binning* es adecuado o no. Otro factor que también tiene una gran relevancia sobre el estimador de la tendencia es el ancho de ventana, por lo que también resultaría aconsejable seleccionar un valor adecuado para este parámetro antes de realizar estimaciones de la tendencia.

Basándonos en el análisis previamente realizado, se puede concluir sobre la estimación de la tendencia espacial que es recomendable aplicar el método con *binning* siempre y cuando se escoja un número de *bins* adecuado, ya que este factor va a ser el más determinante a la hora de establecer si la implementación del *binning* es o no beneficiosa a la hora de realizar estudios. Por lo que previamente a su empleo, es aconsejable analizar el tamaño de la muestra de la que se dispone y en base a esto escoger que cantidad de *bins* a utilizar, de manera que no fuera demasiado elevada, provocando muy buenas estimaciones pero con costes computacionales muy elevados, pudiendo en casos extremos incluso superar a los que se tendrían utilizando al conjunto de datos original (debido a los cálculos que se requieren para realizar *binning*), ni que tampoco demasiado pequeña, situación en la que las estimaciones se calcularían de una manera muy rápida pero tendrían con una calidad muy baja. Uno de los criterios más comúnmente utilizados en la práctica es el de escoger como número de *bins* a la raíz cuadrada del tamaño muestral.

Con respecto a la estimación del variograma, los resultados mostraron como su comportamiento en función del tamaño de la muestra y el número de *bins* es similar al de la estimación de la tendencia, es decir, a mayor tamaño muestral mejor actúa tanto el estimador LL que emplea *binning* como el que no. Para la cantidad de *bins* ocurre lo mismo, a mayor cantidad mayor calidad de las estimaciones. Sin embargo, como ya se comentó previamente, la estimación del variograma tiene la gran ventaja de que no hay que preocuparse de los tiempos computacionales al implementar *binning*, ya que la implementación del

algoritmo FFT permitiría reducirlos enormemente. Por lo que, basándonos en esto, para la estimación del variograma se recomendaría implementar *binning* y además utilizando un número elevado de *bins* (en torno a $(30,30)$ en el caso bidimensional), para que la estimación sea lo más precisa posible.

Por otro lado, otro factor que también condiciona la precisión de las estimaciones es la dependencia. En este caso se obtuvo una discordancia entre el caso con y sin *binning*, ya que los estudios realizados mostraron como a mayor dependencia mejor es la estimación del procedimiento con *binning* y peor la del método que emplea el conjunto de datos original, mientras que lo contrario ocurre cuando se reduce la dependencia.

En conclusión, fundamentándonos en todo lo anteriormente nombrado, la implementación del *binning* es por término general recomendable, permite reducir enormemente los tiempos computacionales de las estimaciones sin que se vean afectadas en gran medida la calidad de las estimaciones, en especial cuando el conjunto de datos sobre el que se aplica es de un tamaño grande (superior a los 1000 datos). Pero, previamente a su aplicación es aconsejable hacer un análisis del conjunto de datos a estudiar (cantidad de datos, dependencia, etc.), para así poder escoger valores adecuados para los parámetros (ancho de ventana, número de *bins*), y de este modo maximizar el beneficio conseguido al implementar el *binning*. Una de las opciones para escoger el número de *bins* es optar por la raíz cuadrada del tamaño muestral, mientras que para la selección de la ventana (matriz en caso multidimensional), un criterio adecuado sería el presentado en Francisco-Fernández y Opsomer (2005).

A partir del estudio presentado a lo largo del presente trabajo, se podría plantear llevar a cabo en el futuro estudios sobre el comportamiento del *binning* en otros ámbitos de la estadística espacial. Algunos ejemplos podrían ser, por ejemplo: la precisión de las estimaciones del variograma cuando se emplea un modelo paramétrico en función de si se emplea el conjunto de datos original o el *binned* o también, su influencia en el comportamiento del *kriging*. Este último estudio, por falta de tiempo no se realizó. La idea sería emplear datos agrupados para realizar *kriging* en lugar de las observaciones originales, el tiempo de computación se reduciría enormemente al disminuir el tamaño de las matrices de covarianzas. El estimador (2.3), no es muy adecuado parar aproximar la variabilidad de los datos originales, pero se podría emplear para aproximar la de los datos agrupados (además de que, si se desplaza el variograma sumándole una constante, las predicciones no varían). A parte de estos, también se pueden nombrar otros estudios que podrían resultar de interés, y que serían modificaciones a los realizados a lo largo de este trabajo: la estimación de la tendencia espacial con distintos modelos teóricos, el estudio de la estimación del variograma para distintos modelos de dependencia, estudiar la estimación del variograma cuando la tendencia no es constante, la corrección del sesgo de la estimación del variograma o cantidades auxiliares tales como medidas de grados de libertad.

Bibliografía

Armstrong M (1998). *Basic linear geostatistics*. Springer Science & Business Media.

Bowman AW, Azzalini A (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations* (Vol. 18). OUP Oxford.

Bowman AW, Azzalini A (2003). Computational aspects of nonparametric smoothing with illustrations from the sm library. *Computational statistics & data analysis*, *42*(4), 545-560.

Castillo SA (2017). Aportaciones a la Geoestadística No Paramétrica. Tesis, Universidad de Vigo.

Cressie N (1990). The origins of kriging. *Mathematical Geology*, *22*(3), 239-252.

Cressie N (1993). Statistics for Spatial Data. John Wiley and Sons.

Eubank RL (1999). *Nonparametric Regression and Spline Smoothing*. CRC Press.

Fan J, Gijbels I (1996). Local Polynomial Modelling and its Applications: monographs on statistics and applied probability 66, volumen 66. CRC Press.

Fernández-Casal R (2003). Geoestadística Espacio-Temporal: Modelos flexibles de Variogramas anisotrópicos no separables. Tesis, Universidad de Santiago de Compostela.

Francisco-Fernández M, Opsomer JD (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. Canadian Journal of Statistics, 33:539-558.

Francisco-Fernández M y Vilar-Fernández JM (2001). Local polynomial regression estimation with correlated erors. Communications in Statistics -Theory and Methods, 30:1271-1293.

Kuo FY, Sloan IH (2005). Lifting the curse of dimensionality. *Notices of the AMS*, *52*(11), 1320-1328.

Liu X (2001). *Kernel* Smoothing for Spatially Correlated Data. Tesis doctoral, Department of Statistics, Iowa State University.

Marcotte D (1996). Fast variogram computation with FFT. Computers & Geosciences, 22(10), 1175-1186.

Nadaraya EA (1964). On estimating regression. *Theory of Probability & Its Applications*, *9*(1), 141-142.

Opsomer JD, Wang Y, y Yang Y (2001). Nonparametric regression with correlated errors. Statistical Science, 16:134-153.

Reyes MA (2010) Estimación Paramétrica y No Paramétrica de la Tendencia en Datos con Dependencia Espacial. Un estudio de simulación. Technical report, Universidad Santiago de Compostela.

Rupert D y Wand, MP (1994). Multivariate locally weighted least squares regression. Annals of Statistics, 22:1346-1370.

Seifert B, Gasser T (1996). Finite-sample variance of local polynomials: analysis and solutions. *Journal of the American Statistical Association*, *91*(433), 267-275.

Turlach BA, Wand MP (1996). Fast computation of auxiliary quantities in local polynomial regression. *Journal of Computational and Graphical Statistics*, *5*(4), 337-350.

Wand MP (1994). Fast computation of multivariate kernel estimators. *Journal of Computational and Graphical Statistics*, *3*(4), 433-445.

Watson GS (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359-372.

Lista de figuras

Lista de tablas