

EPG3343 - Seminario de Estadística III

Clase 2

Jonathan Acosta

Pontificia Universidad Católica de Chile

Segundo Semestre, 2022



1 Introducción

- Análisis Exploratorio
 - Medidas sobre Lattice
 - Estadísticos de Moran y Geary
 - Variograma Empírico

1 Introducción

- Análisis Exploratorio
 - Medidas sobre Lattice
 - Estadísticos de Moran y Geary
 - Variograma Empírico

Principio (Primera Ley de la geografía)

Cantidades cercanas (vecinas) tienden a ser más parecidas que las cantidades que están más apartadas

Tipos de Datos Espaciales:

- 1 Datos Geoestadísticos: El dominio es un conjunto continuo y fijo.
- 2 Datos sobre Grillas: El dominio es fijo y contable.
- 3 Patrones de Puntos: El dominio es aleatorio.

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).
- (ii) Construyendo del Histograma y Box-Plot de la variable atributo.

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).
- (ii) Construyendo del Histograma y Box-Plot de la variable atributo.
- (iii) Graficar las coordenadas.

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).
- (ii) Construyendo del Histograma y Box-Plot de la variable atributo.
- (iii) Graficar las coordenadas.
- (iv) Graficar $Z(\mathbf{s})$ versus \mathbf{s} .

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).
- (ii) Construyendo del Histograma y Box-Plot de la variable atributo.
- (iii) Graficar las coordenadas.
- (iv) Graficar $Z(\mathbf{s})$ versus \mathbf{s} .
- (v) Detectar *outlier* y *datos influyentes*.

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).
- (ii) Construyendo del Histograma y Box-Plot de la variable atributo.
- (iii) Graficar las coordenadas.
- (iv) Graficar $Z(\mathbf{s})$ versus \mathbf{s} .
- (v) Detectar *outlier* y *datos influyentes*.
- (vi) Obtener las curvas de nivel.

Suponga que tiene una variable atributo Z Georreferenciada. Dada una muestra de esta variable, $Z(\mathbf{s}_1), Z(\mathbf{s}_2), \dots, Z(\mathbf{s}_n)$, podemos resumir (describir) la información contenida en ella:

- (i) Calculando medidas de Tendencia Central, Dispersión y Forma (Asimetría y Curtosis).
- (ii) Construyendo del Histograma y Box-Plot de la variable atributo.
- (iii) Graficar las coordenadas.
- (iv) Graficar $Z(\mathbf{s})$ versus \mathbf{s} .
- (v) Detectar *outlier* y *datos influyentes*.
- (vi) Obtener las curvas de nivel.
- (vii) Detectar Cluster.

Mantel (1967) definió un índice para detectar clusters.

Mantel (1967) definió un índice para detectar clusters.

- Este índice está construido para datos espacio-temporales.
- Puede ser aplicable a datos espaciales.

Mantel (1967) definió un índice para detectar clusters.

- Este índice está construido para datos espacio-temporales.
- Puede ser aplicable a datos espaciales.

Sean $T(\mathbf{s}_1), T(\mathbf{s}_2), \dots, T(\mathbf{s}_n)$ los tiempos en los cuales los eventos de interés ocurren. Considere :

$$W_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|; \quad U_{ij} = |T(\mathbf{s}_i) - T(\mathbf{s}_j)|$$

Nociones de Estadística Descriptiva

Mantel (1967) definió un índice para detectar clusters.

- Este índice está construido para datos espacio-temporales.
- Puede ser aplicable a datos espaciales.

Sean $T(\mathbf{s}_1), T(\mathbf{s}_2), \dots, T(\mathbf{s}_n)$ los tiempos en los cuales los eventos de interés ocurren. Considere :

$$W_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|; \quad U_{ij} = |T(\mathbf{s}_i) - T(\mathbf{s}_j)|$$

Mantel sugirió los estadísticos:

$$M_1 = \sum_{i=1}^{n-1} \sum_{j=i+1}^n W_{ij} U_{ij}; \quad M_2 = \sum_{i=1}^n \sum_{j=1}^n W_{ij} U_{ij}$$

Nociones de Estadística Descriptiva

Mantel (1967) definió un índice para detectar clusters.

- Este índice está construido para datos espacio-temporales.
- Puede ser aplicable a datos espaciales.

Sean $T(\mathbf{s}_1), T(\mathbf{s}_2), \dots, T(\mathbf{s}_n)$ los tiempos en los cuales los eventos de interés ocurren. Considere :

$$W_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|; \quad U_{ij} = |T(\mathbf{s}_i) - T(\mathbf{s}_j)|$$

Mantel sugirió los estadísticos:

$$M_1 = \sum_{i=1}^{n-1} \sum_{j=i+1}^n W_{ij} U_{ij}; \quad M_2 = \sum_{i=1}^n \sum_{j=1}^n W_{ij} U_{ij}$$

Obs: Para procesos Espaciales $U_{ij} = |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|$

La distribución asintótica de M_2 es aproximadamente normal, es decir,

$$\frac{M_2 - \mathbb{E}[M_2]}{\sqrt{\mathbb{V}[M_2]}} \longrightarrow \mathcal{N}(0, 1) \quad \text{cuando } n \rightarrow \infty$$

Luego, considere el test

$$H_0 : \text{cov} [Z(\mathbf{s}_i), Z(\mathbf{s}_j)] = 0 \quad \forall i \neq j$$

\therefore Se rechaza H_0 si

$$|Z_{obs}| = \left| \frac{M_{2(obs)} - \mathbb{E}[M_2]}{\sqrt{\mathbb{V}[M_2]}} \right| > z_{1-\frac{\alpha}{2}}$$

Suponga que el atributo Z es binario y esta definido sobre una grilla, obteniendo

$$\omega_{ij} = \begin{cases} 1 & \text{Si los sitios } i \text{ y } j \text{ están conectados} \\ 0 & \text{en caso contrario} \end{cases}$$

$$Z(\mathbf{s}_i) = \begin{cases} 1 & \text{Si el evento ocurre en el sitio } i. \\ 0 & \text{en caso contrario} \end{cases}$$

Supuesto: $\mathbb{P}[Z(\mathbf{s}_i) = 1] = p, \quad 0 \leq p \leq 1.$

Además, $\mathbb{E}[Z(\mathbf{s}_i)^k] = p$

Asumiendo que $Z(\mathbf{s}) = 1$ se colorea de color negro, mientras que $Z(\mathbf{s}) = 0$ se colorea de color blanco. Considere los estadísticos

$$BB = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} Z(\mathbf{s}_i) Z(\mathbf{s}_j); \quad BW = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$$

Asumiendo que $Z(\mathbf{s}) = 1$ se colorea de color negro, mientras que $Z(\mathbf{s}) = 0$ se colorea de color blanco. Considere los estadísticos

$$BB = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} Z(\mathbf{s}_i) Z(\mathbf{s}_j); \quad BW = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$$

Observaciones:

- BB es un caso particular del Estadístico de Mantel con $U_{ij} = Z(\mathbf{s}_i)Z(\mathbf{s}_j)$.
- La cantidad de puntos negros es una variable aleatoria binomial con parámetros n y p .

Bajo H_0 , se tiene que

$$\mathbb{E}[Z(\mathbf{s}_i)Z(\mathbf{s}_j)] = p^2; \quad \mathbb{V}[Z(\mathbf{s}_i)Z(\mathbf{s}_j)] = p^2 - p^4$$

Medidas sobre Lattice

Bajo H_0 , se tiene que

$$\mathbb{E}[Z(\mathbf{s}_i)Z(\mathbf{s}_j)] = p^2; \quad \mathbb{V}[Z(\mathbf{s}_i)Z(\mathbf{s}_j)] = p^2 - p^4$$

Finalmente,

$$\mathbb{E}[BB] = \frac{p^2}{2} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij}; \quad \mathbb{V}[BB] = \frac{1}{4}p^2(1-p)[(1-p)S_1 + pS_2]$$

donde,

$$S_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\omega_{ij} + \omega_{ji})^2; \quad S_2 = \sum_{i=1}^n \left[\sum_{j=1}^n \omega_{ij} + \sum_{j=1}^n \omega_{ji} \right]$$

Obs: Los calculos son similares para BW .

Estadísticos de Moran y Geary

Si Z es un atributo continuo tal que su media no varía espacialmente ($\mathbb{E}[Z(\mathbf{s})] = \mu$), existen varias formas de medir la cercanía entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$.

Estadísticos de Moran y Geary

Si Z es un atributo continuo tal que su media no varía espacialmente ($\mathbb{E}[Z(\mathbf{s})] = \mu$), existen varias formas de medir la cercanía entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$.

$$\begin{aligned}U_{ij}^{(1)} &= [Z(\mathbf{s}_i) - \mu] [Z(\mathbf{s}_j) - \mu] \\U_{ij}^{(2)} &= |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)| \\U_{ij}^{(3)} &= [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}] \\U_{ij}^{(4)} &= (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2\end{aligned}$$

Estadísticos de Moran y Geary

Si Z es un atributo continuo tal que su media no varía espacialmente ($\mathbb{E}[Z(\mathbf{s})] = \mu$), existen varias formas de medir la cercanía entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$.

$$U_{ij}^{(1)} = [Z(\mathbf{s}_i) - \mu] [Z(\mathbf{s}_j) - \mu]$$

$$U_{ij}^{(2)} = |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|$$

$$U_{ij}^{(3)} = [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}]$$

$$U_{ij}^{(4)} = (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$$

- $U_{ij}^{(1)}$ se descarta porque en la práctica μ es desconocido.

Estadísticos de Moran y Geary

Si Z es un atributo continuo tal que su media no varía espacialmente ($\mathbb{E}[Z(\mathbf{s})] = \mu$), existen varias formas de medir la cercanía entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$.

$$U_{ij}^{(1)} = [Z(\mathbf{s}_i) - \mu] [Z(\mathbf{s}_j) - \mu]$$

$$U_{ij}^{(2)} = |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|$$

$$U_{ij}^{(3)} = [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}]$$

$$U_{ij}^{(4)} = (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$$

- $U_{ij}^{(1)}$ se descarta porque en la práctica μ es desconocido.
- $U_{ij}^{(2)}$ se descarta por ser matemáticamente intratable.

Si se escoge $U_{ij}^{(3)}$, este se podría estandarizar por la varianza muestral, de modo que el estadístico:

$$\frac{(n-1) [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}]}{\sum_{i=1}^n [Z(\mathbf{s}_i) - \bar{Z}]^2}$$

es un estimador de la correlación entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$.

Si se escoge $U_{ij}^{(3)}$, este se podría estandarizar por la varianza muestral, de modo que el estadístico:

$$\frac{(n-1) [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}]}{\sum_{i=1}^n [Z(\mathbf{s}_i) - \bar{Z}]^2}$$

es un estimador de la correlación entre $Z(\mathbf{s}_i)$ y $Z(\mathbf{s}_j)$.

Inconvenientes:

- 1 No usa los pesos ω_{ij} .
- 2 Propiedades estadísticas Pobres.

Moran (1950) define utilizando $U_{ij}^{(3)}$ y los pesos ω_{ij}

$$I = \frac{n}{(n-1)S^2\omega_{..}} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}]$$

donde $\omega_{..} = \sum_{i=1}^n \sum_{j=1}^n \omega_{ij}$ y $S^2 = \frac{1}{n-1} \sum_{i=1}^n [Z(\mathbf{s}_i) - \bar{Z}]^2$

Moran (1950) define utilizando $U_{ij}^{(3)}$ y los pesos ω_{ij}

$$I = \frac{n}{(n-1)S^2\omega_{..}} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} [Z(\mathbf{s}_i) - \bar{Z}] [Z(\mathbf{s}_j) - \bar{Z}]$$

donde $\omega_{..} = \sum_{i=1}^n \sum_{j=1}^n \omega_{ij}$ y $S^2 = \frac{1}{n-1} \sum_{i=1}^n [Z(\mathbf{s}_i) - \bar{Z}]^2$

Similarmente, Geary (1954) utiliza $U_{ij}^{(4)}$ y los pesos ω_{ij} y define

$$C = \frac{1}{2S^2\omega_{..}} \sum_{i=1}^n \sum_{j=1}^n \omega_{ij} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2$$

Estadísticos de Moran y Geary

Bajo Normalidad (Cliff & Ord, 1981)

$$\mathbb{E}[I] = -\frac{1}{n-1}; \quad \mathbb{E}[C] = 1$$

Bajo Normalidad (Cliff & Ord, 1981)

$$\mathbb{E}[I] = -\frac{1}{n-1}; \quad \mathbb{E}[C] = 1$$

Interpretación de I :

- Si $I > \mathbb{E}[I]$, entonces un sitio tiende a estar conectado a los sitios que tienen atributos similares.
La correlación espacial es positiva y aumenta a medida que $|I - \mathbb{E}[I]|$ crece.
- Si $I < \mathbb{E}[I]$, entonces los valores de sitios conectados tienden a ser diferentes.

Bajo Normalidad (Cliff & Ord, 1981)

$$\mathbb{E}[I] = -\frac{1}{n-1}; \quad \mathbb{E}[C] = 1$$

Interpretación de I :

- Si $I > \mathbb{E}[I]$, entonces un sitio tiende a estar conectado a los sitios que tienen atributos similares.
La correlación espacial es positiva y aumenta a medida que $|I - \mathbb{E}[I]|$ crece.
- Si $I < \mathbb{E}[I]$, entonces los valores de sitios conectados tienden a ser diferentes.

Interpretación de C : Es opuesta a la interpretación de I

La librería *spdep* contiene las funciones:

`moran.test`

`geary.test`

Las cuales sirven para calcular los estadísticos de Moran y Geary respectivamente.

Variograma Empírico

- En el caso de datos Geoestadísticos (dominio fijo y continuo) una de las principales herramientas descriptiva para analizar si existe dependencia espacial es el variograma o semivariograma.

Variograma Empírico

- En el caso de datos Geoestadísticos (dominio fijo y continuo) una de las principales herramientas descriptiva para analizar si existe dependencia espacial es el variograma o semivariograma.
- El variograma consiste en analizar el comportamiento

$$\text{var}[Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})], \quad \mathbf{s}, \mathbf{s} + \mathbf{h} \in D$$

- El estimador de Matheron (1965) del semivariograma es :

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} [Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2$$

donde $N(\mathbf{h})$ es el conjunto de todos los pares de localizaciones que tienen diferencia \mathbf{h} y $|N(\mathbf{h})|$ es el número de pares distintos en este conjunto.

Obs: Para construir el estimador considere una partición

$$t_0, t_1, \dots, t_K \quad \text{del intervalo} \quad (0, t_{max}),$$

donde $t_{\max} = 0,5 \cdot \max\{\|\mathbf{s}_i - \mathbf{s}_j\|\}^1$. Además, considere los intervalos

$$I_k = (t_{k-1}, t_k)$$

En este caso, se tiene que

$$N(t_k) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in I_k\}$$

¹Recuerde que en una partición $t_0 = 0$ y $t_K = t_{\max}$.

Variograma Empírico

Cresie & Hawking (1980) sugirieron un estimador robusto del variograma.

$$\begin{aligned}Z(\mathbf{s}_i) &\sim \mathcal{N}(0, \sigma^2) \\ Z(\mathbf{s}_i) - Z(\mathbf{s}_j) &\sim \mathcal{N}(0, 2\gamma(\mathbf{s}_i - \mathbf{s}_j))\end{aligned}$$

Luego,

$$\frac{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2}{2\gamma(\mathbf{s}_i - \mathbf{s}_j)} \sim \chi^2_{(1)}; \quad \sqrt[4]{\frac{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^2}{2\gamma(\mathbf{s}_i - \mathbf{s}_j)}} \sim \mathcal{Normal}$$

Variograma Empírico

Cresie & Hawking (1980) sugirieron un estimador robusto del variograma.

$$\begin{aligned}Z(\mathbf{s}_i) &\sim \mathcal{N}(0, \sigma^2) \\ Z(\mathbf{s}_i) - Z(\mathbf{s}_j) &\sim \mathcal{N}(0, 2\gamma(\mathbf{s}_i - \mathbf{s}_j))\end{aligned}$$

Luego,

$$\frac{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2}{2\gamma(\mathbf{s}_i - \mathbf{s}_j)} \sim \chi^2_{(1)}; \quad \sqrt[4]{\frac{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^2}{2\gamma(\mathbf{s}_i - \mathbf{s}_j)}} \sim \mathcal{Normal}$$

Obteniendo

$$\hat{\gamma}(\mathbf{h}) = \frac{\left(\frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} \right)^4}{0,914 + \frac{0,988}{|N(\mathbf{h})|}}$$

Variograma Empírico

Ejemplo: Sea Z una variable atributo que ha sido observada en $\mathbf{s} = (x, y)$. Los datos son:

$$Z(1,1) = 1; \quad Z(2,2) = 2; \quad Z(3,4) = 20; \quad Z(1,4) = 4; \quad Z(3,1) = 3$$

- 1 Grafique adecuadamente los datos.
- 2 Obtenga y grafique el estimador de Momentos de Matheron del variograma.
- 3 Obtenga y grafique el estimador de Robusto de Cresie & Hawking del variograma.

- Banerjee S., Carlin B., and Gelfand A. (2015) *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton: Chapman Hall/CRC.
- Cressie, N. (1993) *Statistics for Spatial Data*. New York: Wiley.
- Matheron, G., (1965) Les variables régionalisés et leur estimation. Masson, Paris.
- Shabenberger, O., Gotway, C. A. (2005) *Statistical Methods for Spatial Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

¿Alguna Consulta?