Tarea 1 - MAT467

Alejandro Villazón G.

Considere el estadístico de Mantel dado por

$$M_Y = \sum_{i=1}^{n} \sum_{j=1}^{n} \|s_i - s_j\| \, |Y(s_i) - Y(s_j)|$$

Asumiendo que $\mathbb{P}(Y(s_i) = 1) = p$ y que $\mathrm{Cov}(Y(s_i), Y(s_j)) = 0$:

1. Calcule $\mathbb{E}[M_Y]$.

   Primero notemos que dada la simetría, el estadístico de Mantel puede ser escrito como:

   $$M_Y = 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|s_i - s_j\| \, |Y(s_i) - Y(s_j)|$$

   Luego, por definición de esperanza de una variable aleatoria discreta, si $i \neq j$

   $$\begin{aligned}
   \mathbb{E}[|Y(s_i) - Y(s_j)|] &= 1 \cdot \mathbb{P}(|Y(s_i) - Y(s_j)| = 1) + 0 \cdot \mathbb{P}(|Y(s_i) - Y(s_j)| = 0) \\
   &= \mathbb{P}(|Y(s_i) - Y(s_j)| = 1) \\
   &= \mathbb{P}([Y(s_i) = 1 \wedge Y(s_j) = 0] \vee [Y(s_i) = 0 \wedge Y(s_j) = 1]) \\
   &= 2p(1-p)
   \end{aligned}$$

   Finalmente, dada la linealidad de la esperanza, se tiene el siguiente resultado,

   $$\begin{aligned}
   \mathbb{E}[M_Y] &= \mathbb{E}\left[2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|s_i - s_j\| \, |Y(s_i) - Y(s_j)| \right] \\
   &= 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|s_i - s_j\| \, \mathbb{E}[|Y(s_i) - Y(s_j)|] \\
   &= 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|s_i - s_j\| \, 2p(1-p) \\
   &= 4p(1-p) \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \|s_i - s_j\| .
   \end{aligned}$$

2. Calcule $\text{Var}\left[M_Y\right]$.

   Consideremos en adelante $w_{ij} = \|s_i - s_j\|$, $Y_i := Y(s_i)$. Luego,

$$\text{Var}\left[M_Y\right] = 4\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{n-1}\sum_{l=k+1}^{n} w_{ij}w_{kl}\,\text{cov}\left(|Y_i - Y_j, Y_k - Y_l|\right)$$

   Si $i \neq j$ y $k \neq l$, tenemos que,

$$\begin{aligned}
\text{cov}\left(|Y_i - Y_j|, |Y_k - Y_l|\right) &= \text{cov}\left(\left(Y_i - Y_j\right)^2, \left(Y_k - Y_l\right)^2\right) \\
&= \text{cov}\left(Y_i^2 - 2Y_iY_j + Y_j^2, Y_k^2 - 2Y_kY_l + Y_l^2\right) \\
&= \left[\text{cov}\left(Y_i^2, Y_k^2\right) + \text{cov}\left(Y_i^2, Y_l^2\right) + \text{cov}\left(Y_j^2, Y_k^2\right) + \text{cov}\left(Y_j^2, Y_l^2\right)\right] \\
&\quad - 2\left[\text{cov}\left(Y_i^2, Y_kY_l\right) + \text{cov}\left(Y_j^2, Y_kY_l\right) + \text{cov}\left(Y_iY_j, Y_k^2\right)\right. \\
&\quad \left. + \text{cov}\left(Y_iY_j, Y_l^2\right)\right] + 4\,\text{cov}\left(Y_iY_j, Y_kY_l\right)
\end{aligned}$$

   Utilizando la definición $\text{cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$, la naturaleza binaria de la variable Y_i y recordando que $i \neq j$ y $k \neq l$, tenemos que,

$$\begin{aligned}
\text{cov}\left(Y_i^2, Y_k^2\right) &= \text{cov}\left(Y_i, Y_k\right) = p(1-p)1_{\{i=k\}} \\
\text{cov}\left(Y_iY_j, Y_k^2\right) &= \text{cov}\left(Y_iY_j, Y_k\right) = p^2(1-p)\left[1_{\{i=k\}} + 1_{\{j=k\}}\right] \\
\text{cov}\left(Y_iY_j, Y_kY_l\right) &= p^2(1-p^2)1_{\{(i=k\ \wedge\ j=l)\ \vee\ (j=k\ \wedge\ i=l)\}} \\
&\quad + p^3(1-p)1_{\{(i=k\ \wedge\ j\neq l)\ \vee\ (j=k\ \wedge\ i\neq l)\ \vee\ (i=l\ \wedge\ j\neq k)\ \vee\ (j=l\ \wedge\ i\neq k)\}}
\end{aligned}$$

   De forma análoga se obtiene el resto de términos cruzados, reemplazando tenemos que,

$$\begin{aligned}
\text{cov}\left(|Y_i - Y_j|, |Y_k - Y_l|\right) &= p(1-p)\left[1_{\{i=k\}} + 1_{\{i=l\}} + 1_{\{j=k\}} + 1_{\{j=l\}}\right] \\
&\quad - 2p^2(1-p)\left[1_{\{i=k\}} + 1_{\{i=l\}} + 1_{\{j=k\}} + 1_{\{j=l\}} + 1_{\{i=k\}} + 1_{\{j=k\}}\right. \\
&\quad \left. + 1_{\{i=l\}} + 1_{\{j=l\}}\right] + 4p^2(1-p^2)1_{\{(i=k\ \wedge\ j=l)\ \vee\ (j=k\ \wedge\ i=l)\}} \\
&\quad + 4p^3(1-p)1_{\{(i=k\ \wedge\ j\neq l)\ \vee\ (j=k\ \wedge\ i\neq l)\ \vee\ (i=l\ \wedge\ j\neq k)\ \vee\ (j=l\ \wedge\ i\neq k)\}} \\
&= \left[p(1-p) - 4p^2(1-p)\right]\left[1_{\{i=k\}} + 1_{\{i=l\}} + 1_{\{j=k\}} + 1_{\{j=l\}}\right] \\
&\quad + 4p^2(1-p^2)1_{\{(i=k\ \wedge\ j=l)\ \vee\ (j=k\ \wedge\ i=l)\}} \\
&\quad + 4p^3(1-p)1_{\{(i=k\ \wedge\ j\neq l)\ \vee\ (j=k\ \wedge\ i\neq l)\ \vee\ (i=l\ \wedge\ j\neq k)\ \vee\ (j=l\ \wedge\ i\neq k)\}}
\end{aligned}$$

   Finalmente, al reemplazar lo anterior en la expresión de la $\text{Var}[M_Y]$ obtenemos lo pedido.

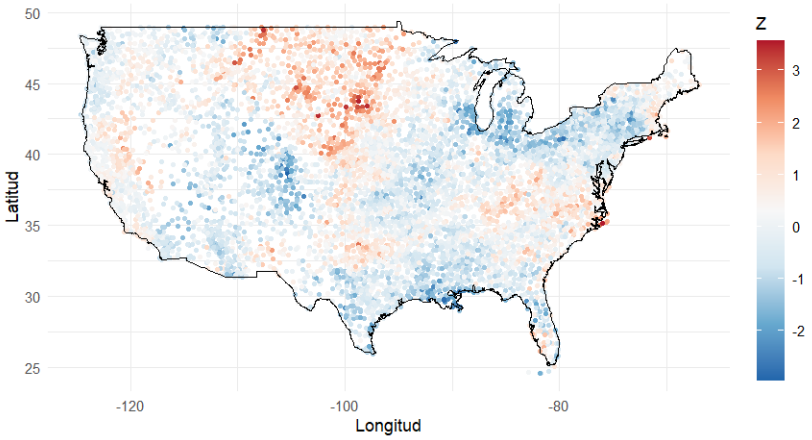


Figure 1: Mapa de calor de *anomalies*.

Considere el set de datos *anomalies* de la librería **GeoModels**. Realice un análisis descriptivo de datos que incluya:

3. Una breve descripción de los datos.

   anomalies es un extracto de la colección de datos espaciales irregularmente espaciados de mediciones de estaciones meteorológicas en Estados Unidos recopiladas por la National Climatic Data Center (NCDC) entre los años 1895 a 1977.

   En específico, consideramos las mediciones del año 1962 de las anomalías de precipitación total anual, es decir, totales anuales estandarizados por la media y la desviación estándar de largo plazo para cada estación meteorológica. El conjunto de datos consiste de 7.352 sitios (longitud y latitud) junto a sus observaciones.

   La Figura 1 muestra los datos representados en un mapa de calor según la variable considerada, observamos una concentración de los mayores valores en la zona centro norte de Estados Unidos.

4. Indicadores de resumen apropiados (promedio, varianza, y otros).

   En la Tabla 1 se presentan algunas estadísticas relevantes. Podemos obtener información de los sitios gracias al máximo y mínimo, las estaciones meteorológicas están en el rectángulo $[-124, -67] \times [24, 49]$ del plano longitud, latitud de la Tierra.

   También podemos observar que la media de nuestra variable observada es 0, lo que tiene sentido con lo descrito en la pregunta 3, ya que la variable observada viene estandarizada. Además, observamos que la varianza es 0.76, esto se puede explicar por el procedimiento de estandarización utilizado, respecto a valores de largo plazo en cada estación meteorológica.

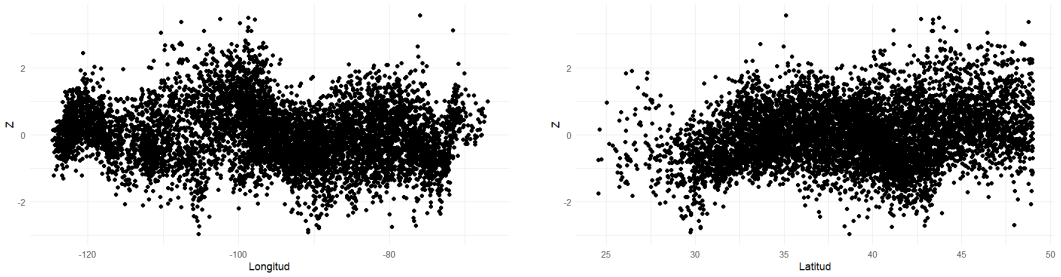
5. Gráficos relevantes.

   En la Figura 2 podemos observar gráficos de dispersión de las coordenadas de los sitios versus nuestra variable observada. No observamos ninguna tendencia clara, mas bien se observa nubes de puntos centradas horizontalmente en $Z = 0$.

   Por otra parte, en la Figura 3 se presenta el histograma de la variable observada, podemos notar que la distribución se asemeja a una normal centrada en 0.

	Longitud	Latitud	Z
Mínimo	-124.73	24.55	-2.96
Máximo	-67	49	3.56
Promedio	-96.25	38.83	0
Varianza	217.80	25.79	0.76

Table 1: Indicadores de resumen.



(a) (b)

Figure 2: Gráficos de dispersión.

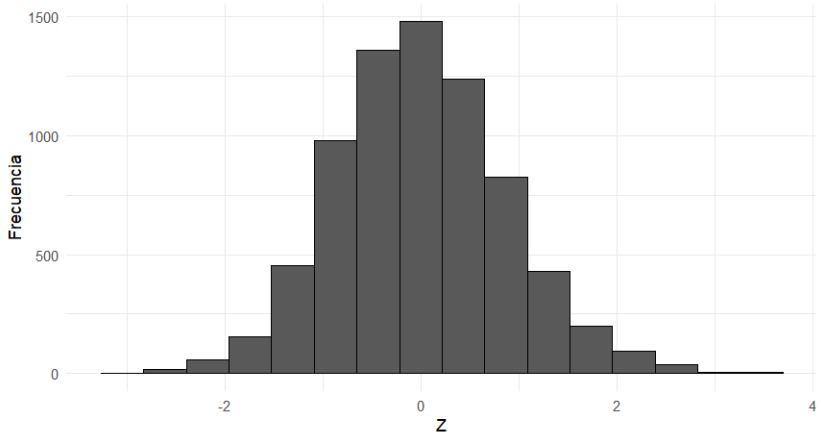


Figure 3: Histograma variable observada.

Sea τ un número real. Considere la siguiente transformación

$$Y_\tau(s_i) = \begin{cases} 1 & \text{si } X(s_i) \geq \tau \\ 0 & \text{si } X(s_i) < \tau \end{cases}$$

6. Calcule el estadístico de Mantel estandarizado para la variable Y_τ, es decir

$$M^s_{Y_\tau} = \frac{M_{Y_\tau} - \mathbb{E}[M_{Y_\tau}]}{\sqrt{\text{Var}[M_{Y_\tau}]}}$$

Reporte sus resultados para diferentes valores de τ, realizando el (o los) test de hipótesis correspondientes.

Dada la gran cantidad de observaciones, no fue posible calcular lo solicitado ni siquiera para un valor de τ, esto debido al intensivo cálculo de la varianza. En consecuencia, proponemos realizar el ejercicio en un subconjunto de los datos, para esto consideramos el rectángulo $[-102, -99] \times [31, 34]$ quedando $N = 82$ observaciones con media 0.012 y desviación estándar 0.924.

Para obtener los diferentes valores de τ consideraremos una grilla regular de 100 valores en el intervalo $[\min(X)+\varepsilon, \max(X)-\varepsilon]$, con $\varepsilon = 0.1$, esto con el objetivo de que el estadístico esté bien definido, i.e. la varianza sea no nula, resultando el intervalo $[-1.62, 1.57]$.

Por otro lado, dado que $p = \mathbb{P}(Y_\tau(s_i) = 1) = \mathbb{P}(X(s_i) \geq \tau)$, consideraremos el siguiente estimador de p,

$$\hat{p}_\tau = \frac{1}{N} \sum_{i=1}^{N} Y_\tau(s_i).$$

En la Figura 4 se presentan los resultados obtenidos, notamos que el valor máximo del estadístico de Mantel estandarizado se encuentra muy cerca de $\tau = \text{mediana}(X)$, lo cual tiene sentido pues para ese valor de τ la proporción de los valores que toma Y_τ es pareja, i.e., $\hat{p}_\tau = 0.5$.

Bajo normalidad es posible probar que $M^s_{Y_\tau} \overset{\text{Approx.}}{\sim} N(0,1)$ obteniendo el test de hipótesis con $H_0 : \text{cov}(Y_\tau(s_i), Y_\tau(s_j)) = 0$ para todo $i, j = 1, \ldots, N$. Rechazando H_0 si $\left|M^s_{Y_\tau}\right| > z_{\alpha/2}$ con α un nivel de significancia.

Considerando un nivel de significancia $\alpha = 0.05$ se rechaza la hipótesis nula para valores de τ en el intervalo $[-0.20, 0.38]$, es decir, para estos valores de τ estamos en presencia de autocorrelación espacial. En la Figura 5 se pueden observar los resultados obtenidos, en color rojo tenemos los valores de τ para los cuales se rechaza H_0 considerando $\alpha = 0.05$.

7. Calcule el estadístico de Geary[1]. Compare sus resultados.

Junto al cálculo del estadístico de Geary realizamos el test de hipótesis considerando la alternativa 'two-sided' para una prueba de dos colas, ya que estamos interesados en determinar si hay autocorrelación espacial significativa, ya sea positiva o negativa. Este test de hipótesis prueba que el estadístico de Geary sea significativamente diferente de 1.

Consideramos los mismos valores de τ que en la pregunta 6 y el mismo nivel de significancia $\alpha = 0.05$, los valores obtenidos para el estadístico de Geary se presentan en la Figura 6.

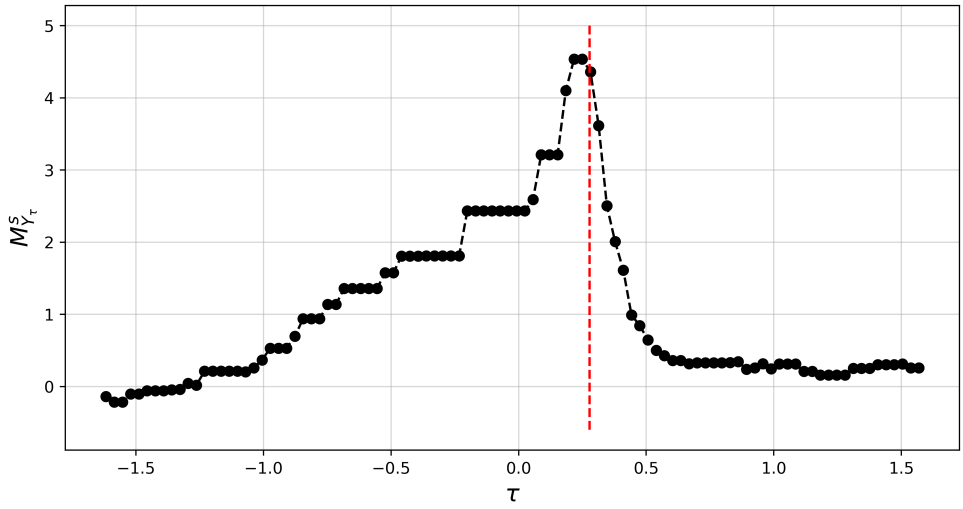


Figure 4: Mantel estandarizado para distintos valores de τ. En color rojo se presenta la mediana de X.

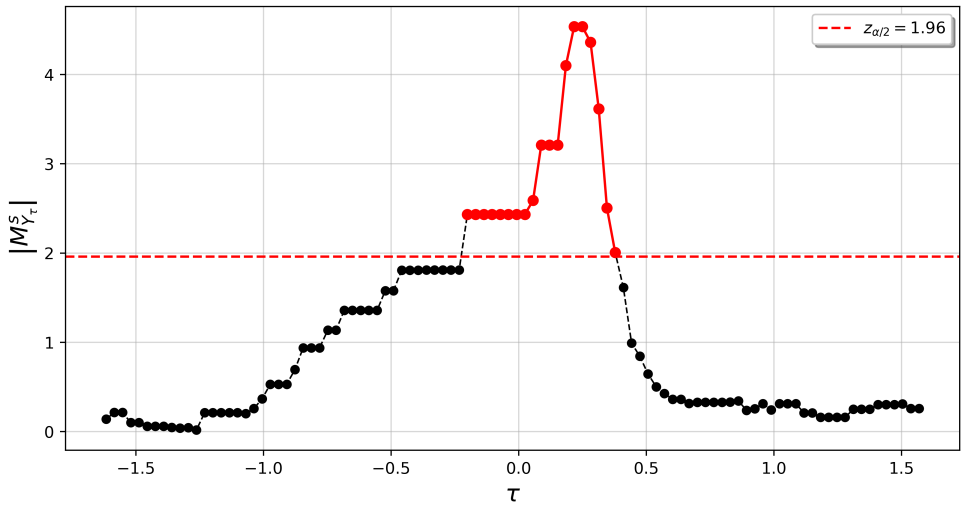


Figure 5: Mantel estandarizado en test de hipótesis de autocorrelación espacial con $\alpha = 0.05$.

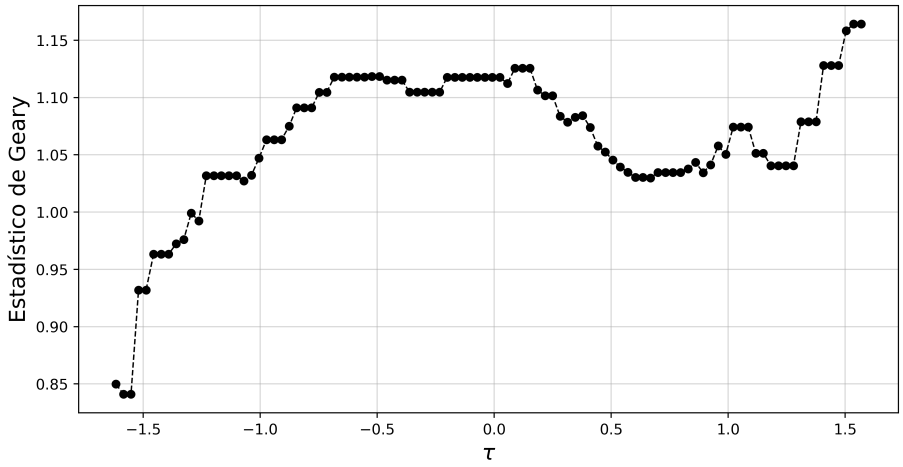


Figure 6: Estadístico de Geary para distintos valores de τ.

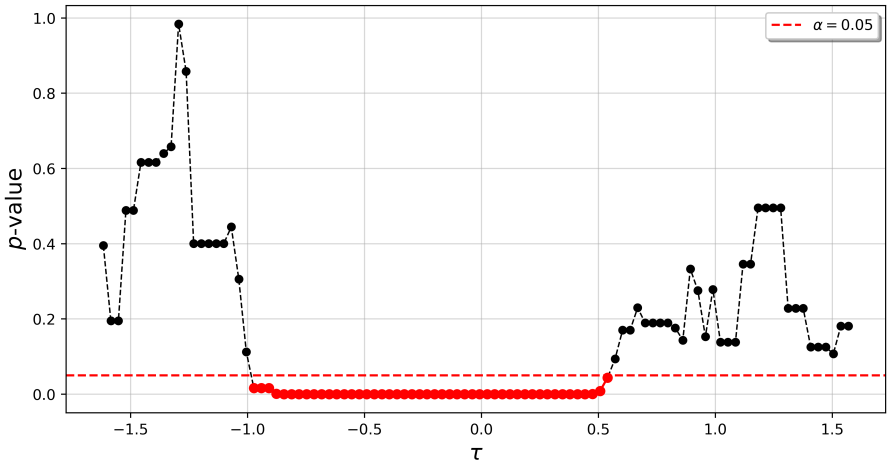


Figure 7: p-valores para los diferentes valores de τ en test de hipótesis de autocorrelación espacial utilizando el estadístico de Geary.

En la Figura 7 se presentan los p-valores obtenidos en el test de hipótesis para los diferentes valores de τ, rechazando la hipótesis nula si $p \leq \alpha$. En la figura, estos valores han sido pintados en rojo. Observamos que se rechaza la hipótesis nula para valores de τ en el intervalo $[-0.98, 0.54]$. Esto significa que para estos valores de τ, la variable Y_τ presenta autocorrelación espacial.

Al comparar los resultados obtenidos utilizando los estadísticos de Geary y Mantel, notamos que la región de rechazo según el estadístico de Geary es más amplia y contiene a la de Mantel. Esto significa que el estadístico de Geary identifica una mayor cantidad de valores de τ donde se rechaza la hipótesis nula. Por lo tanto, dada la contención, podemos decir que los resultados dados por el estadístico de Geary respaldan la presencia de autocorrelación espacial en la variable Y_τ para los valores de τ obtenidos mediante el test de hipótesis utilizando Mantel.

[1]Se le recomienda utilizar la función *geary.test* del paquete **spdep**.