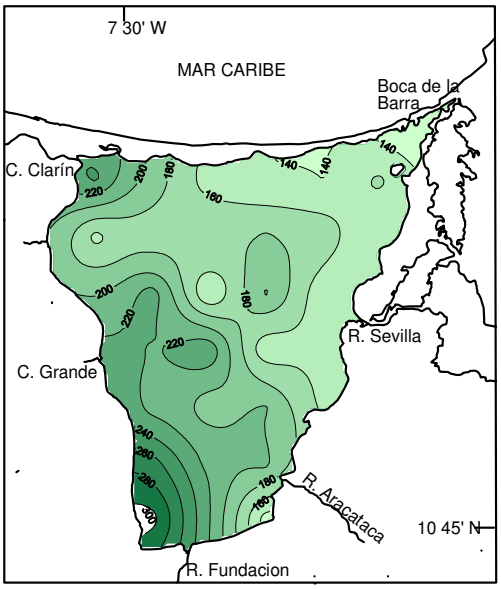
# Estadística Espacial




## NOTAS DE CLASE

Ramón Giraldo Henao

Profesor Asociado

Departamento de Estadística

Universidad Nacional de Colombia

Bogotá, Agosto de 2009

# Prefacio

La necesidad de acudir a herramientas estadísticas para el análisis de datos en todas las áreas del conocimiento, ha hecho que aparezcan con el correr de los años nuevas metodologías que, no obstante se centran en fundamentos probabilísticos comunes, son específicas para cada una de las diversas disciplinas del saber. Algunos ejemplos son econometría, psicometría o bioestadística. Desde hace unas dos décadas se ha dedicado mucho interés a la modelación de datos espaciales, entendidos como aquellos en los que además de los valores de las variables de interés se registran también lo sitios donde estos ocurren. El análisis de datos espaciales se ha hecho frecuente en muchas áreas del conocimiento, incluyendo ente otras geografía, geología, ciencias ambientales, economía epidemiología o medicina. Las técnicas estadísticas clásicas suponen que al estudiar un fenómeno se realizan observaciones bajo circunstancias idénticas y además que las observaciones son independientes y por ello no son convenientes para analizar fenómenos que varían en tiempo o espacio. Cuando se tienen datos espaciales, intuitivamente se tiene la noción de que observaciones cercanas están correlacionadas y por ello es necesario acudir a herramientas de análisis que contemplen dicha estructura.

El presente documento ha sido preparado como guía de clase para estudiantes del curso de estadística espacial del pregrado en estadística de la Universidad Nacional de Colombia. Este cubre los elementos esenciales de cada una de las técnicas de análisis de datos espaciales, es decir, geoestadística, datos de áreas (regionales) y patrones puntuales. Estas notas también tienen el propósito servir de consulta a geólogos, biólogos, agrónomos, ingenieros, meteorólogos, ecólogos, epidemiólogos y todos aquellos profesionales que se encargan del estudio de información georreferenciada. El documento tiene un enfoque teórico-práctico.

Para el seguimiento completo de la teoría descrita se requiere tener conocimientos básicos de álgebra lineal, estadística matemática (univariable y multivariable) y de modelos lineales. Un resumen no exhaustivo de estos temas es hecho al final en el apéndice. Aquellos lectores que estén poco familiarizados con los métodos estadísticos, podrán obviar las secciones en las que se hacen desarrollos teóricos y centrar su atención en la filosofía de los métodos presentados y en las aplicaciones mostradas en cada uno de los capítulos del documento. No obstante en el escrito se cubren diversos temas de la estadística espacial y se hacen aplicaciones de métodos recientes, es necesario acudir a la lectura de artículos científicos y textos avanzados para lograr un buen dominio de estas metodologías.

En el texto no se hace una redacción en castellano riguroso, en el sentido de que en muchas ocasiones se prefiere evitar la traducción de palabras técnicas. Por ello en muchas partes del documento se encuentran expresiones como *kriging*, *test*, *nugget* o *kernel*, en vez de traducciones como *krigeage*, *prueba*, *pepita* o *núcleo*, respectivamente..

# Índice general

# Índice de figuras

# Capítulo 1

# Datos Espaciales y Análisis Exploratorio

En este capítulo se presentan conceptos básicos de probabilidad que permiten posteriormente enmarcar las áreas de la estadística espacial dentro del contexto de los procesos estocásticos. Se definen algunas medidas de autocorrelación espacial y se dan dos ejemplos de como la dependencia espacial afecta la inferencia estadística.

## 1.1. Conceptos básicos de probabilidad y procesos estocásticos

**Definition 1.1.** *Sea $\Omega \neq \emptyset$. Un sistema $\mathcal{F}$ de subconjuntos de $\Omega$ se llama $\sigma$-álgebra si satisface las siguientes condiciones*

1. $\Omega \in \mathcal{F}$
2. Si $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$
3. Si $A_1, \cdots, A_n \in \mathcal{F} \Rightarrow \bigcup_{i=1}^{n} A_i \in \mathcal{F}$.

**Definition 1.2.** *Sea $\Omega \neq \emptyset$, $\mathcal{F}$ una $\sigma$-álgebra de subconjuntos de $\Omega$. La pareja $(\Omega, \mathcal{F})$ se llama espacio medible.*

**Definition 1.3.** *Sea* $(\Omega, \mathcal{F})$ *un espacio medible.* $P : \mathcal{F} \to [0, 1]$ *se llama medida de probabilidad si satisface*

1. $P(A) \geq 0, \forall A \in \mathcal{F}$
2. $P(\Omega) = 1$
3. Si $A_1, \cdots, A_n \in \mathcal{F}$ con $A_i \cap A_j = \emptyset, \forall i \neq j \Rightarrow P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.

**Propiedades de** $P$

1. $P(\emptyset) = 0$
2. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
3. Si $A \subseteq B \Rightarrow P(A) \leq P(B)$.
4. $P(A) = 1 - P(A^c)$.

**Definition 1.4.** *La tripla* $(\Omega, \mathcal{F}, P)$*, donde* $\Omega \neq \emptyset, \mathcal{F}$ *$\sigma$-álgebra sobre* $\Omega$ *y* $P$ *es una medida de probabilidad sobre* $(\Omega, \mathcal{F})$*, se denomina espacio de probabilidad.*

**Definition 1.5.** *Sea* $(\Omega, \mathcal{F}, P)$*, un espacio de probabilidad.* $X : \Omega \to \mathbb{R}$ *se llama variable aleatoria.*

**Definition 1.6** (**Proceso Estocástico**)**.** *Es una familia de variables aleatorias* $\{Z(s) : s \in D \subset \mathbb{R}^P\}$ *definida sobre un espacio de probabilidad* $(\Omega, \mathcal{F}, P)$*. El conjunto* $D$ *deíndices del procesos se denomina espacio de parámetros. Los valores que toma* $Z(s)$ *se llaman estados y el conjunto de todos los posibles valores de* $Z(s)$ *se llama espacio de estados.*

Los procesos estocásticos son clasificados de acuerdo con el espacio de parámetros (discreto y continuo) y el espacio de estados (discreto y continuo). Algunos ejemplos de procesos estocásticos no espaciales son los siguientes

1. Espacio de parámetros discreto y espacio de estados discreto
   - Z(n): Preferencia del consumidor en el n-ésimo mes, con n $\in \mathbb{N}$.

- Z(n): Número de individuos de la n-ésima generación de una población, con n $\in \mathbb{N}$.

2. Espacio de parámetros continuo y espacio de estados discreto

   - Z(t): Número de partículas de una sustancia acuosa de volumen t, con t $\in T \subset \mathbb{R}$.
   - Z(t): Número de individuos que esperan el bus por periodo de tiempo t, con t $\in T \subset \mathbb{R}$.

3. Espacio de parámetros discreto y espacio de estados continuo

   - Z(n): Tiempo de espera hasta que el n-ésimo estudiante arribe a la parada de bus, con n $\in \mathbb{N}$.
   - Z(n): Utilidad en pesos de un jugador después del n-ésimo lanzamiento de una moneda, con n $\in \mathbb{N}$.

4. Espacio de parámetros continuo y espacio de estados continuo

   - Z(t): Contenido de un embalse sobre un periodo de tiempo t, con t $\in T \subset \mathbb{R}$.
   - Z(t): Temperatura en el instante t, con t$\in T \subset \mathbb{R}$.

## 1.2. Datos espaciales y áreas de la estadística espacial

Estadística espacial es la reunión de un conjunto de metodologías apropiadas para el análisis de datos que corresponden a la medición de variables aleatorias en diversos sitios (puntos del espacio o agregaciones espaciales) de una región. De manera más formal se puede decir que la estadística espacial trata con el análisis de realizaciones de un proceso estocástico $\{Z(s) : s \in D \subset \mathbb{R}^P\}$, en el que $s$ es la ubicación en el espacio Euclidiano $P$-dimensional y $Z(s)$ es una variable aleatoria en la ubicación $s$.

**Observaciones**

- El proceso estocástico $\{Z(s) : s \in D \subset \mathbb{R}^P\}$, en el que $s$ es sitio del espacio, también se denomina *proceso aleatorio* o *campo aleatorio*.

- El proceso estocástico $\{\mathbf{Z}(s) : s \in D \subset \mathbb{R}^P\}$, en el que $\mathbf{Z}(s)$ es un vector aleatorio se denomina en el contexto espacial *proceso aleatorio multivariable* o *campo aleatorio multivariable*

- El proceso estocástico $\{Z(s) : s \in D \subset \mathbb{R}^P\}$ en el que tanto el espacio de estados como el espacio de parámetros es continuo (es decir que las variables aleatorias $Z(s)$ son continuas y $D \subset \mathbb{R}^P$ es un conjunto continuo) se denomina *variable regionalizada*. Este término es particularmente usado en aplicaciones de la estadística espacial en ingeniería y geología.

- Cuando se tiene una observación del proceso estocástico $\{Z(s) : s \in D \subset \mathbb{R}^P\}$ se dispone de una muestra de tamaño $\{\mathbf{Z}(s) = (Z(s_1), Z(s_2), \cdots, Z(s_n))\}$ (con $n$ el número de sitios donde se hace la medición de la variable aleatoria $Z(s)$) y no de una muestra de tamaño $n$ de una variable aleatoria. Por ello puede ser carente de sentido práctico hacer inferencia estadística clásica (intervalos de confianza, pruebas de normalidad, etc) con los datos obtenidos. Desconocer esto hace que se cometan errores intentando validar los supuestos necesarios para la aplicación de métodos estadísticos espaciales. En general en estadística espacial, como en el caso clásico, es deseable tener normalidad para hacer inferencia. Sin embargo lo que se asume en este contexto es que la muestra corresponde a la observación de vector aleatorio con distribución normal multivaluada y no que se tiene una muestra $n$-variada de una variable aleatoria con distribución normal. Usar una prueba de normalidad univariada (por ejemplo la de Shapiro-Wilk) para comprobar si los datos siguen una distribución normal es ciertamente equivocado en el contexto espacial, puesto que además de desconocer que no se tiene una muestra *iid* (puesto que hay dependencia espacial), lo que en realidad habría que probar es normalidad multivariada.

La estadística espacial se subdivide en tres grandes áreas. La pertinencia de cada una de ellas está asociada a las características del conjunto $D$ de índices del proceso estocástico

de interés. A continuación se mencionan dichas áreas y se describen las propiedades de $D$ en cada una de éstas.

**Geoestadística**: Estudia datos de procesos estocásticos en los que el espacio de parámetros $D \subset \mathbb{R}^P$ es continuo. Algunos ejemplos de datos espaciales que son tratados con métodos geoestadísticos son

- $\{Z(s) : s \in D \subset \mathbb{R}^P\}$, donde $Z(s)$ mide el contenido de nitrógeno en sitios de un parcela experimental. En este caso los sitios pertenecen a $D \subset \mathbb{R}^2$.
- $\{Z(s) : s \in D \subset \mathbb{R}^P\}$, donde $Z(s)$ corresponde a la precipitación en sitios de Colombia.

En los dos ejemplos anteriores hay infinitos sitios donde medir y por ello el conjunto de parámetros es continuo. Sin embargo en la practica es potestad del investigador seleccionar en que sitios de la región de interés hace la medición de las variables, es decir, el investigador puede hacer selección de puntos del espacio a conveniencia o puede seleccionar los sitios bajo algún esquema de muestreo probabilístico. En este sentido se dice que el conjunto $D \subset \mathbb{R}^P$ es *fijo*. Un ejemplo de un conjunto de datos analizado con métodos geoestadísticos es presentado en la Figura 1.1. Es importante resaltar que en geoestadística el propósito esencial es la interpolación y si no hay continuidad espacial pueden hacerse predicciones carentes de sentido.

**Datos de áreas o regionales**: En este caso el proceso estocástico tiene espacio de parámetros $D \subset \mathbb{R}^P$ discreto y la selección de los sitios de medición depende del investigador ($D$ fijo). Las ubicaciones de muestreo pueden estar regular o irregularmente espaciadas. Dos ejemplos de datos regionales son

- $\{Z(s) : s \in D \subset \mathbb{R}^P\}$, donde $Z(s)$ es la variable aleatoria correspondiente a la tasa de mortalidad y los sitios son departamentos de Colombia, es decir $D$ es el conjunto discreto formado por los departamentos del país.

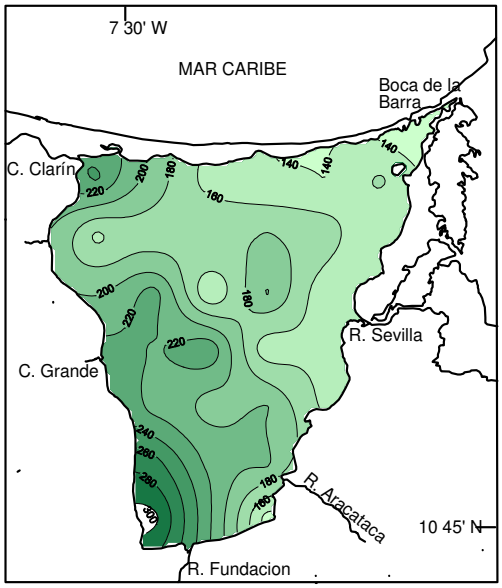



Figura 1.1: Distribución espacial de clorofila en la Ciénaga Grande de Santa Marta (costa norte de Colombia). Datos medidos en un jornada de muestreo realizada en marzo de 1997.

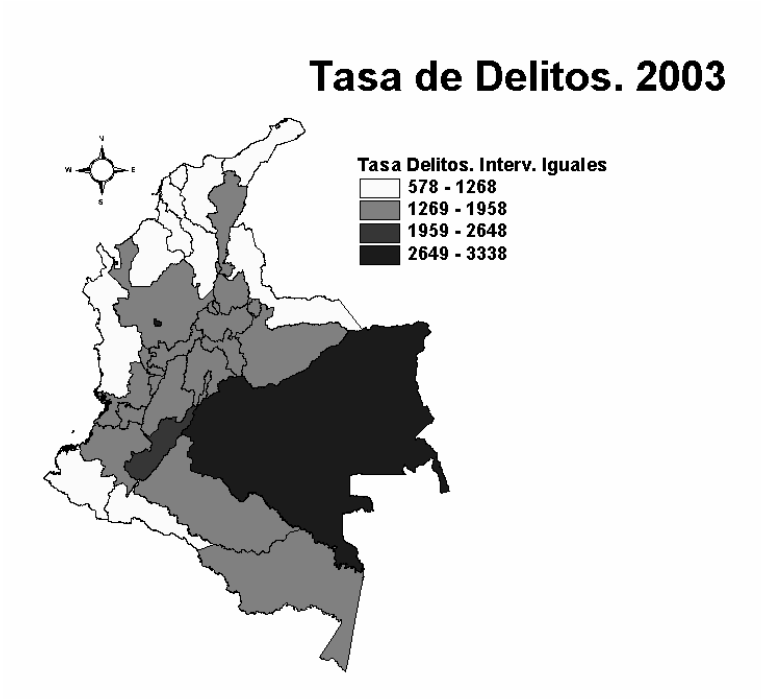



Figura 1.2: Mapa cloroplético de la tasa de delitos en Colombia en el año 2003.

- $\{Z(s) : s \in D \subset \mathbb{R}^P\}$, donde $Z(s)$ corresponde a la producción cafetera (en kilogramos) y $D$ es el conjunto de todas las fincas productoras de café del país.

Nuevamente el investigador puede decidir donde (en que departamentos o en que fincas en los ejemplos) hace la medición de las variables de interés, es decir en datos de áreas también el conjunto $D \subset \mathbb{R}^P$ es *fijo*. En la Figura 1.2 se presenta un ejemplo de un conjunto de datos que corresponde a la observación de un proceso aleatorio de datos regionales. Un ejemplo de datos de área con sitios regularmente espaciados es el de colores de pixeles en interpretación de imágenes de satélite. En ese caso el conjunto de ubicaciones de interés es discreto y estas corresponden a agregaciones espaciales más que a un conjunto de puntos del espacio. Es obvio que la interpolación espacial puede ser carente de sentido con este tipo de datos. Sus principales aplicaciones se encuentran en el campo epidemiológico.

**Patrones Puntuales**: La diferencia central del análisis de patrones puntuales con las

técnicas geoestadísticas y el análisis de datos de áreas radica en el hecho de que el conjunto $D \subset \mathbb{R}^P$ es aleatorio, es decir que la decisión al respecto de donde se hace la medición no depende del investigador. Dicho conjunto puede ser discreto o continuo, pero la ubicación de los sitios donde ocurre el fenómeno a estudiar es dada por la naturaleza. En general el propósito de análisis en estos casos es el de determinar si la distribución de los individuos dentro de la región es aleatoria, agregada o uniforme. Algunos ejemplos de datos correspondientes a patrones puntuales son dados a continuación

- Ubicación de nidos de pájaros en una región dada.
- Localización de imperfectos en una placa metálica
- Sitios de terremotos en Colombia
- Municipios de Colombia con mayorías negras

En los tres primeros ejemplos $D \subset \mathbb{R}^P$ es continuo y en el último es discreto. Cuando en cada sitio se hace medición de alguna variable (por ejemplo del número de huevos en los nidos, de la forma del imperfecto en la placa o de la tasa de analfabetismo de los municipios de mayorías negras) se dice que ese tiene un *patrón espacial marcado*. Dos ejemplos de datos correspondientes a patrones espaciales son dados en la Figura 1.3.

## 1.3. Medidas de dependencia espacial

La dependencia espacial hace referencia a la estructura de correlación de las variables aleatorias del proceso $\{Z(s) : s \in D \subset \mathbb{R}^P\}$. Cuando hay dependencia espacial los sitios cercanos tienen valores más similares que los distantes. Por el contrario la ausencia de correlación espacial se refleja en el hecho de que la distancia entre los sitios no tiene influencia en la relación de sus valores. A continuación se presentan algunos test y funciones que permiten establecer estadísticamente o de manera empírica si existe dependencia (correlación) espacial.

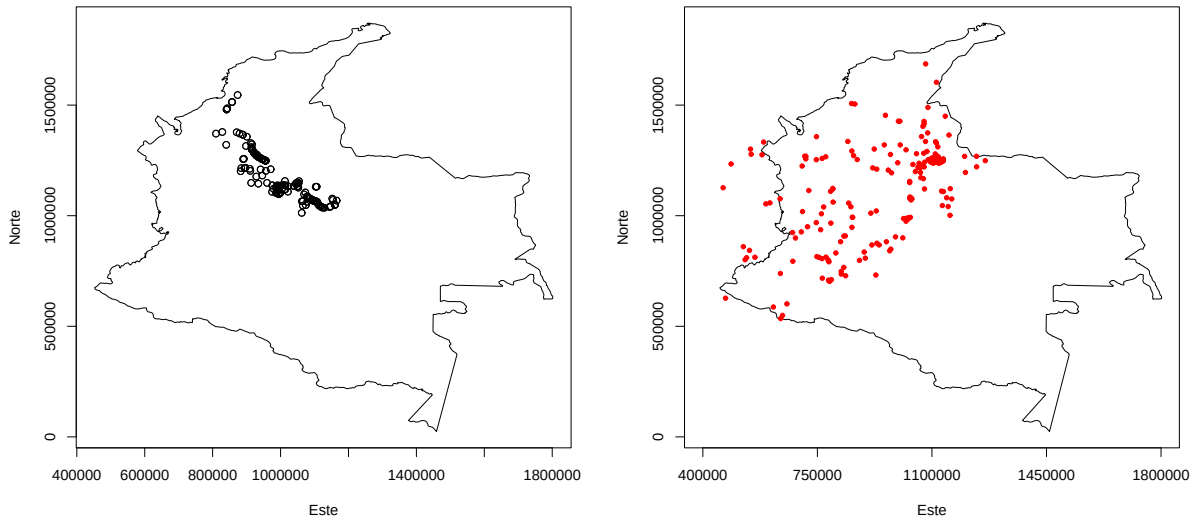



Figura 1.3: Ubicación de deslizamientos en el corredor Caño Limón-Coveñas en 2008 (*panel izquierdo*) y ubicación de sismos de baja magnitud en Colombia en el periodo Julio a Diciembre de 2008 (*panel derecho*).

### 1.3.1. Test de Mantel

Permite comprobar estadísticamente si las observaciones provienen de un proceso estocástico en el que las variables son correlacionadas espacialmente.

*Hipótesis*

$$H_0 : \text{Hay aleatoriedad espacial}$$
$$H_a : \text{Hay correlación espacial}$$

*Estadística de prueba*

$$M = \sum_{i=1}^{n} \sum_{i=1}^{n} W_{ij} U_{ij},$$

donde $Wij = \|s_i - s_j\|$ y $U_{ij} = (Z(s_i) - Z(s_j))^2$. La estadística de mantel está relacionada con la pendiente del modelo de regresión simple $U_{ij} = \beta W_{ij} + e_{ij}$ a través de $\beta = M / \sum_{i=1}^{n} \sum_{i=1}^{n} W_{ij}^2$, es decir que intuitivamente se tiene que a mayor $M$, mayor dependencia espacial positiva. La significancia de la prueba puede establecerse por varios caminos. Puede emplearse un *test de permutaciones* en el que asumiendo aleatoriedad se

encuentran las $n!$ posibles asignaciones de sitios a valores y con cada una de ellas se calcula $M$, obteniéndose por consiguiente su distribución bajo $H_0$. También en el caso de $n$ grande puede usarse un *test de Monte Carlo* en el que solo se toman $k$ de las asignaciones aleatorias de sitios a valores de la variable. En ambos casos (permutaciones, Monte Carlo) podría usarse una aproximación a la normal estimando $E(M)$ y $V(M)$ a través de $\bar{M} = 1/n \sum_{i=1}^{n} M_i$ y $s_M^2 = 1/(n-1) \sum_{i=1}^{n} (M_i - \bar{M})^2$. En el caso de asumir normalidad y aleatoriedad, es decir si $Z(s_1), \cdots .Z(s_n)$ son *iid*, con $Z(s_i) \sim N(\mu, \sigma^2)$, pueden obtenerse expresiones para $E(M)$ y $V(M)$ y establecer el nivel de significancia basándose en un *test normal*.

### 1.3.2. Test de Moran

Este test es especialmente usado en datos de áreas. Sean $Z(s_1), \cdots, Z(s_n)$, las variables medidas en las $n$ áreas. La noción de autocorrelación espacial de estas variables está asociada con la idea de que valores observados en áreas geográficas adyacentes serán más similares que los esperados bajo el supuesto de independencia espacial. El índice de autocorrelación de Moran considerando la información de los vecinos más cercanos es definida como

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij}(Z(s_i) - \bar{Z})(Z(s_j) - \bar{Z})}{\sum_{i=1}^{n}(Z(s_i) - \bar{Z})^2} \tag{1.1}$$

Valores positivos (entre 0 y 1) indican autocorrelación directa (similitud entre valores cercanos) y valores negativos (entre -1 y 0) indican autocorrelación inversa (disimilitud entre las áreas cercanas). Valores del coeficiente cercanos a cero apoyan la hipótesis de aleatoriedad espacial.

Para el cálculo del índice de Moran es necesario definir la proximidad entre las áreas. Lo anterior se lleva a cabo por medio del cálculo de una matriz de proximidad espacial. Dado un conjunto de $n$ áreas $(A_1, \cdots A_n)$ se construye una matriz $W^{(1)}$ de orden $(n \times n)$ donde cada uno de los elementos $W_{ij}$ representa una medida de proximidad entre $A_i$ y $A_j j$. Dicha medida puede ser calculada con alguno de los siguientes criterios:

- $W_{ij} = 1$ si el centro de $A_i$ se encuentra a una distancia determinada de $A_j$ o $W_{ij} = 0$ en caso contrario.
- $W_{ij} = 1$ si $A_i$ comparte frontera con $A_j$ y en caso contrario $W_{ij} = 0$.
- $W_{ij} = I_{ij}/I_i$, donde $I_{ij}$ es la longitud de la frontera entre $A_i$ y $A_j$ y $I_i$ es el perímetro de $A_i$.
- $W_{ij} = d_{ij}$, con $d_{ij}$ la distancia entre los centros de las dos áreas.

En todos los casos anteriores $W_{ii} = 0$. La idea de la matriz de proximidad espacial puede ser generalizada a vecinos de mayor orden (vecinos de vecinos) construyéndose así las matrices $W^{(2)}, \cdots, W^{(n)}$. Se acostumbra a normalizar las filas de la matriz, es decir que la suma por fila de los $W_{ij}$ sea igual a uno.

Una vez obtenido el valor del coeficiente es necesario evaluar su significancia estadística. En otras palabras se requiere probar la hipótesis de aleatoriedad espacial con base en el valor observado. Para llevar a cabo esto es necesario establecer la correspondiente distribución de probabilidad de la estadística de prueba $I$. Bajo normalidad, es decir asumiendo que $Z(s_1), \cdots, Z(s_n)$ son *iid* con $Z_(s_i) \sim N(\mu, \sigma^2)$, la estadística

$$Z = \frac{I - \mathbb{E}(I)}{\sqrt{\mathbb{V}(I)}}$$

sigue una distribución normal estándar, en la que el valor esperado y la varianza están dados por

$$\mathbb{E}(I) = -\frac{1}{(n+1)},\ \mathbb{V}(I) = \frac{n^2 S_1 - n^2 S_2 + 3S_0^2}{(n^2-1)S_0^2} - \frac{1}{(n-1)^2},$$

donde

$$S_0 = \sum_{i \neq j}^{n} W_{ij},\quad S_1 = \sum_{i \neq j}^{n} (W_{ij} + W_{ji})^2,\quad S_2 = \sum_{i=1}^{n} (W_{i0} + W_{0i})^2,$$
$$W_{i0} = \sum_{j=1}^{n} W_{ij},\quad W_{0i} = \sum_{j=1}^{n} W_{ji}.$$

Otra posibilidad para establecer la significancia estadística, con menos supuestos, es llevando a cabo un test de permutación o de Monte Carlo como los descritos para la estadística de Mantel.

### 1.3.3. Variograma

El variograma, denotado por $2\gamma(h)$, se define como la varianza de la diferencia entre variables separadas por una distancia $h = \|s_i - s_j\|$. Asumiendo que $\mathbb{E}(Z(s)) = \mu$ se tiene

$$\begin{aligned} 2\gamma(h) &= \mathbb{V}(Z(s+h) - Z(s)) \\ &= \mathbb{E}(Z(s+h) - Z(s))^2. \end{aligned} \tag{1.2}$$

La mitad del variograma se llama semivariograma y caracteriza las propiedades de dependencia espacial de un fenómeno espacial. Esta función es usualmente empleada para tratar datos de un fenómeno con continuidad espacial (datos geoestadísticos). Usando el método de momentos se tiene que un estimador del semivariograma es

$$\bar{\gamma}(h) = \frac{1}{n(h)} \sum^{n(h)} (Z(s+h) - Z(s))^2, \tag{1.3}$$

donde $n(h)$ representa el número de parejas de sitios $(s_i, s_j)$ que se encuentran separados por una distancia $h$. En la práctica, debido a irregularidad en el muestreo y por ende en las distancias entre los sitios, se toman intervalos de distancia $\{[0, h], (h, 2h], (2h, 3h], \cdots\}$ y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo y no a una distancia $h$ específica. Obviamente el número de parejas de puntos n dentro de los intervalos no es constante. Para interpretar el semivariograma experimental se parte del criterio de que a menor distancia entre los sitios mayor similitud o correlación espacial entre las observaciones. Por ello en presencia de autocorrelación se espera que para valores de $h$ pequeños el semivariograma experimental tenga magnitudes menores a las que este toma cuando las distancias se incrementan.

Como se verá en el capítulo 4 la solución del problema de predicción espacial requiere del conocimiento de la estructura de autocorrelación para cualquier posible distancia entre sitios dentro del área de estudio. De la ecuación (1.3) es claro que el semivariograma muestral es calculado sólo para algunas distancias promedios particulares. Por ello se hace necesario el ajuste de modelos que generalicen la dependencia espacial para cualquier distancia (Figura 1.3). Existen diversos modelos teóricos de semivarianza que pueden ajustarse al semivariograma muestral. En Cressie (1993) se presenta una discusión respecto

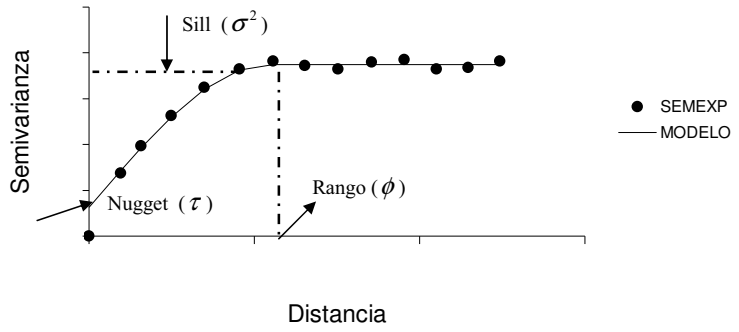



Figura 1.4: Comportamiento típico de un semivariograma acotado con una representación de los parámetros básicos. SEMEXP corresponde al semivariograma experimental y MODELO al ajuste de un modelo teórico.

a las características y condiciones que éstos deben cumplir. En general dichos modelos pueden dividirse en no acotados (lineal, logarítmico, potencial) y acotados (esférico, exponencial, Gaussiano) (Samper and Carrera, 1993). Los del segundo grupo garantizan que la covarianza de los incrementos es finita, por lo cual son ampliamente usados cuando hay evidencia de que presentan buen ajuste. La mayoría de modelos empleados para ajustar el semivariograma muestral, tienen tres parámetros en común (Figura 1.4) que son descritos a continuación:

- *Nugget* ($\tau$): Representa una discontinuidad puntual del semivariograma en el origen (Figura 1.3). Puede ser debido a errores de medición en la variable o a la escala de la misma. En algunas ocasiones puede ser indicativo de que parte de la estructura espacial se concentra a distancias inferiores a las observadas.

- *Sill* ($\sigma^2$): Es un estimador de la varianza de las variables del proceso. También puede definirse como el limite del semivariograma cuando la distancia $h$ tiende a infinito.

- *Rango*($\phi$). En términos prácticos corresponde a la distancia a partir de la cual dos observaciones son independientes. El rango se interpreta como la zona de influencia. Existen algunos modelos de semivariograma en los que no existe una distancia finita para la cual dos observaciones sean independientes; por ello se llama rango efectivo

a la distancia para la cual el semivariograma alcanza el 95 % de la meseta (*sill*).

## 1.4. Efectos de la correlación en inferencia estadística

Muchas métodos estadísticos están basados en el supuesto de que las variables aleatorias involucradas en la muestra son independientes. La violación de dicho supuesto tiene consecuencias en todos los procesos inferenciales. En esta sección se ilustra como la correlación entre las variables (por consiguiente la no independencia entre las mismas) afecta la estimación y la predicción en el modelo de regresión simple (sin covariables).

### 1.4.1. Efecto en la estimación

Sea $Y_1, \cdots, Y_n$ una muestra aleatoria de $Y \sim N(\mu, \sigma^2$. El estimador de $\mu$ es $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. El valor esperado y la varianza de este estimador son $\mu$ y $\sigma^2/n$, respectivamente. Ahora suponga que las variables $Y_1, \cdots, Y_n$ son correlacionadas y que $Cov(Y_i, Y_j) = \sigma^2\rho$. En este caso nuevamente el estimador de $\mu$ es $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ y su valor esperado es $\mu$, sin embargo la correlación aumenta (en este caso) la varianza del estimador. Veamos

$$\begin{aligned}
V(\bar{Y}) &= V(\frac{1}{n}\sum_{i=1}^{n} Y_i) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} Cov(Y_i, Y_j\right) \\
&= \frac{1}{n^2}\left[(\sigma^2 + \sigma^2\rho, \cdots, +\sigma^2\rho), \cdots, (\sigma^2 + \sigma^2\rho, \cdots, +\sigma^2\rho)\right] \\
&= \frac{1}{n^2}\left(n\sigma^2 + (n-1)\sigma^2\rho, \cdots, (n-1)\sigma^2\rho\right) \\
&= \frac{1}{n^2}\left(n\sigma^2 + n(n-1)\sigma^2\rho\right) \\
&= \frac{\sigma^2}{n}\left(1 + (n-1)\rho\right). \qquad (1.4)
\end{aligned}$$

Si $\rho > 0$ en (1.4), $V(\bar{Y}) > \sigma^2/n$, es decir la varianza del estimador de $\mu$ cuando hay correlación es mayor que la de este mismo cuando las variables son independientes.

### 1.4.2. Efecto en la predicción

Sean $Y_1, \cdots, Y_n$ variables aleatorias tales que $Y_i \sim N(\mu, \sigma^2)$ y $Cov(Y_i, Y_j) = \sigma^2\rho$. Un modelo lineal para representar este escenario es

$$\begin{aligned}\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} &= \begin{pmatrix} \mu \\ \vdots \\ \mu \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \\ &= \mu\mathbf{1} + \boldsymbol{\epsilon}, \end{aligned} \tag{1.5}$$

donde

$$V(\boldsymbol{\epsilon}) = \Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

Suponga que se quiere predecir una nueva observación $Y_0$. Definiendo el predictor por

$$Y_0^* = \sum_{i=1}^{n} \lambda_i Y_i, \tag{1.6}$$

los pesos $\lambda_i$ se obtienen de tal forma que se minimice la esperanza de una función de pérdida. Bajo pérdida cuadrática, el mejor predictor (lineal en este caso), será el que minimiza la función

$$\min_{\lambda_1,\ldots,\lambda_n} \mathbb{E}(Y_0^* - Y_0)^2, \text{sujeto a } \mathbb{E}(Y_0^*) = \mathbb{E}(Y_0).$$

De acuerdo con lo anterior, la función objetivo es

$$\min_{\boldsymbol{\lambda},m} \mathbb{V}(Y_0^* - Y_0), \text{ sujeto a } \sum_{i=1}^{n} \lambda_i = 1.$$

Se tiene que $Cov(\mathbf{Y}, Y_0) = \sigma^2\rho\mathbf{1}$. Desarrollando la varianza e incluyendo un multiplicador de Lagrange para la condición de insesgadez la función a optimizar es

$$\begin{aligned}
&\min_{\boldsymbol{\lambda},m} \mathbb{V}(Y_0^*) + \mathbb{V}(Y_0) - 2Cov(Y_0^*, Y_0) - 2m(\sum_{i=1}^{n} \lambda_i - 1) \\
&\min_{\boldsymbol{\lambda},m} \mathbb{V}(\boldsymbol{\lambda}^T\mathbf{Y}) + \sigma^2 - 2Cov(\boldsymbol{\lambda}^T\mathbf{Y}, Y_0) - 2m(\boldsymbol{\lambda}^T\mathbf{1} - 1) \\
&\min_{\boldsymbol{\lambda},m} \boldsymbol{\lambda}^T\Sigma\boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}^T\mathbf{c} - 2m(\boldsymbol{\lambda}^T\mathbf{1} - 1),\ \mathbf{c} = \begin{pmatrix} \sigma^2\rho \\ \vdots \\ \sigma^2\rho \end{pmatrix} = \sigma^2\rho\mathbf{1}.
\end{aligned}$$

Tomando derivadas respecto a $\lambda$ y $m$ se obtiene el siguiente sistema

$$\begin{aligned}
\Sigma\boldsymbol{\lambda} - \mathbf{c} - m\mathbf{1} &= 0 \\
\boldsymbol{\lambda}^T\mathbf{1} - 1 &= 0.
\end{aligned} \tag{1.7}$$

Despejando $\boldsymbol{\lambda}$ en la primera ecuación del sistema (1.7), se obtiene

$$\boldsymbol{\lambda} = \Sigma^{-1}(\mathbf{c} + m\mathbf{1}). \tag{1.8}$$

Reemplazando esta expresión en la segunda ecuación del sistema (1.7) se encuentra

$$\begin{aligned}
&(\Sigma^{-1}(\mathbf{c} + m\mathbf{1}))^T\mathbf{1} = 1 \\
&(\Sigma^{-1}\mathbf{c} + \Sigma^{-1}m\mathbf{1})^T\mathbf{1} = 1 \\
&\mathbf{1}^T(\Sigma^{-1}\mathbf{c}) + \mathbf{1}^T(\Sigma^{-1}m\mathbf{1}) = 1 \\
&\mathbf{1}^T(\Sigma^{-1}m\mathbf{1}) = 1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c}) \\
&m = \left(1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c})\right)(\mathbf{1}^T\Sigma^{-1}\mathbf{1})^{-1} \\
&m = \frac{1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}
\end{aligned} \tag{1.9}$$

Sustituyendo (1.9) en la ecuación (1.8) se obtiene

$$\begin{aligned}
\boldsymbol{\lambda} &= \Sigma^{-1}\left(\mathbf{c} + \frac{1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\mathbf{1}\right) \\
\boldsymbol{\lambda}^T &= \left(\mathbf{c} + \mathbf{1}\frac{1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\right)^T (\Sigma^{-1})^T \\
\boldsymbol{\lambda}^T &= \left(\mathbf{c} + \mathbf{1}\frac{1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\right)^T \Sigma^{-1}. \qquad (1.10)
\end{aligned}$$

De acuerdo con la solución obtenida en (1.10), el predictor en (1.6) es definido por

$$\begin{aligned}
Y_0^* &= \sum_{i=1}^{n} \lambda_i Y_i \\
&= \boldsymbol{\lambda}^T\mathbf{Y} \\
&= \left[\left(\mathbf{c} + \mathbf{1}\frac{1 - \mathbf{1}^T(\Sigma^{-1}\mathbf{c})}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}\right)^T \Sigma^{-1}\right]\mathbf{Y}
\end{aligned}$$

Haciendo algunas manipulaiones de álgebra se obtiene que

$$Y_0^* = \hat{\mu} + \mathbf{c}^T\Sigma^{-1}\left(\mathbf{Y} - \mathbf{1}\hat{\mu}\right), \qquad (1.11)$$

donde $\hat{\mu}$ es el estimador de mínimos cuadrados generalizados de $\mu$ en la ecuación (1.5). La varianza del predictor en (1.11) está dada por

$$\sigma_p^2 = \sigma^2 - \mathbf{c}^T\Sigma^1\mathbf{c} + \frac{(1 - \mathbf{1}^T\Sigma^{-1}\mathbf{c})^2}{(\mathbf{1}^T\Sigma^{-1}\mathbf{1})}. \qquad (1.12)$$

**Observación**

- Del modelo lineal general $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ se tiene que el estimador de mínimos cuadrados generalizados del vector de parámetros es $\boldsymbol{\beta} = \left(\mathbf{X}^T\Sigma^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\Sigma^{-1}\mathbf{Y}\right)$. Definiendo $\mathbf{X} = \mathbf{1}$ y $\boldsymbol{\beta} = \mu$, se obtiene que $\hat{\mu} = \left(\mathbf{1}^T\Sigma^{-1}\mathbf{1}\right)^{-1}\left(\mathbf{1}^T\Sigma^{-1}\mathbf{Y}\right)$.

Ahora considérese el caso de predicción teniendo una muestra aleatoria. Sean $Y_1, \cdots, Y_n$ variables aleatorias independientes e idénticamente distribuidas, con $Y_i \sim N(\mu, \sigma^2)$. Planteando el mismo predictor dado en (1.6) y reemplazando $\Sigma = \sigma^2\mathbf{I}$ y $\mathbf{c} = \mathbf{0}$ en las ecuaciones

(1.8) y (1.9), se encuentra que $m = (\mathbf{1}^T(\sigma^2\mathbf{I})^{-1}\mathbf{1})^{-1}$ y que

$$\begin{aligned}\boldsymbol{\lambda} &= (\sigma^2\mathbf{I})^{-1}(\mathbf{1}^T(\sigma^2\mathbf{I})^{-1}\mathbf{1})^{-1}\mathbf{1} \\ &= \begin{pmatrix} 1/\sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma^2 \end{pmatrix}\begin{pmatrix} \sigma^2/n \\ \vdots \\ \sigma^2/n \end{pmatrix} = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix}. \end{aligned} \tag{1.13}$$

Al sustituir (1.13) en (1.6) se obtiene

$$\begin{aligned} Y_0^* &= \sum_{i=1}^{n} \lambda_i Y_i \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}. \end{aligned} \tag{1.14}$$

Tomando $\Sigma = \sigma^2\mathbf{I}$ y $\mathbf{c} = \mathbf{0}$ en (1.12) se obtiene que $\sigma_p^2 = \sigma^2(1 + 1/n)$, es decir la varianza de predicción del modelo bajo independencia.

# Capítulo 2

# Patrones Puntuales

## 2.1. Tipos de Patrones

## 2.2. Estimación Kernel de la intensidad

## 2.3. Métodos basados en cuadrantes

## 2.4. Métodos basados en distancias

## 2.5. Modelos

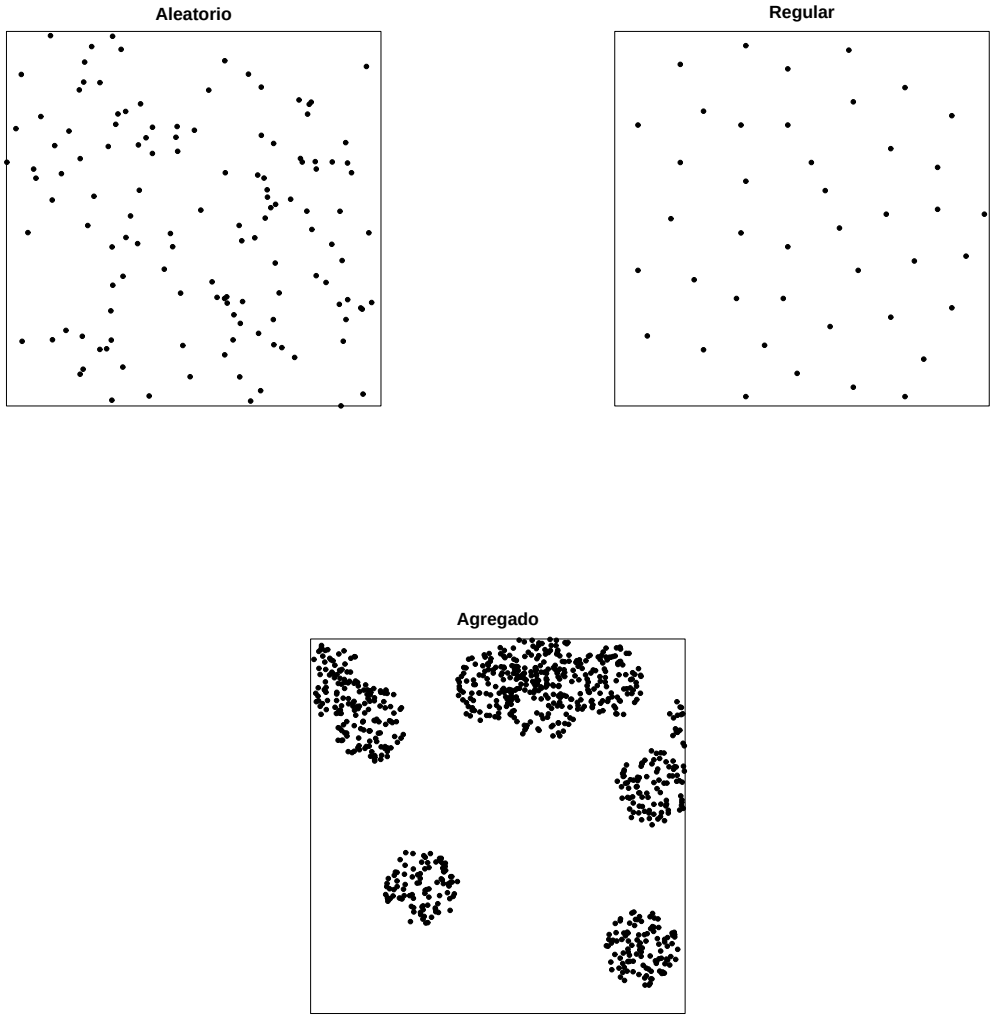



Figura 2.1: Ejemplos simulados de patrones puntuales con distribución aleatoria (*izquierda*), regular (*derecha*) y agregada (*centro*).

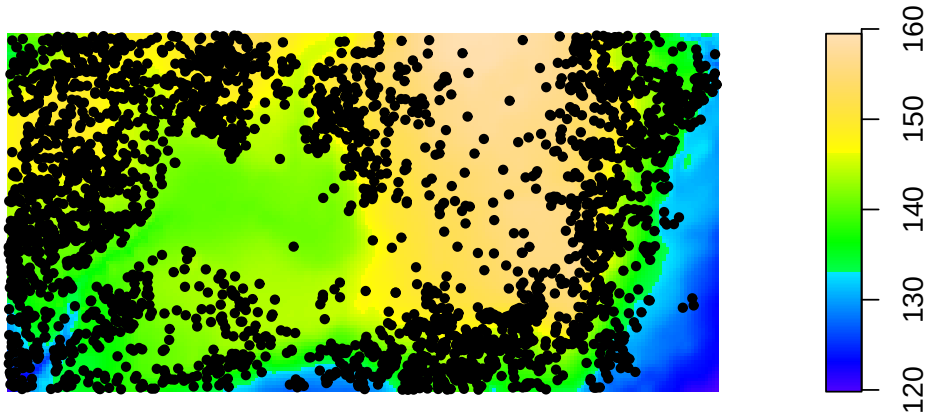



Figura 2.2: Ejemplo de patrón marcado.

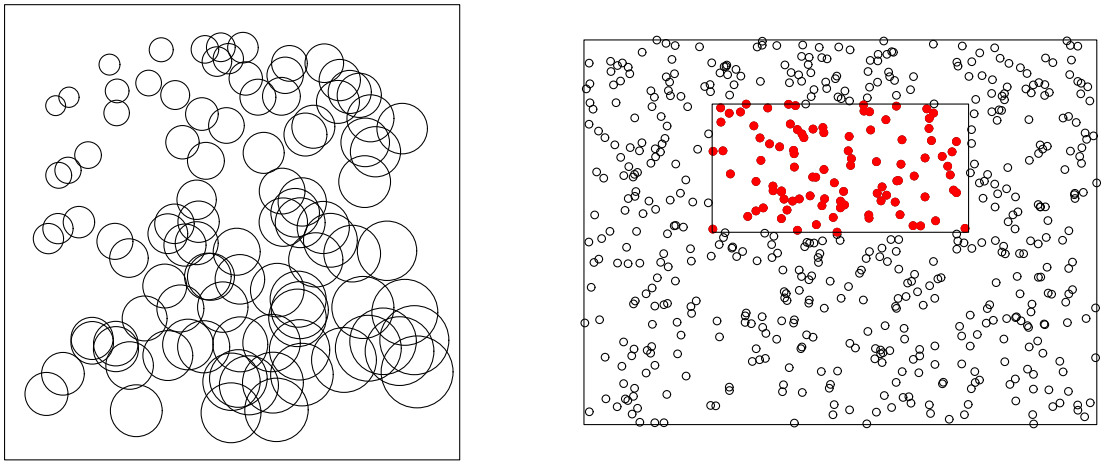


Figura 2.3: Ejemplos de patrones que incluyen covariables (*izquierda*) y del efecto de borde (*derecha*). En la figura de la izquierda la escala de valores corresponde a la elevación en metros

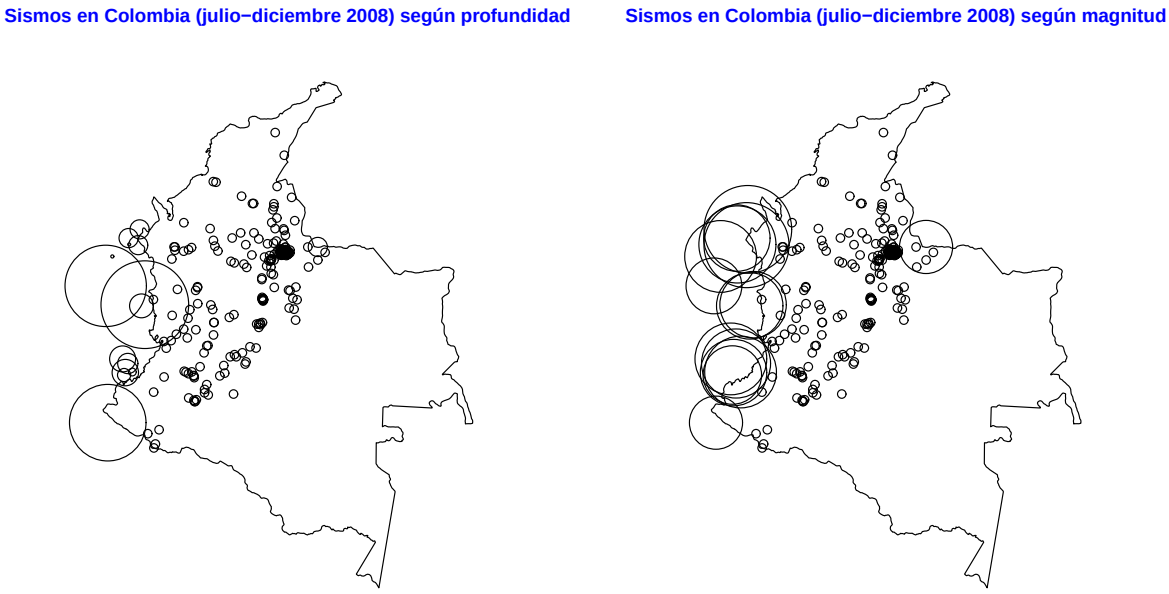



Figura 2.4: Marcas asociadas a los datos de sismos en Colombia

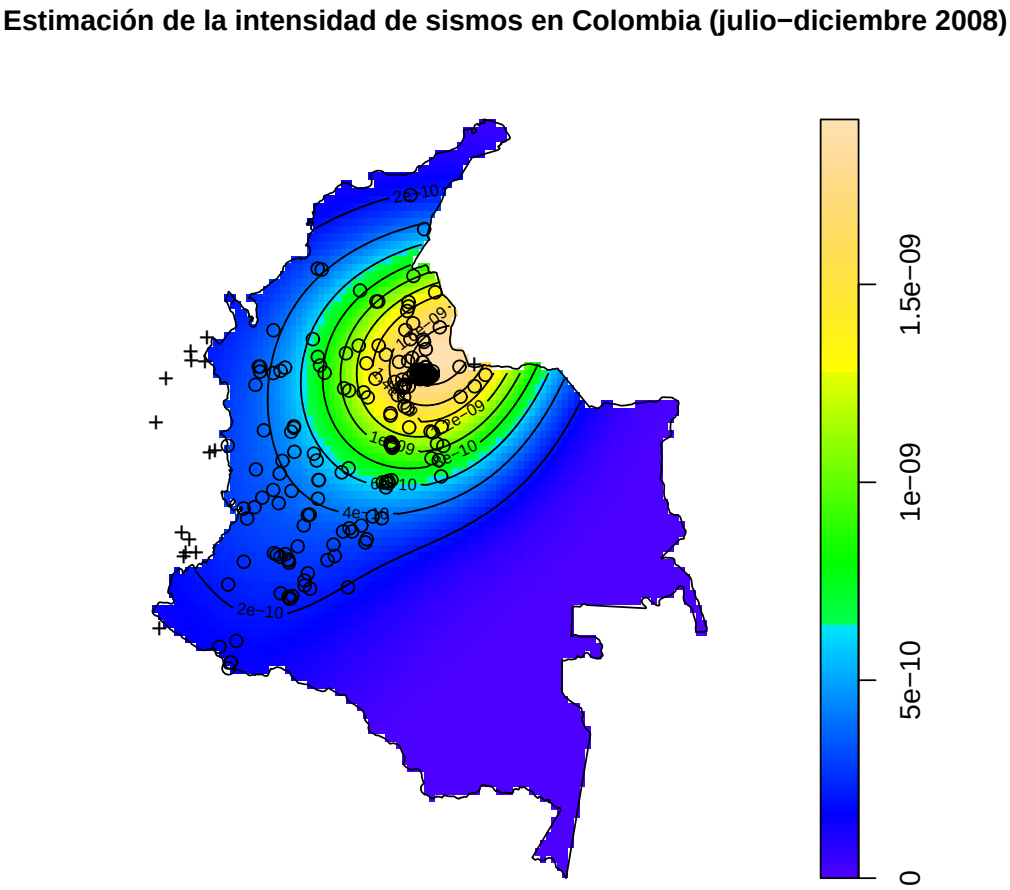



Figura 2.5: Estimación de la intensidad de sismos en Colombia (Julio-Diciembre 2008).

# Capítulo 3

# Datos de Areas

## 3.1. Visualización de datos de áreas

La forma más usual de visualización de datos de agregaciones espaciales es a través de mapas coropléticos o temáticos donde las áreas son sombreadas (coloreadas) de acuerdo con sus valores en la variable de interés. El número de clases y las correspondientes clases pueden ser basadas en diferentes criterios.

### 3.1.1. Mapas de datos originales y tasas

Los sistemas de información geográfica usualmente disponen de tres métodos de corte de los valores de la variable: Intervalos iguales, cuantilas (cuartiles, deciles, percentiles) y desviaciones estándar. En el caso de intervalos iguales se toma la diferencia entre el mínimo y el máximo de la variable (rango) y se divide por el número de clases o intervalos deseados, obteniéndose así la amplitud de los mismos. Este procedimiento es similar a como se hacía hace unos años, cuando no había mucho acceso a computadores, para obtener los intervalos en la construcción de los histogramas de frecuencia. Si la distribución de frecuencias es asimétrica, esta división deja una gran parte del mapa sombreado con el mismo color y por consiguiente impide una buena descripción del fenómeno de estudio. De otro lado pueden emplearse las cuantilas. Tradicionalmente se usan la mediana y los cuartiles. Solo

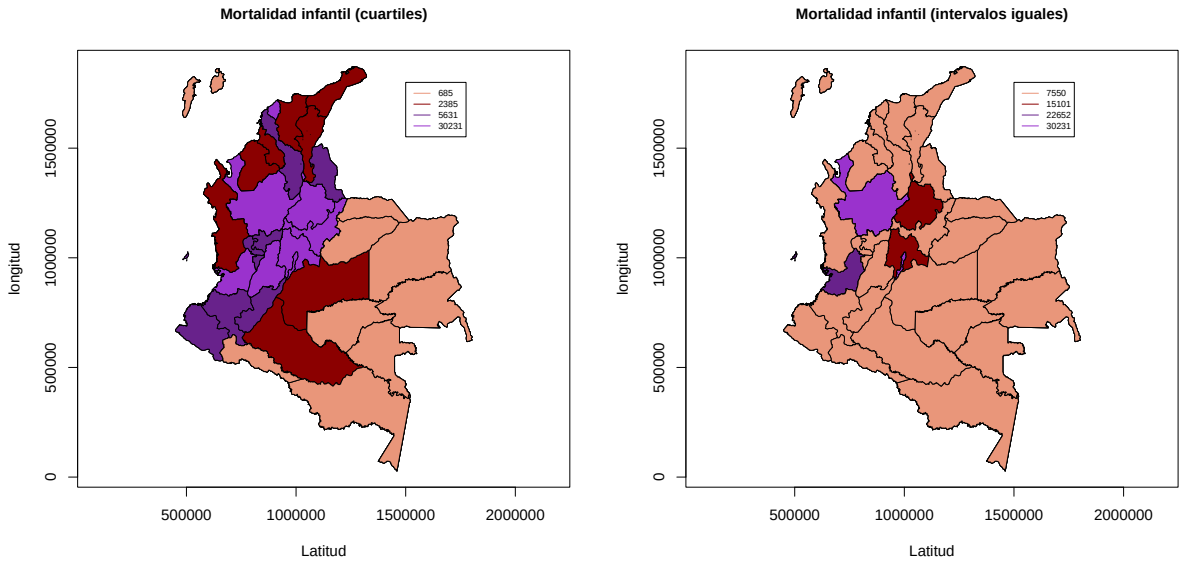



Figura 3.1: Mapas temáticos de mortalidad infantil en Colombia en 2003.

en aquellos casos en que haya un gran número de áreas en el mapa podría pensarse en el uso de deciles o percentiles. En estos casos hay un número igual de áreas sombreadas con cada color. La desventaja de este procedimiento es que los valores extremos son enmascarados, puesto que caen dentro de alguna categoría grande, lo cual en algunas situaciones prácticas podría impedir la detección de zonas críticas, por ejemplo en el caso de que las variables sean tasas. Finalmente en el caso de desviaciones estándar los límites de los intervalos son construidos con base en desviaciones estándar respecto a la media. Este procedimiento es válido sólo en aquellas situaciones en los que el supuesto de simetría pueda ser razonable. Las diferencias entre los distintos modos de construcción de mapas son ilustradas con base en información mortalidad infantil (número total y tasa) por departamentos de Colombia en el año 2003.

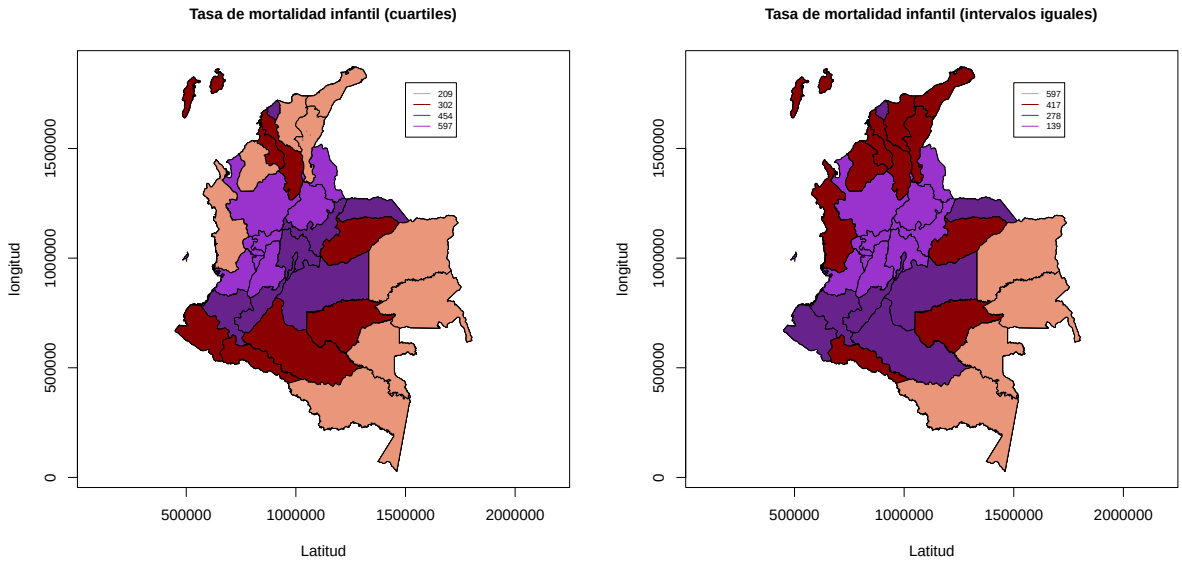



Figura 3.2: Mapas temáticos de tasas de mortalidad infantil en Colombia en 2003.

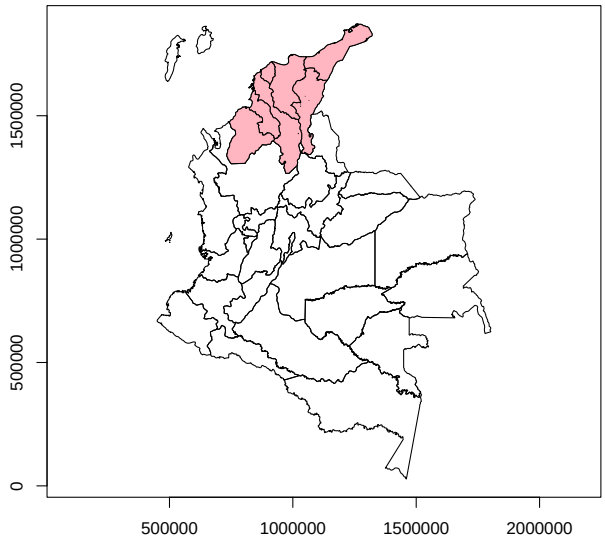



Figura 3.3: Las áreas sombreadas corresponden a los departamentos de la costa Caribe de Colombia (sin incluir San Andrés y Providencia).

## 3.2. Exploración de datos de áreas

### 3.2.1. Medidas de proximidad espacial

### 3.2.2. Medias móviles

### 3.2.3. Estimación kernel

## 3.3. Correlación espacial

### 3.3.1. I de Moran y C de Geary

### 3.3.2. Correlograma

## 3.4. Modelos autorregresivos espaciales

En el contexto de series temporales los modelos autorregresivos (AR) indican que existe una relación lineal entre lo que ocurre con la variable en un periodo de tiempo $t$ y lo que ocurrió con ella en periodos anteriores, es decir si $\{Z(t), t \in \mathbb{R}\}$ representa una serie de tiempo (estacionaria, con media cero), un modelo AR($p$), indica que se tiene el modelo $Z(t) = \rho_1 Z(t-1) + \cdots + \rho_p Z(t-p) + \epsilon(t)$, con $\epsilon(t)$ un ruido blanco. El análogo a dicho modelo en el contexto espacial se presenta cuando para el proceso espacial $\{Z(s), s \in D \subset \mathbb{R}^2\}$ se plantea el modelo lineal que relaciona $Z(s_i)$ con las variables en los sitos vecinos ($\mathbf{Z}(s)_{-i}$), donde $\mathbf{Z}(s)_{-i}$ representa le vector de de todas las variables excepto $Z(s_i)$. Existen varios tipos de modelos autorregresivos espaciales. Los dos más usados en la practica son

- Modelo autorregresivo simultáneo
- Modelo autorregresivo condicional

A continuación se describen dichos modelos y se muestra como hacer estimación y pruebas de hipótesis en ambos casos.

### 3.4.1. Modelo autorregresivo simultáneo (SAR)

En este modelo se asigna una estructura de autocorrelación espacial a los residuales de un modelo de regresión lineal, usando para ello un matriz de proximidad espacial entre las áreas. Este también es conocido como *modelo de errores correlacionados* o de *variables espacialmente rezagadas*. El modelo SAR es planteado como

$$\begin{aligned} \mathbf{Z}(s) &= \mathbf{X}(s)\boldsymbol{\beta} + \mathbf{U}(s), \\ \mathbf{U}(s) &= \rho\mathbf{W}\mathbf{U}(s) + \boldsymbol{\epsilon}(s), \end{aligned} \tag{3.1}$$

donde $\mathbb{E}(\boldsymbol{\epsilon}(s)) = \mathbf{0}$ y $\mathbb{V}(\boldsymbol{\epsilon}(s)) = \sigma^2\mathbf{I}$. En la ecuación (3.1) $\mathbf{W}$ es una matriz de proximidad (ver sección 3.2.1). Es claro que si $\rho = 0$ se tiene un modelo de regresión tradicional. El modelo también puede ser expresado como

$$\begin{aligned} \mathbf{Z}(s) &= \mathbf{X}(s)\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{U}(s) + \boldsymbol{\epsilon}(s) \\ &= \mathbf{X}(s)\boldsymbol{\beta} + \rho\mathbf{W}(\mathbf{Z}(s) - \mathbf{X}(s)\boldsymbol{\beta}) + \boldsymbol{\epsilon}(s) \\ &= \mathbf{X}(s)\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{Z}(s) + \rho\mathbf{W}\mathbf{X}(s)\boldsymbol{\beta} + \boldsymbol{\epsilon}(s) \qquad (3.2) \\ \mathbf{Z}(s) - \rho\mathbf{W}\mathbf{Z}(s) &= \mathbf{X}(s)\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{X}(s)\boldsymbol{\beta} + \boldsymbol{\epsilon}(s) \\ \boldsymbol{\epsilon}(s) &= (\mathbf{I} - \rho\mathbf{W})(\mathbf{Z}(s) - \mathbf{X}(s)\boldsymbol{\beta}). \qquad (3.3) \end{aligned}$$

De (3.1), se tiene que $\mathbb{E}(\mathbf{U}(s)) = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbb{E}(\boldsymbol{\epsilon}(s)) = \mathbf{0}$ y por consiguiente $\mathbb{E}(\mathbf{Z}(s)) = \mathbf{X}(s)\boldsymbol{\beta}$. Además la varianza de $\mathbf{Z}(s)$ está dada por

$$\begin{aligned} \Sigma_{SAR} &= \mathbb{V}(\mathbf{Z}(s)) \\ &= \mathbb{V}(\mathbf{U}(s)) \\ &= \mathbb{V}\left[(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}(s)\right] \\ &= (\mathbf{I} - \rho\mathbf{W})^{-1}\sigma^2\mathbf{I}((\mathbf{I} - \rho\mathbf{W})^{-1})^T \\ &= (\mathbf{I} - \rho\mathbf{W})^{-1}\sigma^2\mathbf{I}(\mathbf{I} - \rho\mathbf{W}^T)^{-1} \\ &= \sigma^2\left((\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{I} - \rho\mathbf{W}^T)^{-1}\right) \end{aligned} \tag{3.4}$$

En la ecuación (3.2) los términos $\rho\mathbf{W}\mathbf{Z}(s)$ y $\rho\mathbf{W}\mathbf{X}(s)\boldsymbol{\beta}$ son llamados variables espacialmente rezagadas y de ahí que el modelo SAR sea conocido con este nombre. Dos simplificaciones del modelo (3.2) son $\mathbf{Z}(s) = \mathbf{X}(s)\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{Z}(s) + \boldsymbol{\epsilon}(s)$ y $\mathbf{Z}(s) = \rho\mathbf{W}\mathbf{Z}(s) + \boldsymbol{\epsilon}(s)$. El

segundo de ellos es conocido como *modelo autorregresivo puro*. El modelo (3.1) está bien especificado si la matriz $(\mathbf{I} - \rho\mathbf{W})$ es invertible. Esto implica que se pongan algunas condiciones sobre $\mathbf{W}$ y sobre $\rho$. Si la matriz $\mathbf{W}$ es definida de tal forma que las filas sumen uno (dividiendo cada $w_{ij}$ por la suma $\sum_j w_{ij}$), se garantiza que $\rho < 1$.

**Estimación e inferencia en el modelo SAR**

En el caso de que se haya definido una matriz de proximidad entre las áreas ($\mathbf{W}$) y asumiendo conocido $\rho$ puede usarse el método de mínimos cuadrados generalizados para hacer la estimación del vector $\boldsymbol{\beta}$ y del parámetro $\sigma^2$ en el modelo (3.1). Los correspondientes estimadores están dados por

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}(s)^T \Sigma_{SAR}^{-1} \mathbf{X}(s))^{-1} \mathbf{X}(s)^T \Sigma_{SAR}^{-1} \mathbf{Z}(s),$$
$$\hat{\sigma}^2 = \frac{(\mathbf{Z}(s) - \mathbf{X}(s)\hat{\boldsymbol{\beta}})^T \Sigma_{SAR}^{-1} (\mathbf{Z}(s) - \mathbf{X}(s)\hat{\boldsymbol{\beta}})}{n - k},$$

donde $\Sigma_{SAR}^{-1}$ está definido como en (3.4).

### 3.4.2. Modelo autorregresivo condicional (CAR)

# Capítulo 4

# Geoestadística

# Bibliografía

Cressie, N. (1993). *Statistic for spatial data.* John Wiley & Sons.

Samper, F. and J. Carrera (1993). *Geoestadística. Aplicaciones a la hidrogeología subterránea.* UPC Barcelona: Centro Internacional de Métodos Numéricos en Ingeniería.