Melo y Mora 2022. La geoestadística: la estadística aplicada a las ciencias de la tierra. Revista Sigma, 18 (1). Páginas 15–22.

REVISTA SIGMA

Departamento de Matemáticas y Estadística

Universidad de Nariño

Volumen XVIII Nº1(2022), páginas 15–22

La geoestadística: la estadística aplicada a las ciencias de la tierra

Arturo Melo [1]
Brayan Mora [2]

Abstract: Geostatistics or statistics applied to Earth sciences, it is a branch of statistics used as fundamental tools to treat, describe, analyze, and infer data obtained from spatial or temporal phenomena. Geostatistics has origin in the mining and oil industries where it has been successfully applied to solve geolocation problems, which involve decisions about exploitation, exploration and drilling operations that are highly costly economically, socially, and environmentally. Geostatistics has been extended to other fields of application such as hydrology, meteorology, oceanography, geography, geosciences, forestry, ecology, and others. This work presents, after a short review of the literature, what Geostatistics is, its theoretical and applied approaches. In this sense, it is used two cases to explain basic concepts and their applications.

Keywords: geostatistics, geosciences, applied statistics, spatio-temporal data.

Resumen: La Geoestadística o estadística aplicada a las ciencias de la tierra es una rama de la estadística usada como herramienta fundamental para tratar, describir, analizar e inferir sobre datos obtenidos de fenómenos de tipo espaciales o temporales. La Geoestadística tiene su origen en las industrias minera y petrolera donde ha sido aplicada de manera satisfactoria para resolver problemas de geolocalización, las cuales involucran decisiones sobre operaciones de explotación, exploración y perforación que son altamente costosas económica, social y ambientalmente. La Geoestadística se ha extendido a otros campos de aplicación como la, meteorología, geografía, geociencias, silvicultura, ecología y otros. Este trabajo presenta, después de una corta revisión de la literatura, que es la Geoestadística, su aproximación aplicada, los conceptos base y dos casos de uso que retoman los elementos planteados e ilustran su aplicación.

Palabras Clave: geoestadística, geociencias, estadística aplicada, datos espacio temporales.

[1]Ingeniero en electrónica. Investigador, estudiante de la Maestría en Estadística Aplicada, Universidad de Nariño. Pasto-Colombia. Correo electrónico: hamelor@unal.edu.co

[2]Estadístico del CIAT-Colombia, estudiante de la Maestría en Estadística Aplicada, Universidad de Nariño. Pasto-Colombia. Correo electrónico: b.mora@cgiar.org

1. Introducción

La geoestadística es una rama de la estadística originada en actividades mineras y petroleras, esta área de conocimiento tiene como objetivo analizar y modelar datos espacio-temporales (donde existen coordenadas de ubicación) con el fin de encontrar patrones pronosticar e interpolar datos faltantes, pero en especial encontrar la mejor distribución de probabilidad de una propiedad que caracterice cualquier espacio geolocalizado, minimizando la incertidumbre.

La geoestadística es usada en diferentes áreas y sus aplicaciones se pueden observar en la minería e industria petrolera, en la ciencias del medioambiente, en el estudio de suelos para la agricultura, silvicultura, el estudio del clima, logística, llegando a la epidemiologia, e incluso a caracterizar fenómenos sociales. En su desarrollo se ha fundamentado en principios estadísticos y probabilísticos para construir modelos de estimación, así como el uso intensivo de la simulación in silico para predecir con alta precisión los mapas o superficies optimas que dan respuesta a los estudios requeridos.

Este documento es una revisión sucinta acerca de la geoestadística, por tanto continúa con el marco teórico acerca de la materia, la relación de la estadística con la geoestadística, las metodologías de predicción (estimación y simulación), algunos términos importantes de la geoestadística, dos casos de uso y las conclusiones.

2. ¿Qué es la geoestadística?

La Geoestadística es la rama de la estadística que se enfoca en analizar y modelar datos espaciales o espacio-temporales, en otras palabras la variabilidad espacial que tienen los datos que referencian localizaciones mediante coordenadas. La Geoestadística fue desarrollada originalmente para describir patrones e interpolar valores de localizaciones donde no se tomaron muestras, pero especialmente para predecir distribuciones de probabilidad, así como determinar la incertidumbre para operaciones mineras y petroleras, de tal manera que se pudieran tomar decisiones informadas acerca de los posibles valores espacio-temporales que se interpolaban o se predecían por medio de los modelos construidos para tal fin ([5], [9]).

La geoestadística es ampliamente usada en muchas áreas de la ciencia, la ingeniería ([9], [6]), entre las que podemos destacar:

La industria minera, donde la geoestadística se usa para cuantificar los recursos minerales que puede tener un nuevo yacimiento, de tal manera que permita evaluar la factibilidad de los diferentes proyectos. También es importante en las de las operaciones de transporte de material sobre cómo generar estrategias logísticas acerca del flujo de materiales hacia las plantas de procesamiento o almacenamiento.

En las ciencias de medioambiente, la geoestadística es usada para estimar los niveles de contaminación del aire, del agua, con el fin de proponer soluciones a este tipo de problemas, pero en, especial si se hace necesario tomar decisiones acerca de la aplicación de conductas y tratamientos medioambientales que disminuyen las concentraciones de contaminación, o en casos más críticos tratar problemas de salud en poblaciones afectadas, para remediar o mejorar estas situaciones.

En el campo de las ciencias de suelo, la geoestadística se encarga de mapear la cantidad y la calidad de los nutrientes que contienen los terrenos estudiados ,entre ellos: nitrógeno, fósforo,

potasio, etc., y otros indicadores como humedad y conductividad del suelo, con el propósito de estudiar la relación de estas variables con la productividad por unidad de área de ciertos cultivos, como maíz, soya, arroz, caña de azúcar, café y muchos más, y así determinar con precisión las cantidades de fertilizantes necesarias para la producción agrícola en espacios geolocalizados.

Otra de las aplicaciones de la geoestadística es la meteorología, donde el foco de estudio incluye las predicciones climáticas, temperaturas, concentraciones de lluvia, acidez de la lluvia y demás, que de hecho en un estudio completo de una zona o región complementa los estudios de las ciencias del suelo.

Dentro de las aplicaciones más recientes de la geoestadística se pueden verificar el aporte al desarrollo de política pública en salud (epidemiología), como ejemplo como cierto tipo y niveles de contaminantes medioambientales están relacionadas con enfermedades respiratorias, hasta con las tasas de incidencia de cáncer. También se ven aplicaciones que aportan soluciones de gran valor en campos como el marketing, comercio internacional, logística, planeación militar, entre otros. Es conveniente decir que los algoritmos geoestadísticos actualmente están incorporados en sistemas de información geográfica (SIG) como en el programa ArcGIS y en paquetes de software estadístico como R.

3. Relación estadística-geoestadística

La estadística centra sus esfuerzos en analizar la variabilidad de los datos con el propósito de obtener correlaciones, dependencias y patrones con los cuales se logre explicar fenómenos tanto físicos como naturales. En cambio, la geoestadística se centra en el análisis y la modelación de variables asociadas a información espacial. Por lo que se puede decir que la estadística aporta todas sus técnicas y son aplicadas a datos georeferenciados o con información de longitud y latitud asociada a los datos. Inicialmente la geoestadística estaba íntimamente relacionada a los métodos de interpolación, pero actualmente las técnicas geoestadísticas se basan en modelos estadísticos para modelar incertidumbre asociada con la estimaciones espaciales y simulaciones ([1], [3]).

Se sabe que el objetivo principal de la geoestadística es predecir la posible distribución espacio-temporal de una o varias propiedades específicas en una región determinada, donde esas predicciones pueden tomar forma de mapa o serie de mapas, los cuales se discriminan e interpretan de acuerdo a su correlación espacial respecto a las distribuciones de las propiedades que se quiere evaluar. En sus inicios la geoestadística se asociaba con los métodos de interpolación, pero hoy por hoy, su desarrollo ha permitido que sobrepase este tipo de problemas simples, de tal manera que las técnicas geoestadísticas usan modelos basados en funciones o variables aleatorias para modelar la incertidumbre asociada a la estimación y simulación de fenómenos espacio-temporales [9].

El fenómeno de datos de localización desconocidos o perdidos es abordado como un conjunto de variables aleatorias correlacionadas de la siguiente manera:

En primera instancia, sea $Z(x)$ el valor de una variable de interés en una localización x el cual a su vez es desconocido (temperatura, rata de lluvia, acidez del suelo, y otros), el cual se considera un valor aleatorio hasta que no sea medido. Adicionalmente la aleatoriedad de $Z(x)$, aunque es definida por la función de distribución acumulada (FDA) depende de cierta información conocida $Z(x) : F(z, x) = Prob\{Z(x) \leq z|$ información $\}$. Generalmente, el valor de Z es conocido en localizaciones cercanas a x, por lo tanto, se puede asumir continuidad espacial, de tal manera que $Z(x)$ puede tomar valores similares a los encontrados en su

entorno cercano o vecindad. En segunda instancia, el modelo de continuidad espacial de una variable aleatoria puede ser una función paramétrica si se usa la varianza de las diferencias de los puntos (variograma) o no paramétrica si se usa otros métodos como simulación de puntos múltiples, técnicas genéticas y otros [8].

El modelamiento para la predicción o estimación de distribuciones para datos geolocalizados, según el contexto aquí presentado tiene dos propósitos: 1) Estimar el valor de $Z(x)$, por medio de media, la mediana o la moda de la FDA $F(z, x)$, y 2) tomando muestras de la totalidad de la función de densidad $f(z, x)$, considerando cada coordenada y su posible resultado, procesos que genera diferentes mapas de $Z(x)$ denominados realizaciones. Recordar que los datos geoespaciales son discretos que forman grillas o mallas, entonces cada realización es una muestra de la función de distribución conjunta N-dimensional:

$$F(z, x) = Prob\{Z(x_1) \leq z_1, Z(x_2) \leq z_2, \ldots, Z(x_N) \leq z_N\}[8].$$

4. Métodos de predicción de la geoestadística

Se ha planteado que la geoestadística, busca predecir las distribuciones espaciales de una propiedad, predicción que toma forma de mapa o realización. En aras de llevar a cabo esa predicción, se hace por medio de la estimación espacial y la simulación [9].

La estimación tiene como objetivo producir una realización o mapa que sea estadísticamente el más optimo posible. Dentro de la estimación, los dos métodos más representativos documentados son:

Método de Kriging: Es un método de interpolación, para el cual los valores interpolados son modelados por un proceso normal o gaussiano, donde las covarianzas previas son los determinantes. Este mecanismo brinda la mejor predicción lineal insesgada para los valores intermedios. En contraste, aquellos métodos que usan criterios de suavización como los splines, no gneran los valores intermedios con mayor probabilidad. El método Kriging es ampliamente utilizado en el dominio del análisis espacial y experimentos informáticos, además puede ser escalado a problemas extensos [11].

Inferencia Bayesiana: Este método utiliza el teorema de Bayes para actualizar un modelo de probabilidad a medida que se dispone más pruebas o información. En este caso la Inferencia Bayesiana implementa el método de Kriging a través de un proceso espacial gaussiano y lo actualiza utilizando el teorema de Bayes para calcular su valor posterior[3].

La simulación es el proceso de replicar la realidad usando un modelo in silico. En geoestadística la simulación es encontrar una realización, mapa o superficie fundamentada en una función aleatoria que tiene las mismas características de los datos de la muestra usados para generar dicha realización. La simulación geoestadística gaussiana (SGG) es muy usada para datos continuos y se asume que los datos provienen de una distribución normal, para esto se supone que los datos son estacionarios, es decir que la media, la varianza y la estructura espacial (semivariograma) no cambian en el dominio espacial de los datos. Otro supuesto clave es que la función aleatoria que se modela es una función aleatoria gaussiana multivariante.

La SGG ofrece mejores estimaciones que el método Kriging debido a que este se basa en un promedio local de los datos y produce una salida suavizada, por lo tanto pierde variabilidad global. Adicionalmente, la SGG produce mejores representaciones ya que conserva la variabilidad local de tal manera que es bien aceptada en la industria del petróleo como método para caracterizar yacimientos heterogéneos. La simulación tiene preferencia con

respecto a los enfoques de interpolación tradicionales, en parte porque captura el carácter heterogéneo y proporciona estimaciones de reservas de hidrocarburos mas precisas. Su enfoque estocástico permite el cálculo de muchas soluciones igualmente probables que pueden procesarse posteriormente para cuantificar y evaluar incertidumbre ([8], [13]).

Como información, existen varios métodos de simulación geoestadística, tales como:

1. Agregación.
2. Desagregación.
3. Descomposición de Cholesky.
4. Gaussiana truncada.
5. Plurigausiano.
6. Simulación espectral.
7. Indicador secuencial.
8. Gaussiana secuencial.
9. Probabilidades de transición.
10. Geoestadística de la cadena de Markov.
11. Modelos de malla de Markov.
12. Maquina de vectores de soporte.
13. Simulación boolena.
14. Modelos genéticos.
15. Modelos pseudogenéticos.
16. Autómatas celulares.
17. Geoestadística de puntos múltiples.
18. Y otras.

5. Elementos y conceptos de la geoestadística

Teoría de variables regionalizadas (TVR): Esta teoría propone que la interpolación desde puntos en el espacio debe basarse en un modelo estocástico que tenga en cuenta las diversas tendencias en el conjunto original de puntos ([8], [10]). La teoría considera que dentro de cualquier conjunto de datos se pueden detectar tres tipos de relaciones:

- Parte estructural, que también se denomina tendencia.
- Variación correlacionada.
- Variación no correlacionada, o ruido.

- Después de definir las tres relaciones anteriores, TVR aplica la primera ley de la geografía respecto a la conservacion de las caracteristicas climáticas, morfológicas, biológicas, etc., de los espacios o territorios adyacentes para predecir los valores desconocidos de los puntos. La principal aplicación de esta teoría es el método Kriging para la interpolación.

Función de covarianza: La covarianza es una medida de cuánto cambian dos variables juntas, y la función de covarianza, o kernel, describe la covarianza espacial o temporal de un proceso o campo de variable aleatoria.

Semivariograma empírico: Las funciones de semivariograma y covarianza son cantidades teóricas que no se puede observar, por lo que se estiman a partir de sus datos usando lo que se llama semivariograma empírico y funciones de covarianza empíricas. A menudo, se puede obtener información sobre las cantidades observando la forma en que se estiman. Suponga que toma todos los pares de datos que tienen una distancia y dirección similares entre sí [10].

Variograma: Se define como la varianza de la diferencia entre los valores de campo en dos ubicaciones a lo largo de las realizaciones del campo [4].

Imagen de entrenamiento: Las imágenes de entrenamiento, son un conjunto de datos que configuran una superficie, la cual puede ser tomada como un parámetro de modelado importante en las geoestadísticas multipunto, determinan directamente el efecto del modelado. Es necesario evaluar y seleccionar la imagen de entrenamiento candidata antes de usar el modelado geoestadístico multipunto. La probabilidad de repetición general no es suficiente para describir la relación de eventos de datos únicos en la imagen de entrenamiento [8].

6. Casos de uso

Caso 1: Estadísticas espaciales y mapeo de suelos: una asociación floreciente bajo presión [4].

Durante la mayor parte del siglo 20, los podólogos mapearon el suelo trazando limites entre diferentes clases de suelo que identificaron a partir de estudios a pie o en vehículos, complementados con interpretación de fotografías aéreas y respaldados por una comprensión del paisaje y surge la necesidad de predecir las condiciones del suelo en sitios no visitados se hizo evidente y en la década de 1980 la introducción a la geoestadística y específicamente el método de Kriging ordinario que revolucionó el pensamiento geoestadístico y en gran medida la práctica.

El kriging ordinario se basa únicamente en datos de muestra de la variable de interés; no tiene en cuenta las covariables relacionadas. Estos últimos fueron incorporados a partir de la década de 1990 como efectos fijos e incorporados como predictores de regresión, dando lugar al kriging con deriva externa y al kriging de regresión. La estimación simultánea de los coeficientes de regresión y los parámetros del variograma se realiza mejor mediante la estimación de máxima verosimilitud residual.

En los últimos años, el aprendizaje automático se ha vuelto factible para predecir las condiciones del suelo a partir de grandes conjuntos de datos ambientales obtenidos de sensores a bordo de satélites y otras fuentes para producir mapas digitales de suelos. Las técnicas se basan en la clasificación y la regresión, pero no tienen en cuenta las correlaciones espaciales. Además, son efectivamente çajas negras"; carecen de transparencia y su producción debe

validarse si se quiere confiar en ellos. Sin duda tienen mérito; Están aquí para quedarse. Sin embargo, también tienen sus deficiencias cuando se aplican a los datos espaciales, que los estadísticos espaciales pueden ayudar a superar.

Los estadísticos espaciales y los podometristas todavía tienen mucho que hacer para incorporar la incertidumbre en las predicciones digitales, los promedios espaciales y los totales de las regiones, y para tener en cuenta los errores de medición y las posiciones espaciales de los datos de muestra. También deben comunicar su comprensión de estas incertidumbres a los usuarios finales de los mapas de suelos, cualquiera que sea el medio por el que se hayan hecho.

Caso 2: Heterogeneidad espacial de las propiedades químicas del suelo en un bosque húmedo tropical de tierras bajas, Panamá [12].

Se evaluó la heterogeneidad espacial para el pH y un conjunto integral de nutrientes y oligoelementos en suelos superficiales (0–0, $1m$ de profundidad) y subsuperficiales (0, 3–0, $4m$ de profundidad) en 26, 6 ha de bosque húmedo tropical antiguo de tierras bajas, establecido en un suelo altamente meteorizado en Panamá.

Poco se sabe acerca de los patrones de heterogeneidad espacial de las propiedades del suelo en los suelos de los bosques tropicales. El suelo era moderadamente ácido (pH 5,28) con bajas concentraciones de cationes básicos intercambiables (13, 4 cmolc/kg), PO4 extraíble con Bray (2, 2 mg/kg), NO3 extraíble con KCl (5, 0 mg/kg) y NH4 extraíble con KCl (15, 5 mg/kg). El coeficiente de variación de las propiedades del suelo osciló entre 24 % y $>$ 200 %, con un valor medio de 84 %.

El análisis geoestadístico reveló dependencia espacial a una escala de 10 – 100 m para la mayoría de las propiedades del suelo; sin embargo, pH, NH4, Al y B tuvieron dependencia espacial a una escala de hasta 350 m. Los modelos de mejor ajuste para variogramas individuales incluyeron funciones aleatorias, exponenciales, esféricas, gaussianas, lineales y de potencia, lo que indica muchos patrones espaciales diferentes entre el conjunto de propiedades del suelo. La correlación entre los elementos individuales fue pobre, lo que indica patrones independientes. Los resultados muestran patrones espaciales complejos en las propiedades químicas del suelo y proporcionaron una base para futuras investigaciones sobre las relaciones suelo-planta y la diferenciación de nichos de nutrientes del suelo.

7. Conclusiones

El objetivo de la geoestadística es predecir la posible distribución espacial de una propiedad (por ejemplo el grado de contaminación, o la cantidad de nutrientes de un suelo y demás) de una región, subregión, o de un espacio claramente localizado. Tal predicción a menudo toma la forma de un mapa o una serie de mapas denominadas realizaciones. En la geoestadística se destacan dos formas básicas de predicción, ellas son la estimación y simulación. En la estimación, se produce una única realización o superficie estadísticamente óptima (mapa) de la ocurrencia espacial estudiada. La estimación se basa tanto en los datos de la muestra como en un modelo preestablecido (variograma) determinado como la representación más precisa de la correlación espacial de los datos de la muestra. Esta estimación única o mapa generalmente se produce mediante la técnica de distribución gaussiana llamada Kriging y brinda la mejor aproximación lineal insesgada. Por otro lado, en la simulación, se produce uno o muchos mapas de igual probabilidad (a veces llamados imágenes, superficies o realizaciones) de la distribución de propiedades que se analiza, adicionalmente, la simulación conserva la varianza local, por lo cual el método es más preciso que Kringing. Finalmente, las diferencias

entre los mapas alternativos, de las simulaciones, proporcionan una medida para cuantificar la incertidumbre, una opción que no está disponible con la estimación de Kriging.

Referencias

[1] Asociación Geoinnova. (2019). *¿Qué es la geoestadística y cuáles son los principales análisis geoestadísticos?* https://geoinnova.org/blog-territorio/que-es-la-geoestadistica-analisis-geoestadisticos/ 17

[2] Bachmaier, M. & Backes, M. (2008). Variogram or semivariogram? Understanding the variances in a variogram. *Precision Agriculture, 9(3)*, 173–175. https://doi.org/10.1007/s11119-008-9056-2

[3] Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CHAPMAN & HALL/CRC. 17, 18

[4] Cressie, N. A. C. (1993). *Statistics for Spatial Data*. JOHN WILEY & SONS, INC. 20

[5] Krige, D. G. (1951). Journal of the Chemical. Metallurgical and Mining Society of South Africa, 52, 119. 16

[6] ESRI. (2022). *What Is Geostatistics?* https://pro.arcgis.com/es/pro-app/2.8/help/analysis/geostatistical-analyst/what-is-geostatistics-.htm 16

[7] Heuvelink, G. B. M. & Webster, R. (2022). Spatial statistics and soil mapping: A blossoming partnership under pressure. *Spatial Statistics*, 100639. https://doi.org/10.1016/j.spasta.2022.100639

[8] Mariethoz, G. & Caers, J. (2015). *Multiple-point Geostatistics. Stochastic Modeling with Training Images*. Wiley-Blackwell. 18, 19, 20

[9] Shaltami, O. R., Fares, F. F., Errishi, H. & Oshebi, F. M. El. (2021). Geostatistics- A Review. *Virtual Conference on Natural Gas Palais Royal , Yekaterinburg , Russia, 10 January 2021.* https://www.researchgate.net/publication/348406123_Geostatistics_-_A_review 16, 17, 18

[10] Wackernagel, H. (1995). *Multivariate Geostatistics. An Introduction with Applications.* Springer Berlin Heidelberg GmbH. 19, 20

[11] Wahba, G. (1990). *Spline Models for Observational Data.* SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS. https://doi.org/10.2307/1269578 18

[12] Yavitt, J. B., Harms, K. E., Garcia, M. N., Wright, S. J., He, F. & Mirabello, M. J. (2009). Spatial heterogeneity of soil chemical properties in a lowland tropical moist forest, Panama. *Australian Journal of Soil Research*, 47(7), 674–687. https://doi.org/10.1071/SR08258 21

[13] Ziegel, E. R. (2000). [Review of Geostatistics and Petroleum Geology, by M. E. Hohn]. Technometrics, 42(4), 444–444. https://doi.org/10.2307/1270983 19