UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

REVISTA UD Y LA GEOMÁTICA

http://revistas.udistrital.edu.co/ojs/index.php/UDGeo/index

DOI:http://dx.doi.org/10.14483/udistrital.jour.udgeo.2013.7.a08

INVESTIGACIÓN

Diseño de una red de muestreo espacial que optimiza la estimación de parámetros de la función de covarianza para la medición pluviométrica en el departamento de Cundinamarca

Spatial sampling design optimizing parameters estimation of the covariance function of the rainfall in Cundinamarca, Colombia

***Roverth Steven Pinzón Rodríguez*[a], *Juan Sebastián Pulido Ávila*[b], *Luis Fernando Santa Guzmán*[c], *Luis Fernando Gómez Rodríguez*[d].**

Citation / Para citar este artículo: Pinzón Rodríguez R. S., Pulido Ávila J. S., Santa Guzmán L. F. & Gómez Rodríguez L. F., (2013). Diseño de una red de muestreo espacial que optimiza la estimación de parámetros de la función de covarianza para la medición pluviométrica en el departamento de Cundinamarca. UD y la Geomática, (7), pp. 75 – 85.

Fecha de recepción: 01 de septiembre de 2013 / **Fecha de aceptación:** 01 de diciembre 2013

RESUMEN

En esta investigación se llevó a cabo una metodología que diseñó una red de muestreo espacial a través de la optimización de la estimación de los parámetros de la función de covarianza usando datos de pluviosidad del departamento de Cundinamarca, obtenidos por el Ideam para el año 2007. Se construyó un modelo de variabilidad mediante un análisis de dependencia espacial que sirvió de base para simular varios tamaños de muestra mediante el algoritmo de simulación Annealing, cuyas matrices de Fisher y de covarianza de los parámetros estimados por máxima verosimilitud (ML) fueron comparadas con el fin de obtener la mejor estimación de los parámetros del proceso espacial. Para el proceso se implementaron rutinas en el software R 3.0.0. A pesar de que la matriz de Fisher no fue una buena aproximación a la matriz de covarianza de los estimadores ML, puede ser usada como criterio de diseño si la relación entre ellas es monótona. Como resultado se encontró una mejor red de muestreo con estimación de parámetros óptima y que cumple con las recomendaciones establecidas por la Organización Mundial Meteorológica (OMM), que sirve como base para reubicar estaciones de la red actual cuya implementación no satisface los propósitos de muestreo óptimo por la falta de criterios técnicos durante su diseño, los cuales provocaron defectos de traslape y discontinuidades espaciales de muestreo. El proceso de reubicación logró minimizar la varianza de la predicción, que mejoraría las mediciones de la variable en la región, con el objeto de que planificadores tomen mejores decisiones para prevenir futuros desastres.

Palabras clave: kriging, matriz de Fisher, redes de muestreo, simulación Annealing.

ABSTRACT

In this study, a methodology was carried out to design a spatial sampling network through the optimization of the estimation of the parameters of the covariance function using rainfall data from the Department of Cundinamarca obtained from the IDEAM for the year 2007. A variability model was constructed using an analysis of spatial dependence which served as the basis to simulate various sample sizes through the Annealing simulation algorithm. The Fisher matrices and covariance matrices of the parameters estimated by maximum likelihood (ML) were compared in order to obtain the best estimate of the parameters of the spatial process. In order to perform the process, routines were implemented in the R 3.0.0 software. Although the Fisher matrix was not a good approximation of the covariance matrix of the ML estimators, it can be used as a design criterion if the relationship between them is monotonic, as a result a better spatial sampling network was found with the estimation of optimal parameters which also comply with the recommendations set by the World Meteorological Organization (WMO). This serves as a basis for relocating the existing network stations whose implementation does not meet the purposes of optimal sampling due to the lack of technical criteria for their design, which causes defects of overlap and spatial sampling discontinuities. The relocation process was able to minimize the variance of the prediction, which will improve the measurements of the variable in the region. As a result, planners can make better decisions to prevent future disasters, improve sustainable development and raise the level of knowledge of the human individual with regards to the natural environment that surrounds them.

Keywords: Kriging Method, Fisher Matrix, Sampling networks, Annealing simulation.

[a]Ingeniero en algo (e-mail: rspinzonr@gmail.com).
[b]Ingeniero en algo (e-mail: jusepa1989@gmail.com).
[c]Ingeniero en algo (e-mail: lfsantag@unal.edu.co).
[d]Ingeniero en algo (e-mail: luuruena@yahoo.es).

1. Introducción

Actualmente, las sociedades de todas partes del mundo se ven enfrentadas al reto de mejorar la predicción y cuantificación de las acciones de la naturaleza que se dan por si solas, con el objeto de prevenir desastres, mejorar el desarrollo sostenible y demás intervenciones, para lo cual crean herramientas que permitan elevar el nivel de conocimiento del ser humano en cuanto al entorno natural que lo rodea.

La meteorología, que es una de las múltiples disciplinas encargadas de estudiar las acciones naturales en el medio atmosférico, tiene como uno de sus objetivos la precipitación, generada en una de las fases del ciclo hidrológico y responsable del depósito de agua dulce en el planeta, que produce como consecuencias negativas inundaciones, erosión, desbordamientos de los ríos, deterioro en construcciones entre otras, que hacen que surja la necesidad de hacer mediciones respecto a la cantidad de agua precipitada en regiones de la superficie terrestre. Debido a la amplia extensión de la superficie, la precipitación no es igual en todas las regiones, por esta razón, para poder medir de manera precisa y reducir los costos de operación, se emplean redes de observación meteorológica que permita interpolar la información obtenida de las estaciones meteorológicas a las demás regiones donde no es posible realizar observaciones directas. En Colombia, el Instituto de Hidrología, Meteorología y Estudios Ambientales (Ideam), es el encargado de implementar, operar y mantener la red meteorológica nacional.

Debido a que actualmente no se han definido metodologías adecuadas y solamente se han utilizado criterios empíricos para establecer las estaciones de medición, es de vital importancia demostrar que la estadística espacial es una herramienta que posibilita la creación de las redes de muestreo espacial de forma óptima, la implementación de criterios de diseño que permiten definir los mejores modelos adecuados que se ajustan a un conjunto de datos, de tal forma que puedan ser evaluados, teniendo en cuenta que previamente han sido simulados con metodologías definidas para la optimización de modelos, con el fin de encontrar redes de muestreo espacial óptimas que puedan ser comparadas con redes de muestreo utilizadas en la actualidad, y que se posibiliten el conocimiento efectivo de los efectos ambientales, además de mejorar las actuaciones a fin de solucionar problemáticas que afectan a la población en general y que no alteren el desarrollo sostenible de la sociedad.

En consecuencia, el propósito del presente proyecto es implementar una metodología para la creación de una red de muestreo espacial basada en la optimización de los parámetros de la matriz de covarianza, de tal forma que sea lo más robusta posible con el fin de hacer mediciones pluviométricas para una zona definida de estudio, que en nuestro caso se concreta en el departamento de Cundinamarca.

2. Metodología

Para el desarrollo del presente proyecto se utilizaron las siguientes herramientas, imprescindibles durante la ejecución del proyecto. Se resalta que son de carácter libre, por lo que son de fácil adquisición:

- R, es un entorno de software libre para computación y gráficos estadísticos
- Quantum GIS (QGIS), es un Sistema de información geográfica, fácil de usar, licenciado bajo la GNU - General Public License.

Se utilizaron datos de precipitación del Ideam del departamento de Cundinamarca para el año 2007, los cuales cuentan con la posición (coordenadas x y y) de 151 estaciones y la medida de precipitación total anual en cada una de ellas.

La zona de estudio es el departamento de Cundinamarca, debido a su alta densidad de estaciones pluviométricas, además de ser una zona con cambios abruptos en la precipitación, reflejada en inundaciones que ocasionan problemas graves; por tanto, es una zona apropiada para hacer el diseño de una red de muestreo espacial de estaciones pluviométricas.

Análisis exploratorio para el proceso

En primer lugar, se hizo un análisis exploratorio donde se buscó detallar las medidas de tendencia central de la variable (a nivel univariado) sin tener en cuenta la dependencia espacial que esta pueda tener, con el fin de evaluar la posible existencia de tendencia de los datos o si, por el contrario, la variable tenía un comportamiento heterogéneo en su distribución de probabilidad.

Se realizaron gráficas para observar si en ellas había clara evidencia de tendencia. Con el apoyo del software R se realizaron histogramas que permitieran ver la posible existencia de asimetría en la distribución de los datos; se realizó un gráfico de caja que detalla datos atípicos y dispersogramas para ver la presencia de tendencia que pudieran tener los datos frente a sus coordenadas Este – Oeste (x) y sus coordenadas Norte – Sur (y); se encontraron las medidas de tendencia central para poder concluir acerca de ellas.

Una vez terminado el análisis univariado de los datos, en el siguiente nivel se hizo otro examen exploratorio de los datos, pero teniendo en cuenta ahora el aspecto espacial (aclarando que aún no se analizaba la dependencia espacial). Para esto se realizaron nuevas gráficas que permitieron tener evidencia de tendencia espacial, entre estas gráficas se obtuvieron: una de contornos en la región de estudio (departamento de Cundinamarca), la cual permitía ver zonas donde la variable presentara cambios considerablemente bruscos; otra gráfica de interpolación en 3D que

permitía observar la posible existencia de tendencia y hacia qué zonas se concentraba; también se obtuvo una gráfica del mapa de símbolos donde se podía ver qué tan extremos llegaban a ser los valores bajos, medios y altos de la variable en cada punto de la muestra que se tiene en la base de datos.

Modelado para la tendencia del proceso

Cuando se analizó la tendencia de los datos en su análisis exploratorio, se procedió a modelar esa tendencia: en caso de no cumplir las condiciones de estacionariedad, se hace necesario hacer una transformación Box-Cox del valor de la precipitación de los datos con el fin de volverlos más heterogéneos; en caso de mantenerse la tendencia, se debe modelar mediante una regresión lineal la tendencia de los datos. Una vez se obtienen los modelos lineales de la regresión, se hallan los datos residuales, se vuelve a hacer un análisis exploratorio a los datos residuales y se observa si la tendencia se ha podido eliminar.

Análisis estructural de la variable

Una vez concluido los datos con los que se hace el estudio (datos residuales), se continuó con el primer paso del análisis estructural (donde se estudia la dependencia espacial), en el cual, mediante el software R, se obtuvieron los semivariogramas empíricos de los datos, empleando los estimadores clásico y robusto. En este análisis se tuvieron en cuenta varios aspectos: ¿En qué valor de φ los semivariogramas alcanzaban su meseta $(\tau^2 + \sigma^2)$?, ¿qué distancia máxima era la apropiada para tener una buena gráfica de cada semivariograma?, ¿cuál de los dos entre el clásico y el robusto iba a poder ajustarse mejor a los modelos teóricos?

Posteriormente a la elección del mejor semivariograma empírico entre el clásico y el robusto, se procede a ajustar el mejor modelo teórico de dependencia espacial. En el software R se utilizó una función para realizar el ajuste eyefit de los parámetros, con la cual se pudieron obtener los parámetros a sentimiento, ajustando lo mejor posible el variograma empírico a los modelos teóricos; se ajustaron los modelos gaussiano, exponencial, esférico y Mattern. Después de ajustar cada uno de los modelos al ojo con eyefit, se aplicó la función likfit de R para encontrar por método de máxima verosimilitud los parámetros de cada modelo (los ajustes a sentimiento se hicieron para obtener unos parámetros iniciales a la hora de ajustar por ML), como esta función requería unos valores iniciales de los parámetros, se le ingresó a la función los que se ajustaron a sentimiento por eyefit. Después de obtener por ML los parámetros para cada modelo, se debía escoger el de mejor ajuste, para ello se usó la validación cruzada, la cual arrojaba el error de predicción kriging para cada modelo. El de menor error fue el modelo escogido y sus parámetros θ se deducen como la mejor información extraída de la realidad.

Simulaciones con Simulated Annealing Algorithm (SAA)

Una vez se obtuvo el modelo teórico que mejor ajustó al semivariograma empírico, se procedió a hacer simulaciones con el algoritmo Annealing. Para poder ejecutar este algoritmo de simulación era necesario primordialmente discretizar la región de estudio en una grilla fina, cuyos nodos iban a ser los puntos donde el algoritmo simularía los valores de la variable. Así el shape de Cundinamarca se cubrió con una grilla lo más fina posible (para el proyecto fue de 1000 x 1000) haciéndola lo más cercana al cubrimiento de la región, sin ir tampoco a un punto en el que computacionalmente el algoritmo demandara mucho tiempo de cálculo computacional.

Para usar el SAA se usó el paquete de R Intamapinteractive, el cual para poder ser ejecutado pedía especificar el modelo de dependencia espacial (modelo teórico de semivariograma), que fue encontrado en el análisis estructural. También solicitaba la cantidad de estaciones de muestreo que se quería eliminar o adicionar, y la cantidad de iteraciones que se quería ejecutar para alcanzar a optimizar la varianza kriging.

Se simularon entonces tamaños de muestra a partir de 71 estaciones, con un incremento de 5 en 5 hasta las 231 estaciones. Se tomaron estos tamaños de muestra, ya que se encuentran alrededor del tamaño de muestra inicial, que era de 151 datos; además, debido a la demanda computacional en tiempo que requiere el algoritmo, por cada tamaño de muestra se hicieron 3 simulaciones.

Cálculo de las matrices de covarianza de los parámetros estimados

Una vez se tuvieron las simulaciones SAA para cada tamaño de muestra (3 por cada tamaño de muestra), se hizo para cada simulación el análisis estructural para encontrar los mejores parámetros que se ajustaron al modelo de dependencia espacial escogido en el análisis estructural inicial, Nota: No se cambia el modelo inicial, solo se ajustan los parámetros, esto se hizo con cada simulación SAA y se tabuló esta información, es decir, para cada simulación se obtuvo un valor único para σ^2, τ^2 y φ.

Una vez se organizó la información, por cada tamaño de muestra (que tenía 3 valores por cada parámetro), se le calculó la matriz de covarianza de los parámetros estimados, la cual se pudo calcular de una manera sencilla con la expresión:

$$\begin{pmatrix} Var\,\sigma^2 & Cov\,\sigma^2\,\varphi & Cov\,\sigma^2\,\tau^2 \\ Cov\,\varphi\,\sigma^2 & Var\,\varphi & Cov\,\varphi\,\tau^2 \\ Cov\,\tau^2\,\sigma^2 & Cov\,\tau^2\,\varphi & Var\,\tau^2 \end{pmatrix} \quad (1)$$

Ya teniendo las matrices de covarianza de los parámetros estimados, se le calculó a cada una el logaritmo de su determinante, para la posterior comparación con las matrices de Fisher.

Cálculo de las matrices de Fisher

Teniendo ya las matrices de covarianza de los parámetros estimados, se necesitaba calcular las matrices de Fisher para comparar ambas matrices. El cálculo de las matrices de Fisher es dispendioso, por lo que se programó una rutina en R para hacer el cálculo de la matriz para cada tamaño de muestra.

De las simulaciones SAA para cada tamaño de muestra, se pudo obtener su matriz de distancias, es decir, el SAA daba un dataframe con las coordenadas y los valores de Z de las nuevas configuraciones simuladas, lo que permitió obtener la matriz de distancias para cada simulación.

Para calcular la matriz de Fisher se hicieron los cálculos sobre el modelo de dependencia espacial inicial (el semivariograma que mejor ajustó en el análisis estructural inicial de los 151 datos de la muestra obtenida). Debido a que de él se obtiene información de la realidad, se extrae el covariograma usando:

$$\gamma(h) = \sigma^2 - C(h) \tag{2}$$

Se derivó parcialmente $C(h)$ respecto a cada parámetro:

$$\frac{\delta C(h)}{\delta \sigma^2}, \frac{\delta C(h)}{\delta \varphi}, \frac{\delta C(h)}{\delta \tau^2} \tag{3}$$

Y con esta información y la matriz de distancias se pudieron calcular las matrices de Fisher para cada tamaño de muestra.

Una vez se tuvieron las matrices de Fisher por cada tamaño de muestra, se calcularon sus inversas (estimación de la matriz de covarianza de los parámetros ML óptimos de cada tamaño de muestra), y se calculó el logaritmo de sus determinantes para compararlos con los de las matrices de covarianza de los estimadores.

Comparación de las matrices de covarianza de las estimaciones ML y las de Fisher

Del proceso anterior se obtuvieron los logaritmos de las matrices de covarianza de los parámetros estimados $\hat{\theta}$ y de las matrices de Fisher, se graficaron los resultados en función de los tamaños de muestra, y se observaron los tamaños de muestra donde estos valores se asemejaban, para realizar un estudio posterior (errores de predicción) en estos puntos y extraer el diseño óptimo para estimar los parámetros de covarianza.

Otra comparación que se hizo fue que se calcularon las raíces de los elementos de las diagonales de la matriz de Fisher, y los RMSE de los parámetros estimados, los cuales son la raíz cuadrada de las desviaciones de los parámetros estimados a los iniciales:

$$RMSE = \sqrt{\frac{\left(\hat{\theta} - \theta_o\right)^2}{\theta_o}} \tag{4}$$

Los RMSE fueron importantes para evaluar la estimación de la varianza de los parámetros con la varianza obtenida de las raíces de las diagonales de las matrices de Fisher.

Error de predicción para selección del mejor diseño

Una vez se estableció cuál muestra fue la que mejor estimó los parámetros en sus simulaciones, se procedió a hacer comparaciones de sus errores de predicción. Esto se hace mediante la validación cruzada para cada simulación del tamaño de la muestra escogido, es decir, se tuvieron 3 errores de predicción para la muestra escogida que estimó bien los parámetros (los parámetros óptimos).

Una vez se hizo la validación del error de predicción de las 3 simulaciones de la muestra que optimizó la estimación de los parámetros, se tomó como diseño óptimo aquel que tuvo menor error de predicción y se dejó como diseño de red de muestreo espacial óptima que estima eficientemente los parámetros de la función de covarianza.

Estudio de la varianza de la predicción para la nueva red de muestreo espacial

Se procedió a elaborar un estudio de predicción para el diseño de red de muestreo espacial cuya estimación de parámetros fue optimizada, por lo que se comparó inicialmente con la red de muestreo inicial de tamaño de muestra de 151 estaciones. Por lo tanto, se utilizó el proceso de validación cruzada, para comparar los valores de varianza de predicción para ambas redes de muestreo y se elaboraron los mapas de varianza de predicción para visualizar el comportamiento de la varianza sobre la región. Para el proceso de elaboración de los mapas de varianza de predicción se tuvo en cuenta el modelo de regresión sobre los datos iniciales para remover la tendencia, por lo que se empleó kriging universal que permite la inclusión de dicho modelo. Además, se aplicó la transformación inversa a las predicciones para volver a la escala original de la variable. Se generaron 10.000 puntos en una grilla regular para realizar los mapas de varianza de predicción.

Teniendo en cuenta una posible optimización de la predicción, se efectuó una nueva simulación Annealing, la cual adicionaría las estaciones que habían sido removidas de la red inicial, de tal forma que se simulara la reubicación de estas estaciones en el departamento de Cundinamarca.

Inicialmente se utilizó como base la red de muestreo espacial cuya estimación de parámetros fue optimizada, de tal forma que utilizando la misma función de modelo de ajuste y los mejores parámetros para 151 estaciones como función objetivo de la simulación, se logró encontrar una nueva configuración de estaciones sobre la región, cuyo valor de validación cruzada fue evaluada con el valor obtenido de las 131 estaciones iniciales. Se realizó nuevamente la gráfica de varianzas de predicción para ambos conjunto de datos y se realizó una comparación. Por último se efectuó el mismo procedimiento para otro conjunto de estaciones iniciales, teniendo en cuenta que la estimación de sus parámetros no hubieran sido optimizada, para posteriormente definir si el uso de una red de muestreo con optimización en la estimación de sus parámetros ofrecería un punto de partida para disminuir la predicción de la variable en la región de estudio.

Figura 1. Ubicación de las estaciones pluviométricas en el departamento de Cundinamarca en el año 2007.

3. Resultados y discusión

Análisis exploratorio de la variable

Análisis sin aspecto espacial

Como primer paso en el desarrollo de la investigación, y teniendo una visión general de la información, se hizo necesario mostrar la distribución espacial de los sitios de muestreo en el departamento de Cundinamarca (figura 1) tomando las 151 estaciones con información de la pluviosidad anual para el año 2007. Se observó que los puntos se encuentran ubicados de manera irregular sobre la superficie y se encontró una acumulación de puntos sobre la zona central, con ausencia de puntos en las zonas oriental y occidental del departamento.

Dentro del análisis descriptivo y exploratorio de los datos de precipitación a nivel univariado y sin análisis espacial, se obtuvieron las medidas de tendencia central y gráficas exploratorias enfocadas a evaluar inicialmente la distribución de los datos (los resultados se presentan en la figura 2 y la tabla 1). Se aprecia una marcada asimetría en la distribución de los datos de precipitación, cuya concentración se manifiesta hacia los valores bajos y unos pocos valores altos de baja frecuencia, por consiguiente una elevada heterogeneidad de los valores de precipitación, con coeficientes de variabilidad relativa superiores al 30% y la presencia de datos atípicos, lo que no es lógico para un proceso estocástico estacionario. Por ende, fue necesario hacer una transformación Box-Cox, con valor de λ igual a -0.650599 y p-valor igual a 1.030941e-05, para resolver el proceso de heterocedasticidad.

Tabla 1. Medidas descriptivas de la pluviosidad

Medidas descriptivas	
Mínimo	418
Máximo	6922
Promedio	1443,93
Mediana	1075
Desviación estándar	966
Desviación mediana	331
Asimetría	2,31
Kurtosis	10,32
Coeficiente Var Promedio (%)	66,9
Coeficiente Var Mediana (%)	30,79

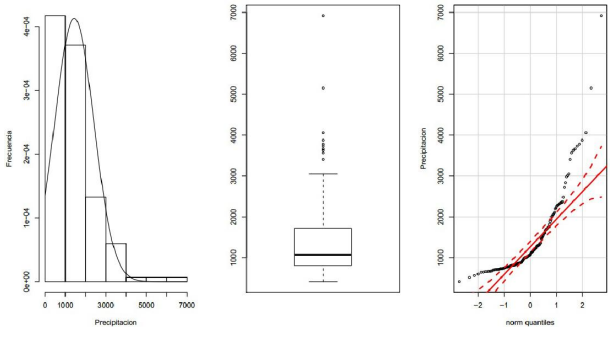

Figura 2. Histograma, diagrama de caja y gráfico de probabilidad normal de la pluviosidad.

Análisis exploratorio del aspecto espacial

Posteriormente, la inclusión del aspecto espacial evaluó la presencia de algún tipo de anomalía o tendencia espacial.

En la figura 3 se muestra el mapa de símbolos, el mapa de contornos, el mapa de interpolación y la superficie de interpolación de los valores originales de la pluviosidad; a partir de ellos se encontró la presencia de tendencia espacial en los datos que sugieren que la media no es constante en la región de estudio, y por lo tanto existe una alta variabilidad de los datos a nivel espacial.

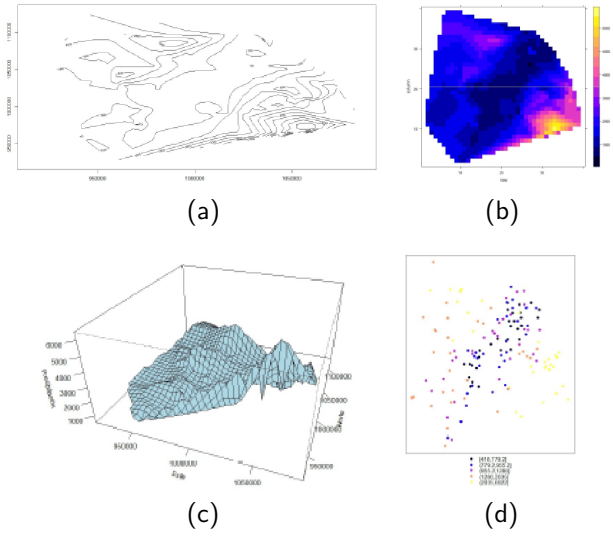

Figura 3. Análisis gráfico para la tendencia espacial.

Los resultados mostraron una tendencia en sentido suroeste a noroeste (figura 3), con presencia de niveles bajos en el centro de la cordillera, y altos al oriente (piedemonte Llanero) y occidente (valle del Magdalena) del departamento, donde las alturas sobre el nivel del mar son menores. Adicionalmente, se apreciaron tres concentraciones de valores altos con precipitaciones mayores a 3000 milímetros/metro cuadrado que parecen ser valores atípicos, pues en la mayor parte del departamento los valores de pluviosidad son bajos.

A pesar de que se utilizó la transformación Box-Cox para disminuir la distribución heterogénea de los datos, esto no redujo la tendencia de la variable. Por lo tanto, uno de los métodos utilizados para eliminar la tendencia fue el de utilizar los residuales, por lo que fue necesario utilizar un modelo de regresión polinómico de segundo orden, para lo cual se encontraron significativas las variables x, xy, x^2 y y^2, donde x es la coordenada Este–Oeste y y la coordenada Norte—Sur, con un R^2 igual a 0.4964 y un R2 ajustado igual a 0.4826.

Para los datos transformados y posterior a la aplicación del modelo de regresión, se comprobó que el proceso tiene un comportamiento más homogéneo en su distribución (figura 4), de igual forma se comprobó que la tendencia fue eliminada, tanto en perfiles Oriente–Occidente como Norte–Sur (figura 5), por lo tanto se decide trabajar sobre los residuales, ya que se ajustan más adecuadamente con el concepto de variable regionalizada utilizada en procesos geoestadísticos.

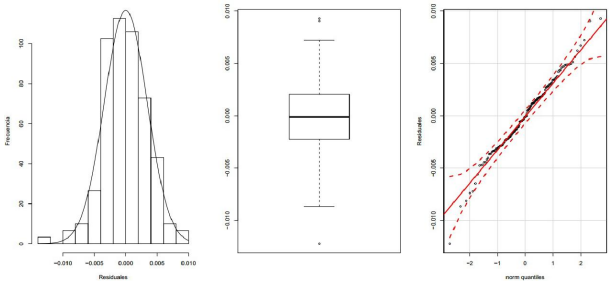

Figura 4. Análisis gráfico univariado.

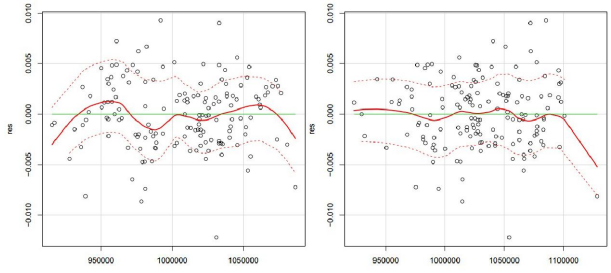

Figura 5. Diagrama de dispersión de los residuales.

Análisis estructural

Los residuales fueron el resultado de todo el análisis exploratorio. Estos datos fueron posteriormente objeto del proceso de elaboración del semivariograma empírico (experimental), se calcularon los semivariogramas robusto (figura 6a) y clásico (figura 6 b) mediante las funciones del paquete GeoR de R, el cual calcula los semivariogramas omnidireccionalmente. De estos, se decidió utilizar el semivariograma clásico, por tener una variabilidad más homogénea después de una distancia de 50.000 metros, donde se asume la meseta ya ha sido alcanzada.

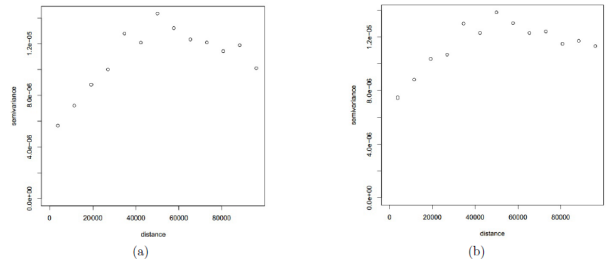

Figura 6. Semivariogramas empíricos.

Luego de la estimación del semivariograma experimental, se identificó el mejor modelo teórico de semivarianza que se ajustara al comportamiento de la estructura de correlación espacial de los datos, a través del proceso de validación cruzada, teniendo en cuenta los posibles mejores

modelos teóricos ajustados, entre los cuales se utilizaron el circular, el exponencial, el gaussiano, el Matern y el esférico (Figura 7). Se pudo detectar que el mejor modelo de ajuste de semivariograma al conjunto de residuales fue el semivariograma teórico gaussiano, con el menor error de validación cruzada igual a 0.01406277 después de haber ajustado por máxima verosimilitud (ML) todos los valores de los parámetros iniciales para cada modelo. Los parámetros de ajuste iniciales obtenidos con la ayuda de la función eyefit son presentados en la tabla 2.

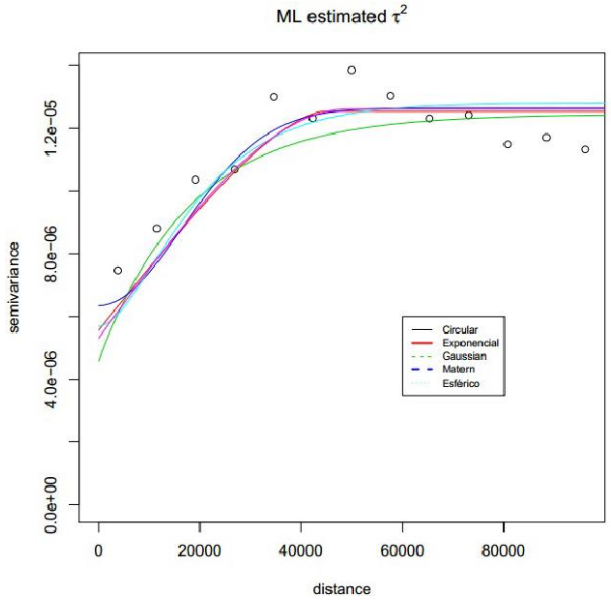

Figura 7. Modelos teóricos de semivarianza ajustados por ML.

Tabla 2. Parámetros de los modelos teóricos de semivarianza justados al semivariograma experimental.

Modelo	σ^2	φ	τ^2	k
Circular	6,8e-06	44176,86	5,8e-06	-
Exponencial	6,3e-06	18190,47	6e-06	-
Gaussiano	7,3e-06	23387,75	5,3e-06	-
Mattern	7,2e-06	10394,56	5,2e-06	1,49
Esférico	6,7e-06	49374,14	5,8e-06	-

Teniendo en cuenta que fue utilizado el método de máxima verosimilitud para el proceso de estimación de parámetros de ajuste, se detectaron los valores de los parámetros para el mejor modelo de ajuste, los cuales son presentados en la tabla 3.

Tabla 3. Valores de los parámetros del semivariograma gaussiano ajustado al semivariograma empírico clásico.

Parámetros	
σ^2	6,297134e-06
φ	23387,75
T	6,348605e-06

Resultados de SAA, matrices de covarianza de $\hat{\theta}$ y matrices inversas de Fisher

Los valores obtenidos para cada una de las tres estimaciones de los parámetros $\hat{\theta} = (\sigma^2, \varphi, \tau^2)$ de cada tamaño de muestra tuvieron un valor cercano. De igual manera, fueron obtenidas las matrices de covarianza de $\hat{\theta}$ y las matrices inversas de información de Fisher de manera satisfactoria para cada tamaño de muestra. La figura 8 presenta los logaritmos de los determinantes de las matrices inversas de información Fisher y los logaritmos de los determinantes de las matrices de covarianza de los de $\hat{\theta}$ estimados por ML para cada tamaño de muestra simulada.

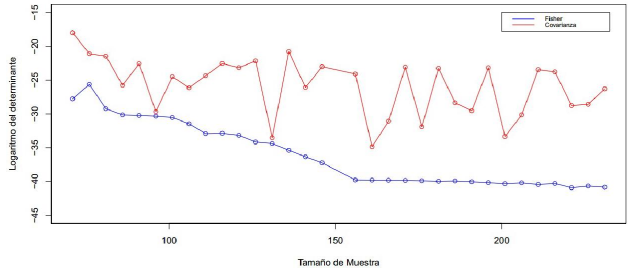

Figura 8. Logaritmos de las matrices de F y $\hat{\theta}$.

Los resultados mostraron que los logaritmos de los determinantes de la matriz de Fisher y la matriz de covarianza $\hat{\theta}$ para cada tamaño de muestra presentaron un comportamiento decreciente a medida que el tamaño de muestra aumenta. Esto indicaría una disminución en la variabilidad de los estimadores, sin embargo, la tendencia del comportamiento de la matriz de covarianza respecto a la matriz de Fisher es diferente, debido al lento decrecimiento que ésta presenta, lo que provoca que a medida que aumenta el tamaño de muestra, la diferencia entre logaritmos también aumenta, y, por lo tanto, la variabilidad de las estimaciones de los parámetros que representa la matriz de covarianza, los cuales representan la realidad, se alejarían de la variabilidad de los estimadores de la matriz de Fisher, que representa la variabilidad de los parámetros que deberían ser óptimos para cada tamaño de muestra.

Comparación de la información de Fisher y los RMSE de los $\hat{\theta}$

Una vez revisada la raíz cuadrada de los elementos diagonales de la matriz inversa de información de Fisher y los RMSE de las estimaciones ML calculados de las simulaciones realizadas por cada tamaño de muestra, se deduce que la matriz de información de Fisher da una estimación bastante precisa de la varianza de los estimadores ML. La aproximación es bastante buena para σ^2 y τ^2 en todos los tamaños de muestreo, y son peores para φ.

Las diferencias entre las varianzas del estimador σ^2 demostraron que a medida que el tamaño de muestra es menor, hay una mayor aproximación entre las varianzas de los estimadores procedentes de la matriz de Fisher y de los estimadores ML (figura 9). Para el caso del estimador φ se pudo observar que las diferencias soy muy pequeñas, mientras que para el estimador τ^2 sus diferencias son considerables. Esto puede deberse al hecho de que, tal como lo consideró Lark (2001), las matrices de covarianzas de los estimadores ML y las matrices de información de Fisher resultan de una combinación de varianzas y covarianzas de las variables con diferentes unidades.

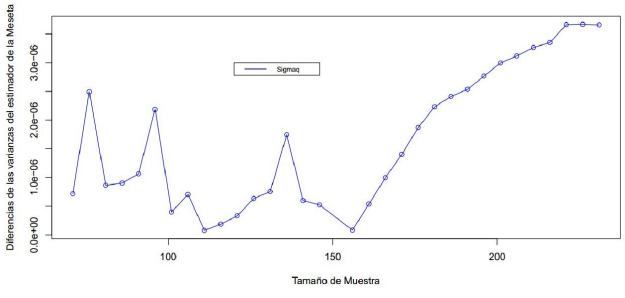

Figura 9. Diferencias de las varianzas de σ^2.

Selección de la red de muestreo espacial óptima

Aunque el estudio de simulación reveló que la matriz inversa de Fisher no es una buena aproximación de la matriz de covarianza de $\hat{\theta}$ por ML cuando el tamaño de muestra es pequeño, todavía puede ser utilizado como un criterio de diseño si la relación entre estos dos es monótona (Zhu y Stein, 2004). Para el presente caso y como lo demuestra la figura 10, si no hay una clara relación monótona entre los logaritmos de las matrices, si se observa que a medida que aumenta el número de la muestra, disminuye el valor de ambos logaritmos, esto quiere decir que aunque la matriz de Fisher para este caso no permitió encontrar un diseño que optimizara los parámetros, se puede tomar una decisión considerando el comportamiento de las diferencias entre los logaritmos de la matriz de información de Fisher y la de covarianza de $\hat{\theta}$ por ML.

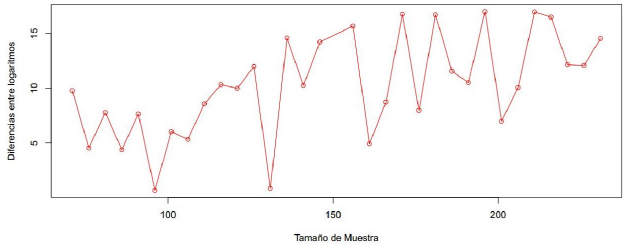

Figura 10. Diferencias entre los logaritmos de los determinantes de las matrices de Fisher y covarianza de los $\hat{\theta}$.

La figura 10 presenta las diferencias entre los logaritmos de las matrices, y revela que la diferencia entre los logaritmos es proporcional al tamaño de muestra, es decir, a medida que el tamaño de muestra aumenta, la diferencia entre ambos logaritmos también aumenta, por lo tanto fue conveniente tomar diseños de red que minimizaran esta diferencia y aproximaran las matrices inversas de Fisher y covarianza de $\hat{\theta}$.

Se evaluaron las diferencias entre los logaritmos de los determinantes para establecer los diseños de red en los que las matrices se aproximan, tales fueron los casos de los conjuntos de datos con tamaños de muestra de 96 y 131 estaciones (figura 10), cuyos valores de diferencia en sus criterios óptimos locales fueron de 0.6665453 y 0.85122658, respectivamente. Dado que la recomendación general de la OMM para redes de muestreo pluviométrico es de una estación cada 100 a 250 kilómetros cuadrados, se evaluó la conveniencia de utilizar ambas redes de muestreo en el departamento de Cundinamarca, que dio como resultado una densidad promedio por estación de 252 kilómetros cuadrados para el conjunto de datos de 96 estaciones, y de 185 kilómetros cuadrados para el conjunto de 131 estaciones, por lo que se hizo necesario solamente considerar el tamaño de muestra de 131 estaciones para el departamento.

Se evaluaron los diseños de las simulaciones (3 simulaciones) con tamaño de muestra de 131 estaciones, y utilizando el criterio de error de predicción por validación cruzada se encontró el mejor diseño, cuyo valor fue de 0.02097591, por lo tanto se eligió el conjunto de datos que minimizaron la varianza del predictor. Los parámetros de la nueva red son presentados en la tabla 7 y, finalmente, la red de muestreo espacial que optimizó los parámetros de la función de covarianza es presentada en la figura 11.

Tabla 4. Parámetros óptimos red de muestreo de 131 estaciones.

Parámetros óptimos	
σ^2	7,9e-07
φ	1,41
T	5,8e-12

Figura 11. Red de muestreo espacial que optimiza la estimación de los parámetros de la función de covarianza.

Estudio de la varianza de la predicción para la nueva red de muestreo espacial

A pesar de que para el presente proyecto no se tiene como objetivo la optimización de la predicción, es relevante evaluar este tipo de comportamientos para la nueva red.

Teniendo en cuenta que se obtuvo un valor para el error obtenido por validación cruzada de 0.02097591 para la nueva red de muestreo espacial de 131 estaciones (figura 12b), al ser comparado con el mismo error de la red existente de 151 estaciones (figura 12a) cuyo valor fue igual a 0.01406277, se pudo deducir que a pesar de tener una diferencia de 0.00691314 los valores de la varianza kriging para la superficie son muy cercanos en ambos casos, ya que su diferencia no es significativa.

Sin embargo, aunque fueron descartadas algunas estaciones originales, las cuales son removidas durante la optimización en los procesos de simulación, podría mejorarse la predicción siempre y cuando estas estaciones sean adecuadamente trasladadas de su sitio original a otro donde la varianza de su predicción podría mejorarse. Por lo tanto, se tomó como base el conjunto de datos de la red de muestreo de 131 estaciones obtenida del proceso de selección de la mejor red para estimar los parámetros de la covarianza, y se realizó una nueva simulación que incorporó las 20 estaciones que habían sido removidas. Se obtuvo finalmente un valor de error para la validación cruzada de 0.006892344, que supera el valor del error de varianza de la predicción del conjunto de estaciones iniciales. La figura 13 representa la disminución de la varianza de la predicción de la superficie para la red de muestreo espacial que optimiza la estimación de los parámetros de la covarianza (131 estaciones) y de la red de muestreo espacial que incorpora estaciones de muestreo, como si se supusiera el traslado de las estaciones meteorológicas que fueron descartadas durante el proceso de simulación (151 estaciones, con reubicación de 20 de ellas).

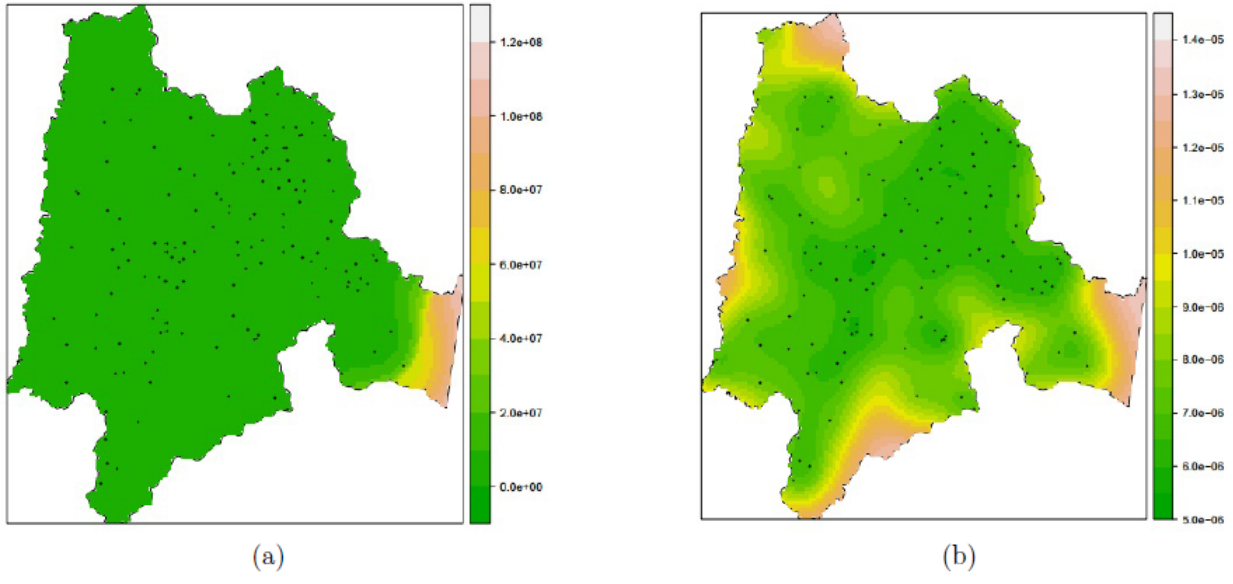

Figura12. Mapas de varianza de predicción.

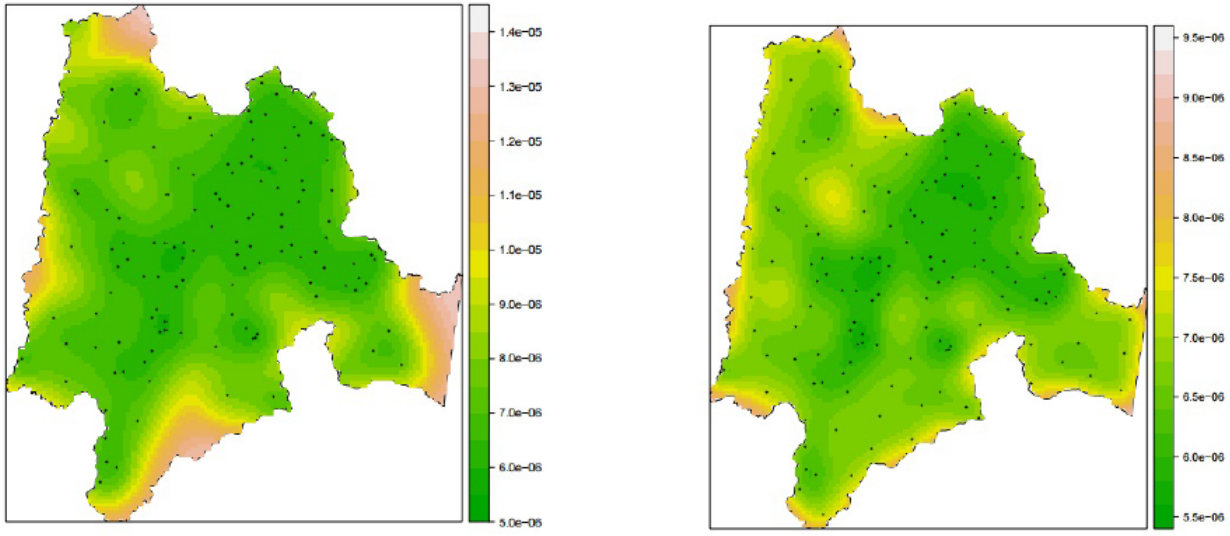

Figura13. Análisis de mapas de varianza para 131 estaciones y 151 estaciones con reubicación.

La nueva red de muestreo con reubicación de 20 estaciones que no cumplían con el proceso de optimización se muestra en la figura 14.

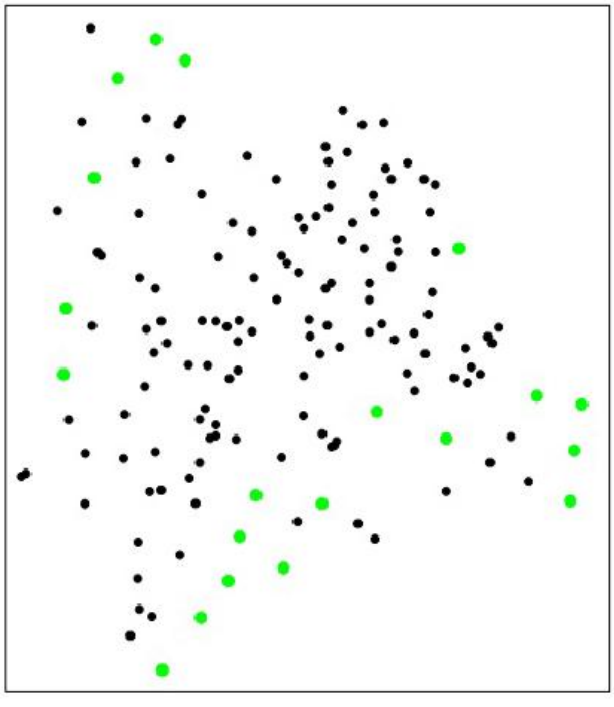

Figura 14. Red de muestreo de 151 estaciones con proceso de reubicación.

Se efectuaron nuevas simulaciones partiendo de otros tamaños de muestra cuya estimación de parámetros no fue óptima, y al realizar el respectivo proceso de inclusión de nuevas estaciones, la varianza de la predicción no disminuía sino que empeoraba la predicción, por lo tanto, debe tenerse en cuenta como punto de partida la adición de nuevas estaciones a partir de una red de muestro espacial cuya estimación de parámetros ha sido optimizada.

4. Conclusiones

Como metodología robusta para optimizar los parámetros del proceso espacial, el criterio de diseño de la matriz de información de Fisher, a pesar de no ser una buena aproximación a la matriz de covarianza de las estimaciones ML, puede ser usada como criterio si la relación entre estas matrices es monótona. Para el presente trabajo, esto facilitó evaluar la necesidad de adicionar o eliminar puntos de muestreo sobre una red existente, de tal forma que los criterios óptimos locales sean un indicador de la aproximación entre ambas matrices.

El algoritmo de simulación Annealing permitió encontrar una serie de resultados óptimos que cumplían con los requerimientos que el proyecto requería, teniendo en cuenta su criterio de optimización y las ventajas que el algoritmo proporcionaba, con lo cual se logró finalmente encontrar diseños óptimos que pudieron ser evaluados para obtener satisfactoriamente la matriz de covarianza de los estimadores ML y posteriormente ser comparado con la matriz de información; sin embargo, a pesar de que el algoritmo de simulación es eficaz, se encontró que en algunos procesos de simulación de datos no es eficiente computacionalmente.

A medida que aumentaba la cantidad de iteraciones en simulación se obtenía una correcta técnica de búsqueda aleatoria para resolver los problemas de optimización que

minimizan la varianza global kriging, de tal forma que el proceso posicionaba los puntos aleatorios sobre espacios sin muestrear de una manera que visualmente podría resultar lógica, con lo que se obtenía un diseño armonioso, ya que cubría la región de estudio de manera efectiva.

Al realizarse un estudio de la varianza de la predicción de una red de muestreo, se pudo concluir que ésta podría mejorar siempre y cuando se inicie con una red de muestreo cuya estimación de parámetros ha sido optimizada adecuadamente, de tal forma que la nueva red con adición de puntos minimice la varianza de predicción y mejore las mediciones de la variable en la región.

Referencias

BAUME, O., MELLES, SJ. y SKOIEN, J. Package 'intamapInteractive'. Software R. 1.1-7 Version. INTAMAP. April 19 2003.

BERTSIMAS, Dimitris y TSITSIKLIS, Jhon. Simulated Annealing. Statistical Science. Vol. 8, No. 1, Report from the Committee on Applied and Theoretical, Statistics of the National Research Council on Probability and Algorithms. Feb. 1993.

CRESSIE. Noel. Statistics for Spatial Data. Wiley. 1991.

GIRALDO, Ramón. Definiciones Básicas de Geoestadística. Introducción a la geoestadística, Teoría y Aplicación. Universidad Nacional de Colombia. Bogotá. 17 p.

MARDIA K.V., MARSHALL R.J. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika (1984). 135-46 p.

MYUNG, Jay y NAVARRO, Daniel. Matrix Information. Department of Psychology. Ohio State University. Mar 11, 2004.

SAMPER, F.J. y CARRERA, J. Geoestadística: Definición y Alcance. En: Geoestadística: Aplicaciones a la Hidrogeología. edit. Centro Internacional de Métodos Numéricos en Ingeniería, Universidad Politécnica de Catalunya, Barcelona, España, 1990. 1 p.

Zhu, Z. y Stein, M. (2004). Spatial sampling design for parameter estimation of the covariance function. University of North Carolina at Chapel Hill and University of Chicago, Chicago. Available online: www.sciensedirect.com.