Universidad Nacional Autónoma de Honduras

Facultad de Ciencias

Escuela de Matemáticas y Ciencias de la Computación

Carrera de Matemática

Técnicas de Muestreo Espacial: Aplicaciones e Implementación en R

Estudiante:
José Moisés Arias Núñez
20141003876

Catedrático:
M. Sc. Roberto Duarte

Presentado en: Tegucigalpa, M.D.C. al 29 de abril de 2024

Índice

Introducción

Los métodos estadísticos para el análisis de datos geográficamente distribuidos han probado ser de gran importancia en temas de valor social y científico desde la segunda mitad de los años 80 y mostraron un crecimiento rápido en popularidad y desarrollo durante la década subsecuente.

Las principales aplicaciones de importancia para estos métodos aparecen en el muestreo agrícola y forestal, pero también han mantenido su relevancia en la búsqueda de recursos energéticos fósiles y recursos minerales. El estudio de la contaminación ambiental y la predicción del clima se han visto fortalecidos gracias a la implementación del muestreo espacial como lo hacen notar (Arlinghaus, 1996) y (Hohn, Liebhold, y Gribko, 1993).

Existe una variedad de metodologías utilizadas para el análisis de datos geográficos, pero tal vez la más popular en la literatura científica y académica es el *kriging*. Esta técnica o, mejor dicho, colección de técnicas prueba ser sumamente versátil en la predicción de variables aleatorias espacialmente variadas. Posee diferentes versiones que se acoplan a distintas condiciones de variabilidad de los datos otorgando una reducción de la varianza en la predicción o estimación de la variable de interés.

El método de *kriging* o métodos de kriging surgen de un modelo básico de minimizar la varianza entre las observaciones regionales muestreadas y el predictor en la región de interés, bajo la condición de proveer un predictor insesgado. En este sentido, el mismo modelo del método ofrece una criterio para construir diseños muestrales. Pues tal como lo señala (Cressie, 1986) los resultados pueden ser independientes del uso de información muestral o datos de un pilotaje.

La implementación de la estadística espacial está dotada de un arsenal bastante completo de métodos en `R` para extraer variogramas y tendencias de los datos, así como representaciones visuales útiles de los datos recolectados en un experimento. Paquetes como `geoR` cuentan con una documentación detallada en la comunidad en línea así como de textos diversos que ilustran la fortaleza de sus métodos.

Este trabajo busca sintetizar los puntos más fuertes de estas técnicas y servir de punto de partida para llevar el estudio de la geoestadística a un nivel más profundo de comprensión y técnica.

Objetivos

- **Objetivo General**
  Ampliar los conocimientos sobre las técnicas de muestreo de una población en el contexto de unidades observacionales distribuidas en un espacio o región geográfica.

- **Objetivos específicos**

  1. Integrar los modelos de muestreo básicos en el contexto de una muestra espacialmente variable.
  2. Conocer las aplicaciones del muestreo espacial.
  3. Implementar los métodos de muestreo espacial en R.

1. Breve Reseña Histórica

La aplicación de la estadística para recoger y analizar información sobre regiones geográficas ha cobrado una mayor importancia con el paso de los años. Instancias particulares de esto se dan en la agricultura(Benedetti, Piersimoni, y Postiglione, 2015), la geología(Thompson, 2012), la meteorología(Cressie, 1986) y la ecología(Brus, 2022).

De acuerdo con (Arlinghaus, 1996), las técnicas de muestreo espacial pueden rastrearse a fechas tan tempranas como las de la revolución industrial. El análisis de decaimiento a distancia, problemas en cartografía y biología datan desde inicios del siglo pasado, según la literatura científica resumida por (Arlinghaus, 1996).

La importancia de los métodos de muestreo espacial va más allá de la academia, pues los estudios de este tipo permiten mejorar políticas sobre el manejo de recursos agrícolas y forestales(Benedetti y cols., 2015), predecir la existencia de recursos minerales y fósiles energéticamente explotables(Thompson, 2012; Cressie, 1986; Thadani, Alabert, y Journel, 1987), monitorear la distribución de especímenes en una región(Brus, 2022), rastrear la propagación de una enfermedad o infección sobre los habitantes de una localidad, estimar la distribución de comunidades en una zona(Benedetti y cols., 2015), etc.

A los métodos estadísticos empleados para la agricultura se les ha conocido globalmente como *agricultura estadística* y aborda estadísticas sobre productos agrícolas, pesca, ganado, seguridad alimenticia y forestación(Benedetti y cols., 2015). Los métodos empleados en el estudio de los suelos y la búsqueda de recursos minerales y fósiles se conocen conjuntamente como *geoestadística*(Thompson, 2012).

En particular, la geoestadística fue desarrollada para analizar datos geológicos distribuidos en el espacio de un cuerpo mineral con la intención de predecir reservas minerales(Cressie, 1986). Sus principales proponentes, de acuerdo con (Cressie, 1986), habrían sido Matheron, Whittle y Gandin.

El método que interesa en este ensayo es el conocido como *kriging*, una técnica desarrollada por Matheron(Benedetti y cols., 2015; Cressie, 1986) nombrada así por su atribución a Krige(Cressie, 1986). El término *kriging* suele usarse en inglés como verbo o sustantivo, dependiendo del contexto. En este trabajo se referirá a *kriging* como un sustantivo, que es el uso extendido que se le dio en la literatura académica angloparlante(Cressie, 1986; Thadani y cols., 1987).

En particular, el término de *kriging* se asigna a la predicción espacial con modelos geoestadísticos, según (Brus, 2022), donde un modelo geoestadístico

es un modelo estadístico de la variación espacial de una variable de estudio. El *kriging* busca una optimización del tamaño muestral o del patrón espacial de las ubicaciones muestrales(Brus, 2022).

2. Predicción Espacial: Kriging

De acuerdo con (Thompson, 2012) y (Arlinghaus, 1996), en el muestreo espacial se contemplan cantidades ambientales, ecológicas o geológicas como una variable aleatoria que denotaremos con y_t para una región t, cuyos valores son recolectados en una extensión de n regiones, $t_1, t_2, \cdots, t_n$. El fin del método kriging, acierta (Benedetti y cols., 2015), es predecir una variable aleatoria y_0 en una región vecina aun no muestreada.

La motivación principal para abordar muestreos en unidades distribuidas en una región con métodos especiales, radica en el hecho de que la información geográficamente distribuida tiene características y peculiaridades para las cuales los modelos muestrales tradicionales son inadecuados, señala (Benedetti y cols., 2015). Los métodos espaciales en estadística fortalecen el análisis de la información geográfica sujeta a variación de observaciones dependientes(Arlinghaus, 1996).

Las *unidades espaciales* del estudio se definen sobre una partición finita del dominio en formas regulares o irregulares que contienen un número de unidades observacionales(Thompson, 2012); de estas unidades espaciales, el investigador selecciona aleatoriamente las que conformarán su espacio muestral.

Tal vez la razón más fuerte para proceder de esta manera en una investigación geográfico-estadística es el hecho que las poblaciones espacialmente distribuidas son, en sí mismas, muy difíciles de muestrear dado que su esparcimiento puede no ser regular o presentar conglomerados dentro del dominio de estudio(Benedetti y cols., 2015).

Debido a que la variable de interés y_0 es una variable aleatoria concerniente a una región más dentro de la población de zonas a estudiar, no se le llama al método kriging uno de *estimación*, sino, propiamente de *predicción*(Thompson, 2012).

El método, en este sentido, busca una solución óptima(Brus, 2022) a una serie de ecuaciones en términos de una función de covarianzas o un variograma (el cual es una varianza de diferencias)(Thompson, 2012).

Refiriéndonos a (Thompson, 2012), ambos acercamientos prueban ser equivalentes cuando se conoce exactamente la función de covarianzas o el variogra-

ma.

Siguiendo a (Cressie, 1986), el método kriging posee dos variantes conocidas como *kriging ordinario* y el *kriging universal*; el primero se caracteriza porque el conocimiento del arrastre o función de media permite que el variograma estacionario pueda estimarse, mientras que en el segundo marco de trabajo, el variograma es conocido para estimar el arrastre.

En (Benedetti y cols., 2015) se señala la existencia de una versión adicional, el *kriging simple*, el cual contempla un conocimiento exacto de la media y la varianza en cada región; no obstante, debido a sus altas restricciones prácticas se utiliza en raras ocasiones.

Aunque existen muchas versiones del *kriging*, la mayoría parte del siguiente modelo genérico en (Brus, 2022)

$$\begin{aligned} y_t =& \mu_t + \epsilon_t && (1)\\ \epsilon_t \sim& \mathcal{N}(0, \sigma^2) && (2)\\ Cov(\epsilon_t, \epsilon_t') =& C(h) && (3) \end{aligned}$$

donde y_t es la variable regionalizada en la ubicación t, μ_t es la media en la ubicación t, ϵ_t es el residual en la ubicación t entendido como la diferencia entre la variable de estudio y y su media μ_t. La función $C(h)$ es la covarianza de los residuales entre dos localidades separadas por un vector de desplazamiento $h = t - t'$.

Este modelo básico se discutirá a mayor extensión en las subsecuentes secciones.

2.1. Función de Covarianza Espacial

Este acercamiento es tradicional en la estadística(Thompson, 2012), no obstante el variograma se ha empleado extensamente en la geoestadística(Cressie, 1986; Thompson, 2012; Brus, 2022).

Siguiendo a (Benedetti y cols., 2015) y en concordancia con (Cressie, 1986), las variables aleatorias y_t (también denotadas $y(\mathbf{z})$ por (Benedetti y cols., 2015)) son observaciones del proceso estocástico real-valuado definido sobre un dominio D

$$\left\{ y_t : t \in D \subset \mathbb{R}^d \right\} \tag{4}$$

donde D es un espacio de dimensión d continuamente variado y que ha sido observado en una serie de puntos dados $t_1,\ t_2,\ \cdots,\ t_n$ (o $\mathbf{z}_1,\ \mathbf{z}_2,\ \cdots,\ \mathbf{z}_n$, en la notación de (Benedetti y cols., 2015)). El índice t es un vector y se utiliza como un índice sobre *el espacio*(Cressie, 1986).

En las aplicaciones geológicas y ecológicas, los valores de la variable de interés en diferentes sitios suelen ser dependientes entre sí, más aún, los valores en regiones vecinas tienden a guardar una correlación, advierte (Thompson, 2012). La medida de relación entre las variables y_1 y y_2 asociadas a los sitios t_1 y t_2 es la covarianza, ec. 5:

$$Cov(y_1, y_2) = E\left[(y_1 - E(y_1))\,(y_2 - E(y_2))\right] \tag{5}$$

Cuando la covarianza entre dos sitios depende solo de sus posiciones relativas, se emplea una *función de covarianza* $C(h)$. De acuerdo con (Thompson, 2012), podemos obtener esta medida de relación entre los *y-valores* de cualesquiera dos sitios separados por h con

$$C(h) = Cov(y_{t+h}, y_t) \tag{6}$$

donde h es el vector de desplazamiento desde la región t_1 hacia la región t_2. Cuando la covarianza depende de la distancia, $d = \|h\|$, mas no de la dirección entre los sitios, se dice que el proceso es *isotrópico* y la función de covarianza se denota $C(d)$, (Thompson, 2012).

En su discusión sobre las propiedades de los sitios, (Benedetti y cols., 2015) agrega que suponiendo que el proceso espacial tiene una media μ_t para cada región t y que la varianza existe para todos los sitios t, dicho proceso se llama *estrictamente estacionario* si para cualquier $n \geq 1$ número de sitios $\{t_i\}_{i=1}^{n}$ y cualquier h, la distribución de observaciones $\{y_t\}_{t=t_1}^{t_n}$ es la misma que $\{y_{t+h}\}_{t=t_1}^{t_n}$.

Para (Thompson, 2012) un proceso se llama *estacionario de segundo orden* cuando el valor esperado de la variable de las observaciones regionales es constante entre ubicaciones, pero la covarianza depende solo del vector de desplazamiento.

Un proceso se llama *débilmente estacionario*, continua (Benedetti y cols., 2015), cuando la media es constante a través de los sitios t y el proceso es isotrópico en el sentido de (Thompson, 2012).

Proposición 2.1 *Si el proceso* $\left\{y_t : t \in D \subset \mathcal{R}^d\right\}$ *es estrictamente estacionario, entonces es débilmente estacionario.*

De acuerdo con (Benedetti y cols., 2015), un proceso en el que y_t no sea débilmente estacionario puede presentar estacionalidad en los incrementos $y_t - y_{t+h}$. Entonces se dice que y_t es *intrínsecamente estacionario* si[1] $\mu_t = \mu$ y

$$Var\left[y_t - y_{t+h}\right] = 2\gamma(h) \tag{7}$$

donde $2\gamma(h)$ recibe el nombre de *variograma*, mientras que $\gamma(h)$ solo se conoce como semivariograma(Thompson, 2012; Cressie, 1986; Benedetti y cols., 2015).

2.2. Predicción Lineal: Kriging

En la predicción lineal, también conocida como *kriging ordinario*(Cressie, 1986; Brus, 2022) se busca predecir el valor de una variable aleatoria y_0 de interés en un sitio t_0 a partir de la información disponible de un proceso estocástico ec. 4(Hohn y cols., 1993).

Utilizaremos la notación de (Thompson, 2012) para formular el procedimiento:

- *y-valor* observado en el iésimo sitio de la muestra de tamaño n: y_i
- iésimo sitio en la muestra de n sitios: t_i
- covarianza entre *y-valores* de un sitio i y un sitio j: $Cov(y_i, y_j) = c_{ij}$
- varianza del *y-valor* en el sitio t_i: $Var(y_i) = c_{ii}$

El objetivo es encontrar una función $\hat{y}_0$ de los n *y-valores* observados insesgada para y_0

$$E\left(\hat{y}_0\right) = E\left(y_0\right) \tag{8}$$

que minimice el error cuadrado medio de la predicción dada por

$$MSPE = E\left(y_0 - \hat{y}_0\right)^2 \tag{9}$$

Es importante notar, señala(Thompson, 2012), que aunque el mejor estimador de y_0 es la esperanza condicional de y_0 dada las observaciones $y_1, \cdots, y_n$, esto es muy difícil de conseguir dado que se precisa una conocimiento exacto de la distribución conjunta de las variables aleatorias.

[1] Aunque la suposición de que el valor esperado de la variable de interés es constante en cada sitio es rara vez cierta, ya que, de acuerdo con (Cressie, 1986), los datos rara vez presentan estacionalidad.

Un criterio más práctico es hallar una función lineal de los *y-valores* insesgada y que minimice ec. 9. Tal proceso de optimización se concreta mediante multiplicadores de Lagrange(Thompson, 2012; Brus, 2022).

Dado

$$\hat{y}_0 = \sum_{i=1}^{n} a_i y_i \tag{10}$$

se quieren hallar valores $\{a_i\}_{i=1}^{n}$ que minimicen ec. 9 sujeto a ec. 8. El problema se reescribe en forma matricial como:

$$\mathbf{f} = \mathbf{G}^{-1}\mathbf{h} \tag{11}$$

donde las matrices $\mathbf{f}_{n+1\times 1}$, $\mathbf{h}_{n+1\times 1}$ y $\mathbf{G}_{n+1\times n+1}$ son

$$\mathbf{f} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ m \end{pmatrix}, \quad \mathbf{h} = \begin{pmatrix} c_{10} \\ c_{20} \\ \vdots \\ c_{n0} \\ 1 \end{pmatrix}, \quad \mathbf{G} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} & 1 \\ c_{21} & c_{22} & \cdots & c_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix} \tag{12}$$

El método de *kriging ordinario* consiste en estimar los pesos de *kriging* $\{a_i\}_{i=1}^{n}$ y el multiplicador de Lagrange m(Brus, 2022; Thompson, 2012). En (Thompson, 2012), el mejor predictor lineal insesgado[2] $\hat{y}_0$ en ec. 10 se conoce como *predictor de kriging*. Y la *varianza de kriging* o *MSPE* se define como

$$MSPE = E\left(y_0 - \hat{y}_0\right)^2 = c_{00} - \sum_{i=1}^{n} a_i c_{i0} - m \tag{13}$$

Ya que la covarianza no se puede conocer exactamente, se utiliza un estimador desde los datos provenientes de la misma investigación o de una base de datos de estudios previos(Thompson, 2012). Para un proceso estacionario e isotrópico la covarianza en sitios a d unidades de distancia puede estimarse con una colección de n_d pares de ubicaciones a tal distancia d de separación de manera simple con

$$\hat{C}(d) = \frac{1}{n_d} \sum \left(y_{t_i} - \overline{y}\right)\left(y_{t_j} - \overline{y}\right) \tag{14}$$

[2]BLUP por sus siglas en inglés

donde la suma se efectúa sobre los pares de sitios a distancia d de separación. Seguido puede emplearse un método de mínimos cuadrados no lineal para obtener una función de covarianza y estimar la covarianza a cualquier distancia(Thompson, 2012).

Pero no cualquier función puede usarse para modelar la semivarianza bajo isotropía, pues se debe asegurar que la varianza del predictor de kriging sea positiva(Brus, 2022). En este sentido, dos modelos preferidos en la práctica son el exponencial y el esférico.

2.3. Variograma

En el caso de un proceso estacionario de segundo orden, (Thompson, 2012) explica que la función de covarianza y el variograma contienen información equivalente en tanto que

$$\gamma(h) = c(0) - c(h) \tag{15}$$

donde $c(0) = Var(y_t)$, la varianza de y en un sitio t arbitrario.

De acuerdo con (Cressie, 1986), el variograma es más general que la función de covarianza, pues este aún existe para algunos procesos que no cumplen estacionalidad de segundo orden. (Cressie, 1989) hace hincapié en que el valor del kriging en el análisis geoestadístico es mucho mayor de lo acreditado, en tanto que la clase de los problemas que que involucran variogramas contiene a la clase de problemas que emplean funciones de covarianza.

Un método relativamente simple de estimar el variograma en función de la distancia entre dos sitios está dado por (Thompson, 2012) como

$$2\hat{\gamma}(d) = \frac{1}{n_d} \sum (y_{t_i} - y_{t_j})^2 \tag{16}$$

Este acercamiento es recomendado(Thompson, 2012) en tanto que ec. 16 es insesgado, pero 14 no lo es.

Las ecuaciones de predicción pueden reescribirse en función del variograma en condiciones similares bajo covarianza estimada. El objetivo es nuevamente predecir el valor de la variable de interés en una región nueva, $\hat{y}_0$ usando las n observaciones de *y-valores* de manera insesgada:

$$E\left(\hat{y}_0\right) = E\left(y_0\right) \tag{17}$$

minimizando el *MSPE*

$$MSPE = E\left(y_0 - \hat{y}_0\right)^2 \tag{18}$$

Escribiendo el estimador lineal como

$$\hat{y}_0 = \sum_{i=1}^{n} a_i y_i \tag{19}$$

el problema se traduce en encontrar valores $a_1,\ a_2,\ \cdots,\ a_n$ que minimicen la ec. 18 sujeto a la condición ec. 17. La solución de (Thompson, 2012) viene dada en forma matricial:

$$\mathbf{a} = \mathbf{\Gamma}^{-1}\gamma \tag{20}$$

donde las matrices $\mathbf{a}_{n+1\times 1}$, $\gamma_{n+1\times 1}$ y $\mathbf{\Gamma}_{n+1\times n+1}$ están dadas abajo:

$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \\ m^* \end{pmatrix}, \quad \gamma = \begin{pmatrix} \gamma_{10} \\ \gamma_{20} \\ \vdots \\ \gamma_{n0} \\ 1 \end{pmatrix}, \quad \mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1n} & 1 \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2n} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{n1} & \gamma_{n2} & \cdots & \gamma_{nn} & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}$$

Dentro de este contexto la predicción ec. 20 satisface el MSPE de la siguiente forma, ec. 21

$$MSPE = E\left(y_0 - \hat{y}_0\right)^2 = \sum_{i=1}^{n} a_i \gamma_{i0} + m^* \tag{21}$$

Tal como se mencionó en la sección anterior, un modelo adecuado para las funciones de covarianza y variograma deben ser definidos positivos, de tal manera que se evite una varianza predictiva negativa(Cressie, 1986).

Para un modelo de *kriging* condicionado a isotropía, (Brus, 2022) dicta los siguientes parámetros para un modelo esférico del semivariograma:

1. *Nugget* c_0: es la intersección del variograma con el eje y.
2. *Partial sill* c_1: es la diferencia entre la máxima semivarianza y el *nugget*.
3. *Range* ϕ: la distancia a la cual la semivarianza alcanza su máximo valor.

Bajo estos criterios, se define el semivariograma esférico de la siguiente forma:

$$\gamma(d) = \begin{cases} 0 & \text{si } d = 0 \\ c_0 + c_1 \left[1 - \frac{3}{2}\left(\frac{d}{\phi}\right) + \frac{1}{2}\left(\frac{d}{\phi}\right)\right] & \text{si } 0 < d \leq \phi \\ c_0 + c_1 & \text{si } d > \phi \end{cases} \tag{22}$$

Donde la suma $c_0 + c_1$ se conoce como *varianza a priori*(Brus, 2022).

Por otro lado, un semivariograma exponencial con 3 parámetros, los mismos que (Brus, 2022) define para el caso esférico, presenta la siguiente forma

$$\gamma(d) = \begin{cases} 0 & \text{si } d = 0 \\ c_0 + c_1 exp(-d/\phi) & \text{si } d > 0 \end{cases} \tag{23}$$

Alternativamente, un modelo exponencial del variograma puede tomar una definición en el régimen no nulo como la utilizada por (Hohn y cols., 1993)

$$\gamma(d) = c_1(1 - \exp(-d/\phi))$$

donde el modelo tiende asintóticamente hacia el *sill*, con un acercamiento notable después de una distancia $d = 3\phi$. Además, (Hohn y cols., 1993) propone un modelo de variograma completo que involucra una descomposición de $\gamma(h)$ en términos aditivos

$$\gamma(h) = \gamma_1(h) + \gamma_2(h) + \cdots + \gamma_n(h)$$

Tal marco de trabajo se conoce como *modelos anidados* y son útiles en escenarios anisotrópicos donde la varianza presenta saturaciones en ciertas direcciones o los rangos varían según las direcciones entre sitios vecinos(Hohn y cols., 1993).

2.4. Predicción del Valor Sobre una Región

En muchas circunstancias se desea predecir el valor de la media de la variable de interés sobre una región, cuando no su total(Thompson, 2012). Las fórmulas discutidas en las secciones anteriores para realizar una predicción puntual $\hat{y}_0$ con su función de covarianza o variograma son válidas sobre una distribución discreta de la información(Thompson, 2012); no obstante, pueden traducirse a un dominio continuo empleando integrales en su cómputo como lo muestran (Cressie, 1986), (Grondona y Cressie, 1991) y (Thompson, 2012).

Si A es una región de estudio particionada en N sitios, (Thompson, 2012) y (Grondona y Cressie, 1991) definen la predicción y_0 como

$$y_0 = \frac{1}{|A|} \int_A y_t \, \mathrm{d}t \tag{24}$$

suponiendo que el área de cada partición tiene el mismo valor A. La semivarianza media entre el iésimo sitio y la región A está dado por:

$$\gamma_{i0} = \frac{1}{|A|} \int_A \gamma(y_i - y_t) \, \mathrm{d}t \tag{25}$$

EL *MSPE* se obtiene mediante

$$E\left(y_0 - \hat{y}_0\right)^2 = \sum_{i=1}^{n} a_i \gamma_{i0} + m^* - \gamma_{00} \tag{26}$$

done la semivarianza media γ_{00} está dada por

$$\gamma_{00} = \frac{1}{N^2} \int_A \int_A \gamma(t - \nu) \, \mathrm{d}t \, \mathrm{d}\nu \tag{27}$$

Un análisis de varianza es posible dentro de este contexto también, tal como lo demuestra (Grondona y Cressie, 1991) en la aplicación al análisis de experimentos que toman en cuentan correlaciones espaciales entre los datos.

Las estimaciones realizadas mediante bloques y aleatorización de las regiones muestrales prueban manejar de mejor manera las mediciones y ofrecen estimadores más eficientes, de acuerdo con (Grondona y Cressie, 1991).

3. Diseño Espacial

Siguiendo a (Thompson, 2012), la función de covarianza o equivalentemente el variograma aportan suficiente información para establecer el número de sitios que se debe muestrear para hacer una predicción aceptable de la variable de interés en una nueva región. Según (Cressie, 1986), incluso, los modelos de kriging permiten utilizar estudios piloto o bases de datos previas sobre una región para calibrar el instrumento de muestreo antes de realizar el kriging pertinente sobre la zona.

En este sentido, un análisis del MSPE, bajo las condiciones básicas del kriging, provee una vía para elegir un tamaño muestral espacial n idóneo para una predicción insesgada aceptable. De la expresión abajo

$$E\left(y_0 - \hat{y}_0\right)^2 = c_0 + \sum_{i=1}^{n}\sum_{i=1}^{n} a_i a_j c_{ij} - 2\sum_{i=1}^{n} a_i c_{i0}$$

se aprecia que las mejores predicciones resultan de las n regiones muestra que tienen la menor covarianza entre sí, que es lo que sucede cuando se particiona la región de estudio en estratos pequeños(Thompson, 2012), pero que poseen la mayor covarianza con el valor que se quiere predecir.

Para variograma cuyas estimaciones de semivarianza decrecen con la distancia, se puede establecer que las regiones que más cerca están de la región de interés son los que mostrarán la mayor covarianza con respecto a dicha región de interés(Thompson, 2012).

4. Implementación en R

4.1. Aplicación del *kriging ordinario*

Ejemplo 1: Predicción por *kriging ordinario* Datos sobre estudios de camarones provenientes del *Alaska Department of Fish and Game* en las proximidades de la Isla Kodiak, Alaska, fueron utilizados para estimar una función de covarianza espacial, la cual se empleó a su vez para predecir la cantidad de capturas en una nueva ubicación no muestreada previamente. Los datos de capturas fueron ploteados por ubicación en una tabla de la región de estudio registrando los pesos en libras (lbs) y las distancias en millas náuticas (nmi). Una unidad de investigación arrastró una línea de pesca por aproximadamente 1 nmi en un patrón grillado. Las covarianzas muestrales se obtuvieron usando pares de datos compactos en intervalos de distancias. Luego se ajustó los estimados de covarianza mediante mínimos cuadrados no lineales a una curva exponencial obteniendo la siguiente función de covarianza:

$$C(x) = 5.1\exp(-0.49x)$$

Supóngase que una de las grúas atrapó $y_1 = 5.526$ klbs y una segunda grúa consiguió capturar $y_2 = 1.417$ klbs a 6 nmi de distancia. ¿Cuál puede ser el tamaño (en libras) de una captura a 1 nmi de la primera captura y 5.4 nmi de la segunda?

Solución La varianza es $C(0) = 5.1$. La covarianza entre las grúas a 6 nmi de separación es $c_{12} = 5.1 \cdot \exp(-0.49 \cdot 6) = 0.3$. Mientras que las covarianzas con respecto al nuevo sitio t_0 son $c_{10} = 3.1$ y $c_{20} = 0.4$, respectivamente para el sitio 1 y el sitio 2. La ecuación predictiva 11 toma la forma definida:

$$\begin{pmatrix} a_1 \\ a_2 \\ m \end{pmatrix} = \begin{pmatrix} 5.1 & 0.3 & 1 \\ 0.3 & 5.1 & 1 \\ 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} 3.1 \\ 0.4 \\ 1 \end{pmatrix}$$

El siguiente script de `R` computa el resultado con valores $a_1 = 0.78$, $a_2 = 0.22$ y $m = -0.95$ para un *BLUP*, ec. 10, de $\hat{y}_0 = 4.622\,\text{klbs} = 4622\,\text{lbs}$. El *MSPE*, ec. 13, es de 3.5 para un error cuadrático medio, *MSE*, de $1.9\,\text{klbs} = 1900\,\text{lbs}$.

```r
#Vector de capturas en las ubicaciones 1 y 2
y <- c(5.526, 1.417)

#Matriz de covarianzas entre sitios de la muestra
G <- cbind(c(5.1, 0.3, 1), c(0.3, 5.1, 1), c(1, 1, 0))

#Vector de covarianzas entre sitios de la muestra con el sitio
#de la prediccion
h <- c(3.1, 0.4, 1)

#Vector de coeficientes optimos y multiplicador de Lagrange
f <- round(solve(G, h), 2)

#Prediccion de kriging en la localidad de interes
yhat_0 <- round(sum(f[1:2]*y), 3)

#Varianza de kriging de la prediccion
MSPE <- round(G[1,1] - sum(f[1:3]*h[1:3]), 1)

#Error medio de la prediccion
MSE <- round(sqrt(MSPE), 1)
```

Listing 1: Ejemplo de predicción mediante kriging ordinario

4.2. Cómputo del variograma

Ejemplo 2: Re-evaluado mediante *semivariograma* Este ejemplo se desarrolla sobre el enunciado del ejemplo 4.1, esta ves incorporando el cálculo del semivariograma calculado a partir de la ec. 15.

$$\gamma(x) = 5.1 - 5.1 \exp(-0.49x)$$

Solución La semivarianza para las grúas a 6 nmi de distancia es

$$\gamma_{12} = 5.1(1 - \exp(-0.49 \cdot 6)) = 4.8$$

Las semivarianzas con respecto al nuevo sitio son $\gamma_{10} = 2.0$ y $\gamma_{20} = 4.7$, respectivamente para la grúa en la ubicación t_1 y la grúa en la ubicación t_2. La ecuación predictiva, ec. 20, cobra la forma

$$\begin{pmatrix} a_1 \\ a_2 \\ m^* \end{pmatrix} = \begin{pmatrix} 0 & 4.8 & 1 \\ 4.8 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}^{-1} \begin{pmatrix} 2.0 \\ 4.7 \\ 1 \end{pmatrix} \tag{28}$$

El siguiente script de R computa el resultado con valores $a_1 = 0.78$, $a_2 = 0.22$ y $m^* = 0.95$ para un *BLUP*, ec. 10, de $\hat{y}_0 = 4.622\,\text{klbs} = 4622\,\text{lbs}$. El *MSPE*, ec. 21, es de 3.5 para un error cuadrático medio, *MSE*, de 1.9 klbs = 1900 lbs.

```
#Vector de capturas en las ubicaciones 1 y 2
y <- c(5.526, 1.417)

#Matriz de variogramas entre sitios de la muestra
Gamma <- cbind(c(0, 4.8, 1), c(4.8, 0, 1), c(1, 1, 0))

#Vector de variogramas entre sitios de la muestra con el sitio
#de la prediccion
gamma <- c(2.0, 4.7, 1)

#Vector de coeficientes optimos y multiplicador de Lagrange
a <- round(solve(Gamma, gamma), 2)

#Prediccion de kriging en la localidad de interes
yhat_0 <- round(sum(a[1:2]*y), 3)

#Varianza de kriging de la prediccion
MSPE <- round(sum(a[1:3]*gamma[1:3]), 1)

#Error medio de la prediccion
MSE <- round(sqrt(MSPE), 1)
```

Listing 2: Ejemplo de predicción mediante cómputo de variograma

El paquete básico de R para analizar datos geoestadísticos es geoR. El ejemplo siguiente utiliza datos sobre una producción de frijol de soja disponibles en la librería geoR.

```
#install.packages("geoR")
library(geoR)

#asignacion de la data de interes
produccion98 <- as.geodata(soja98, coords.col = 1:2, data.col = "PH")
```

```
6
7  #la funcion variog da el variograma para una distancia maxima de 50
       unidades
8  variograma <- variog(produccion98, max.dist = 50)
9  plot(variograma, pch = 19, cex = 1)
10
11 #curva del ajusted para el variograma
12 fitted <- variofit(variograma)
13 lines(fitted)
```

Listing 3: Ejemplo usando geoR

La ejecución del código en listing 3 genera la gráfica en la figura 1 que muestra los valores de semivarianza correspondientes a cada distancia a la cual se recogen las muestras. La línea de tendencia ayuda a determinar la forma particular de la función de semivarianza que debe implementarse para realizar las predicciones.

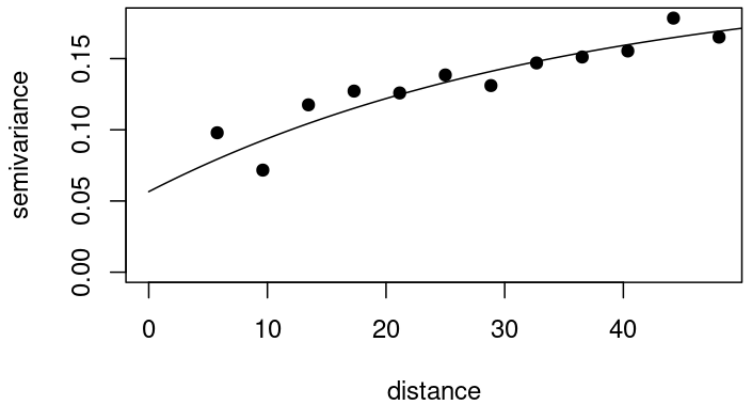

Figura 1: Gráfico de semivarianza contra distancia de los datos de producción de frijol de soja en listing 3.

En particular, para la ejecución de listing 3 la curva que mejor se ajusta (por mínimos cuadrados) a la distribución de puntos es una exponencial con *Range* $\phi = 35.5832$, *partial sill* $c_1 = 0.1520$ y *nugget* $c_0 = 0.0566$. Estos datos pueden ser visualizados con el comando `summary(fitted)`.

Mediante el paquete `geoR` es posible generar un mapa de contornos de la región muestreada, tal como se muestra en listing 4.

```
1  #install.packages("geoR")
2  library(geoR)
3  prod98 <- as.geodata(soja98, coords.col = 1:2, data.col = "PH")
4  locat <- expand.grid(seq(min(soja98$X), max(soja98$X), l=100),
5                       seq(min(soja98$Y), max(soja98$Y), l=100))
6  kc <- krige.conv(prod98, loc=locat,
7                   krige = krige.control(type.krige = "ok", cov.pars=c(1,20)
       ))
```

```
image(kc, col=gray((5:50)/55), axes = T)
contour(kc, axes = T, add = T)
points(prod98, cex.min = 0.1, cex.max = 1.5, pch = 19, add = T)
```

Listing 4: Script para generar un mapa de contornos en R

La figura 2 muestra la distribución de las localidades muestreadas un cultivo de frijol de soja generada por R. Se puede apreciar el mapa de contornos marcando delimitadores para las regiones de interés.

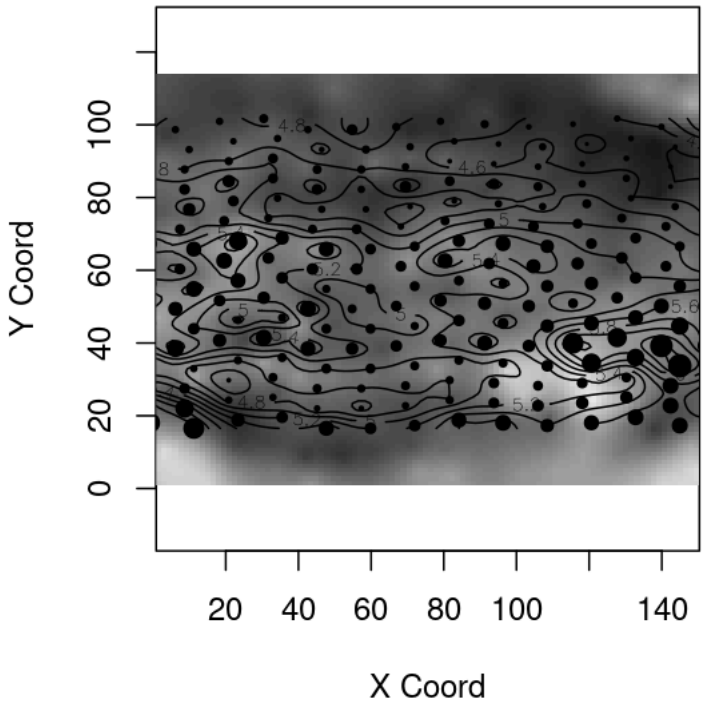

Figura 2: Mapa de contorno de los datos muestreados en el código en listing 4.

Existen formas de estimar el variograma utilizando el método de momentos y máxima verosimilitud sobre una base de datos geoespaciales, como lo explica extensamente (Brus, 2022). Los paquetes `geostat` y `geoR` proveen muchos métodos para concretar tales estimaciones bajo diferentes condiciones de estacionalidad, nuevamente (Brus, 2022) además de (Benedetti y cols., 2015) ofrecen ejemplos puntuales de su uso.

5. Conclusiones

La estadística espacial abre un sinfín de oportunidades para la comprensión de la naturaleza y la dimensión de las actividades humanas en el contexto del cambio climático y la explotación de recursos naturales. Más aún, todavía concede una oportunidad para la construcción de sistemas de información geográfica más precisos, pues de la calidad de los datos depende también en cierta medida la calidad de los resultados provistos por estas técnicas.

La diversidad de aplicaciones y la extensión en que pueden mejorar la calidad de vida de una sociedad que ofrecen estos métodos va más allá de una apreciación del conocimiento en cuanto que es conocimiento. Y es aquí donde radica la relevancia de la geoestadística, la estadística agrícola y ambiental en el presente.

El tratamiento de observaciones recolectadas sobre muestras espacialmente esparcidas toma conceptos prestados de técnicas de muestreo tradicionales como la estratificación y el muestreo por conglomerados, tales como reducir la varianza entre grupos o analizar las fuentes de variabilidad por separado dentro del diseño muestral. Conjuntamente, la técnica de kriging combina el interés de la estadística por encontrar patrones en el caos y la impredictibilidad de la naturaleza con las técnicas de optimización propias de la ingeniería matemática.

Este acercamiento ofrece una reducción sobre el ruido esperado o tolerado para una investigación estadística y promueve la obtención de modelos más precisos para la extracción de resultados significativos de la información que puede ofrecer un proceso al científico o creador de políticas.

La implementación del muestreo estadístico puede concretarse con una aplicación directa de los modelos teóricos desarrollados y documentados en la literatura académica, pero también se puede conseguir a través de paquetes computacionales especializados para el análisis de datos geoespaciales.

Una complementación interesante al uso de `R` para estudios geográficos, a mí parecer, es el de las bases de datos geoposicionales implementadas como gestores de bases de datos como `PostgreSQL` con el paquete `PostGIS`. Personalmente me motiva a seguir estudiando este tema.

Bibliografía

Arlinghaus, S. L. (Ed.). (1996). *Practical handbook of spatial statistics* (1st ed.). CRC Press.

Benedetti, R., Piersimoni, F., y Postiglione, P. (Eds.). (2015). *Sampling spatial units for agricultura surveys* (1st ed.). Springer.

Brus, D. J. (2022). *Spatial sampling with r* (1st ed.). CRC Press.

Cressie, N. (1986, Septiembre). Kriging nonstationary data. *Journal of the American Statistical Association*, *81*(395), 625 - 634. doi: http://dx.doi.org/10.1080/01621459.1986.10478315

Cressie, N. (1989, Noviembre). Geostatistics. *The American Statistician*, *43*(4), 197 - 202. doi: http://dx.doi.org/10.1080/00031305.1989.10475658

Grondona, M. O., y Cressie, N. (1991, Noviembre). Using spatial considerations in the analysis of experiments. *Technometrics*, *33*(4), 381 - 392. doi: http://dx.doi.org/10.1080/00401706.1991.10484867

Hohn, M. E., Liebhold, A. M., y Gribko, L. (1993, Octubre). Geostatistical model for forecasting spatial dynamics of defoliation caused by the gypsy moth (lepidoptera: Lymantriidae). *Environmental Entomoly*, *22*(5), 1066 - 1075. doi: http://dx.doi.org/10.1080/00401706.1991.10484867

Thadani, S. G., Alabert, F., y Journel, A. G. (1987). An integrated geostatistical/pattern recogniction technique for characterization of reservoir spatial variability. *SEG Technical Program Expanded Abstracts*, 372 - 375. doi: http://dx.doi.org/10.1190/1.1892098

Thompson, S. K. (2012). *Sampling* (3rd ed.). John Wiley & Sons, Inc.