

Análisis de Conglomerados Esférico para Comprobar Autocorrelación Espacial Positiva: Una Aplicación al Índice de Marginación en Méjico

Carlos Espino García

Instituto Tecnológico Autónomo de México

Marzo, 2015

Objetivos

- El objetivo principal es detectar una estructura espacial latente entre los municipios de México a partir de clasificarlos en grupos de acuerdo a sus características de marginación.
- Para la agrupación se utilizará el algoritmo de conglomerados de k -medias esféricas.
- Para corroborar la estructura espacial de los grupos, se utilizan medidas de autocorrelación espacial para variables discretas.
- Aprovechando la introducción a la estadística espacial, se hará una breve análisis del índice de marginación propuesto por CONAPO y se harán pruebas de autocorrelación espacial para variables continuas sobre dicho índice.

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

¿Qué es el análisis de conglomerados?

- El análisis de conglomerados es una técnica de estadística multivariada que consiste en agrupar objetos, tomando como base únicamente la información que encontramos en los datos que describen al objeto y a sus relaciones.
- El objetivo es formar grupos (conglomerados) cuyos elementos tengan características similares entre sí, pero que estén poco relacionados con los objetos de otros grupos.

Notación

- $X \in \mathbb{R}^{n \times p}$ denota el conjunto de observaciones de n individuos y p variables. x_{ij} es la medición del atributo j para el individuo i para $i = 1, 2, \dots, n$ y $j = 1, 2, \dots, p$.
- Denotamos al individuo i como x_i , donde $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T \in \mathbb{R}^p$ para $i = 1, 2, \dots, n$. Así $X = (x_1, x_2, \dots, x_n)^T$.
- C_k denota el k -ésimo grupo.
- K denota el número total de grupos.

Enfoques

- **Particional:** Es simplemente una división del conjunto de datos en subconjuntos mutuamente excluyentes de tal forma que cada objeto esté en sólo un subconjunto.
Se busca hacer una partición de X en K grupos,
 $C = \{C_1, C_2, \dots, C_K\}$ tal que:
 - $C_i \neq \emptyset, i = 1, 2, \dots, K$
 - $\bigcup_{i=1}^K C_i = X$
 - $C_i \cap C_j = \emptyset$ con $, i, j = 1, 2, \dots, K, i \neq j$
- **Jerárquico:** Si los grupos tienen subgrupos, entonces obtenemos un conglomerado jerárquico, que es un conjunto de conglomerados anidados

Matrices de Proximidad

- Muchas veces los datos son representados en términos de la proximidad entre pares de objetos. Esto puede ser ya sea por sus similitudes o disimilitudes.
- Así, los datos pueden ser representados en una matriz D de $n \times n$, donde n es el número de individuos y cada entrada d_{ij} representa la proximidad entre el individuo i y el j .

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

K-medias

- Utiliza como medida de similitud entre dos objetos la distancia Euclídea al cuadrado:

$$d(x_i, x_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2. \quad (1)$$

El objetivo es minimizar:

$$E = \sum_{i=1}^n \|x_i - m_{k(x_i)}\|^2, \quad (2)$$

donde m_i es el centroide que corresponde al grupo i para $i = 1, 2, \dots, k$ y $k(x_i) = \operatorname{argmin}_k \|x_i - m_k\|$ es el índice del centroide más cercano a x_i .

Algoritmo: K-medias

Algoritmo 1: Algoritmo de K-medias

Input: Conjunto de n individuos $X = (x_1, x_2, \dots, x_n)$ en \mathbb{R}^p y el número de grupos K .

Output: Una partición de los datos indexado por $Y = (y_1, y_2, \dots, y_n)$ con $y_i \in \{1, 2, \dots, K\}$ para $i = 1, 2, \dots, n$.

- 1 Inicialización: inicializar los centroides de los grupos $\{m_1, m_2, \dots, m_K\}$
 - 2 Asignación: para cada objeto x_i , se toma
 - 3 $y_i = \operatorname{argmin}_k \|x_i - m_k\|$ con $i = 1, 2, \dots, n$
 - 4 Estimación de centroides: para cada grupo k , sea $C_k = \{x_n | y_n = k\}$, el centroide es estimado como $m_k = \frac{1}{n} \sum_{x \in C_k} x_i$
 - 5 Parar si Y no cambia, en otro caso, regresar a paso 2.
-

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- **K-medias esféricas**
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

K-medias esféricas

- Cuando se cuenta con datos de dimensiones altas se ha mostrado que la similitud de cosenos es una métrica superior a la distancia Euclídea.
- Esta implicación se sigue si asumimos que la dirección del vector de una observación, es más importante que su magnitud.

K-medias esféricas

- La medida de distancia utilizada, que se busca minimizar, es la de disimilitud de cosenos:

$$d(x_i, x_{i'}) = 1 - \cos(x_i, x_{i'}) = 1 - \frac{x_i^T x_{i'}}{\|x_i\| \|x_{i'}\|} \quad (3)$$

- Pero minimizar $1 - \cos(x_i, x_{i'})$ es equivalente a maximizar $\cos(x_i, x_{i'})$.
- Si se normaliza a cada x_i de tal forma que $\|x_i\| = 1$ para $i = 1, 2, \dots, n$ de tal forma que las observaciones pertenezcan a la hiperesfera de dimensión p y radio 1,
 $\mathcal{S}^p = \{x \in \mathbb{R}^p : x^T x = 1\}$ entonces la ecuación 3 se convierte en $d(x_i, x_{i'}) = 1 - x_i^T x_{i'}$.

K-medias esféricas

Sean $\mu_1, \mu_2, \dots, \mu_K$ un conjunto de centroides unitarios, el algoritmo de k-medias esféricas (i.e. k-medias en una hiperesfera unitaria) busca maximizar la similitud de cosenos promedio

$$L = \sum_{i=1}^n x_i^T \mu_{k(x_i)} \quad (4)$$

donde $k(x_i) = \underset{k}{\operatorname{argmax}} x_i^T \mu_k$ es el índice del centroide cuyo ángulo tiene mayor similitud al ángulo de x_i .

Algoritmo: K-medias esféricas

Algoritmo 2: Algoritmo K-medias esféricas

Input: Conjunto de n vectores de individuos unitarios

$X = (x_1, x_2, \dots, x_n)$ en \mathbb{R}^p y el número de grupos K .

Output: Una partición de los datos indexado por

$Y = (y_1, y_2, \dots, y_n)$ con $y_i \in \{1, 2, \dots, K\}$ para
 $i = 1, 2, \dots, n$.

- 1 Inicialización: inicializar los centroides unitarios de los grupos $\{\mu_1, \mu_2, \dots, \mu_K\}$
 - 2 Asignación: para cada objeto x_i , se toma
 - 3 $y_i = \operatorname{argmax}_k x_i^T \mu_k$ con $i = 1, 2, \dots, n$
 - 4 Estimación de centroides: para cada grupo k , sea
 $C_k = \{x_n | y_n = k\}$, el centroide es estimado como
$$\mu_k = \sum_{x \in C_k} \frac{x_i}{\| \sum_{x \in C_k} x_i \|}$$
 - 5 Parar si Y no cambia, en otro caso, regresar a paso 2.
-

- **Observación:** Cuando x y μ son vectores unitarios, es equivalente utilizar similitud de cosenos o norma Euclídea para asignar los datos. Razón:

$$\|x - \mu\|^2 = \|x\|^2 + \|\mu\|^2 - 2x^T \mu = 2 - 2x^T \mu = 2(1 - x^T \mu). \quad (5)$$

- Así pues, en una hiperesfera, maximizar la ecuación 4 es equivalente a minimizar la ecuación 2.
- De esta manera, convertimos el problema de optimización cuadrático en uno lineal.

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

Determinar el número de clusters K

- Debemos seleccionar el número de grupos K^* .
- Supongamos que se han agrupado las observaciones en k grupos. Sea

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \quad (6)$$

la suma de la distancia entre pares para todos los puntos en el grupo r , y sea

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r \quad (7)$$

la medida de compacidad de los clusters.

Determinar el número de clusters K

- Idea: estandarizar la gráfica de $\log W_k$ comparándola con su valor esperado bajo una distribución de referencia nula.
- La estimación del valor óptimo de grupos es el valor K para el cual $\log W_k$ cae lo más lejano por debajo de su curva de referencia.
- Así, se define, $\text{Gap}_n(k) = E_n^* [\log(W_k)] - \log(W_k)$ donde E_n^* denota el valor esperado bajo una muestra de tamaño n de la distribución de referencia generada por Monte Carlo.
- El estimador \hat{K} es aquel valor que maximiza $\text{Gap}_n(k)$.
- Se elige la distribución uniforme como distribución de referencia, distribuyendo los datos sobre un rectángulo que contiene los datos.

Implementación Computacional del Estadístico Gap

- 1 Agrupar las observaciones utilizando diferentes números de conglomerados $K = 1, 2, \dots, M$, donde M es un número fijo. Se calcula W_k .
- 2 Generar B conjuntos de datos. Agrupar cada conjunto y calcular W_{Kb}^* , $b = 1, 2, \dots, B$, $K = 1, 2, \dots, M$ y estimar

$$\text{Gap}(K) = \frac{1}{B} \sum_{b=1}^B \log W_{Kb}^* - \log W_k \quad (8)$$

- 3 Calcular la desviación estándar de W_{Kb}^* y definir

$$s_K = sd_K \sqrt{1 + \frac{1}{B}}.$$

Finalmente, escoger la menor K tal que

$$\text{Gap}(K) \geq \text{Gap}(K+1) - s_{K+1}.$$

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

¿Qué es?

- El análisis estadístico espacial es un conjunto de técnicas y modelos que usan la referencia espacial asociada a cada observación.
- Los datos espaciales se distinguen por ser observaciones obtenidas en ubicaciones espaciales que pueden ser coordenadas en el plano \mathbb{R}^2 ó \mathbb{R}^3 , líneas que unen un punto con otro o polígonos que cubren un área determinada.
- Los modelos espaciales intentan modelar la correlación entre observaciones en diferentes posiciones en el espacio.
- Existen tres tipos de análisis de datos:
 - **Geoestadística:** Se mide la variable de interés sobre el espacio de manera continua.
 - **Patrón de Puntos Espaciales:** Se interesa la ubicación de ciertos “eventos” que ocurren en el espacio.
 - **Datos en retícula:** Contamos con una partición de la región de estudio donde se llevan a cabo las observaciones.

Ejemplos

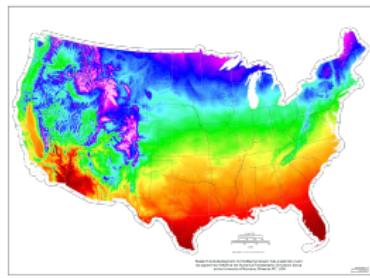


Figure: Geoestadística



Figure: Point pattern analysis

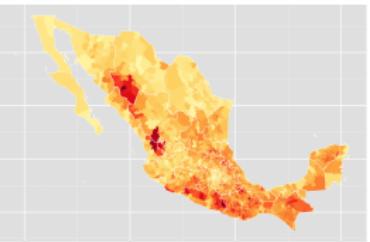


Figure: Datos en retícula

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

Autocorrelación Espacial

- Las observaciones realizadas en diferentes puntos espaciales pueden no ser independientes.
- A esto se le llama autocorrelación espacial, la cual mide el grado en el que un fenómeno de interés se relaciona consigo mismo en el espacio.
- Las pruebas de autocorrelación espacial examinan si el valor observado de una variable en algún lugar es independiente de los valores de la misma variable en lugares cercanos o contiguos.

- **Autocorrelación espacial positiva** indica que los valores similares están cercanos entre sí, o aglomerados, en el espacio.
- **Autocorrelación espacial negativa** indica que valores vecinos no son similares, o equivalentemente, que valores similares están dispersos en el espacio.
- **Autocorrelación espacial nula** indica que el patrón espacial es aleatorio.

Matriz de pesos

Debemos definir primero qué significa que dos observaciones sean cercanas a través de una matriz de pesos W .

- **Contigüidad binaria:**

$$w_{ij} = \begin{cases} 1 & \text{si la región } i \text{ comparte frontera con } j \\ 0 & \text{e.o.c.} \end{cases} \quad (10)$$

- **Distancia:** Distancia entre dos puntos o regiones.
- **Frontera en común:** Longitud de la frontera entre dos regiones.
- **Combinación de frontera y distancia**

Podemos ajustar W de tal forma que la suma de los pesos por renglón sea igual a 1 utilizando una matriz estandarizada por filas.

Pruebas de Autocorrelación Espacial

- Consideraremos un área de estudio particionada en n regiones.
- Sea Y la variable de estudio, y_i es la observación de la variable Y en la región i .
- Si Y es de escala continua u ordinal, utilizamos los coeficientes \mathcal{I} de Moran y \mathcal{C} de Geary.
- Si Y es nominal, utilizamos el estadístico de conteo de fronteras (Join count).

Índice \mathcal{I} de Moran

Se basa en los productos cruzados de las desviaciones de la media y se calcula:

$$\mathcal{I} = \left(\frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \right). \quad (11)$$

\mathcal{I} no es como un coeficiente de correlación común, pues no pertenece necesariamente al intervalo $(-1, 1)$. Usualmente $\mathcal{I} \in (-1, 1)$

Índice \mathcal{C} de Geary

Utiliza la suma de diferencias al cuadrado entre pares de observaciones como medida de variación. Está dado por

$$\mathcal{C} = \left(\frac{n-1}{2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - y_j)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \right). \quad (12)$$

Toma valores en el intervalo $[0, 2]$ donde 0 indica correlación espacial positiva perfecta y 2 indica correlación espacial negativa perfecta.

Distribución de \mathcal{I} y \mathcal{C}

Bajo ciertas condiciones de regularidad, \mathcal{I} y \mathcal{C} se distribuyen normal a medida que n aumenta. Los momentos de \mathcal{I} y \mathcal{C} pueden ser evaluados bajo alguna de los siguientes dos supuestos:

- 1 Normalidad:** Asumimos que las observaciones y_i son resultado de n realizaciones de una población normal.
- 2 Aleatorización:** Consideramos el valor observado de \mathcal{I} y \mathcal{C} relativo al conjunto de todos los valores posibles que pueden tomar \mathcal{I} y \mathcal{C} si y_1, y_2, \dots, y_n fueran permutadas de manera aleatoria repetidamente alrededor de las regiones dentro del área de estudio.

Momentos de \mathcal{I} y \mathcal{C}

Coeficiente \mathcal{I}

$$E[\mathcal{I}] = -\frac{1}{n-1}, \quad E[\mathcal{I}^2] = \frac{n^2 S_1 - n S_2 + 3S_0^2}{S_0^2(n^2 - 1)} \quad (13)$$

Coeficiente \mathcal{C}

$$E[\mathcal{C}] = 1, \quad \text{Var}(\mathcal{C}) = \frac{(2S_1 + S_2)(n-1) - 4S_0^2}{2(n+1)S_0^2} \quad (14)$$

Donde

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}, \quad S_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (w_{ij} + w_{ji})^2, \quad S_2 = \sum_{i=1}^n (w_{i.} + w_{.i})^2.$$

$$w_{i.} = \sum_{j=1}^n w_{ij} \quad \text{y} \quad w_{.j} = \sum_{i=1}^n w_{ij}.$$

Variables Discretas: Estadístico join-count

- Se utiliza si Y es categórica.
- Supongamos que Y cuenta con 2 clases $y_i \in \{0, 1\}$.
Mapeando y en 2 colores , W si $y_i = 0$ y B si $y_i = 1$ cada frontera o unión entre dos regiones es clasificada como WW ($0-0$), BB ($1-1$) ó BW ($1-0$).
- Si el número de uniones BB es significativamente mayor del esperado por sorteo, habrá autocorrelación espacial positiva; si es significativamente menor, autocorrelación espacial negativa; y si es aproximadamente el mismo, autocorrelación espacial nula.

Distribución de los conteos

- BB , BW y WW asintóticamente se distribuyen normal.
- Los parámetros μ y σ^2 de los coeficientes pueden ser evaluados bajo uno de los dos supuestos:
 - 1 **Muestreo con reemplazo:** cada una de las regiones es etiquetada como B o W independientemente con probabilidad p_B y $p_W = 1 - p_B$ respectivamente.
 - 2 **Muestreo sin reemplazo:** cada región tiene la misma probabilidad, a priori, de ser B o W , pero la codificación está sujeta a la restricción de que hay n_B regiones con color B y n_W regiones con color W , y $n_a + n_b = n$.

Generalización a $k > 2$

Generalmente contamos con más de dos clases ($k > 2$), tenemos que cada una de las n regiones pertenece a alguna de las k categorías. Así, n_1 regiones son de tipo 1, n_2 regiones son de tipo 2 y así sucesivamente, y n_k regiones son de tipo k . De tal manera:

$$n_1 + n_2 + \dots + n_k = n.$$

Definimos N_{rr} como el número de uniones entre regiones del tipo rr , N_{rs} el número de uniones del tipo rs , con $r, s \in \{1, 2, \dots, k\}$.

Momentos de los estadísticos N_{rr} bajo muestreo sin reemplazo

$$\mu = \frac{S_0 n_r (n_r - 1)}{2n(n-1)}, \quad (15)$$

$$\sigma^2 = \frac{S_1 n_r^{(2)}}{4n^{(2)}} + \frac{(S_2 - 2S_1)n_r^{(3)}}{4n^{(3)}} + \frac{(S_0^2 + S_1 - S_2)n_r^{(4)}}{4n^{(4)}} - \mu^2. \quad (16)$$

donde $n^{(k)} = n(n-1)(n-2)\dots(n-k+1)$.

Pruebas de hipótesis

- La hipótesis nula H_0 es de no autocorrelación espacial,
- Hay dos procedimientos a seguir:
 - 1 Utilizando el supuesto de normalidad.
 - 2 Si dudamos del supuesto de normalidad, podemos utilizar simulaciones de Monte Carlo. Es recomendable hacer esta prueba ya que la función de densidad del estadístico es sensible a los siguientes factores:
 - La forma de las regiones en el área de estudio.
 - Los pesos w_{ij} utilizados.
 - La distribución de la variable Y .
 - El tamaño de la muestra n .

Simulaciones de Monte Carlo

- 1 Permutamos aleatoriamente las etiquetas y_1, y_2, \dots, y_n a través de las regiones.
- 2 Calculamos el estadístico de interés, digamos \mathcal{I} , con las etiquetas permutadas.
- 3 Repetimos 1. y 2. n_{sim} veces, obteniendo una muestra de tamaño n_{sim} de \mathcal{I} .
- 4 Comparamos el valor observado del estadístico con la muestra obtenida.

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

■ Análisis Exploratorio

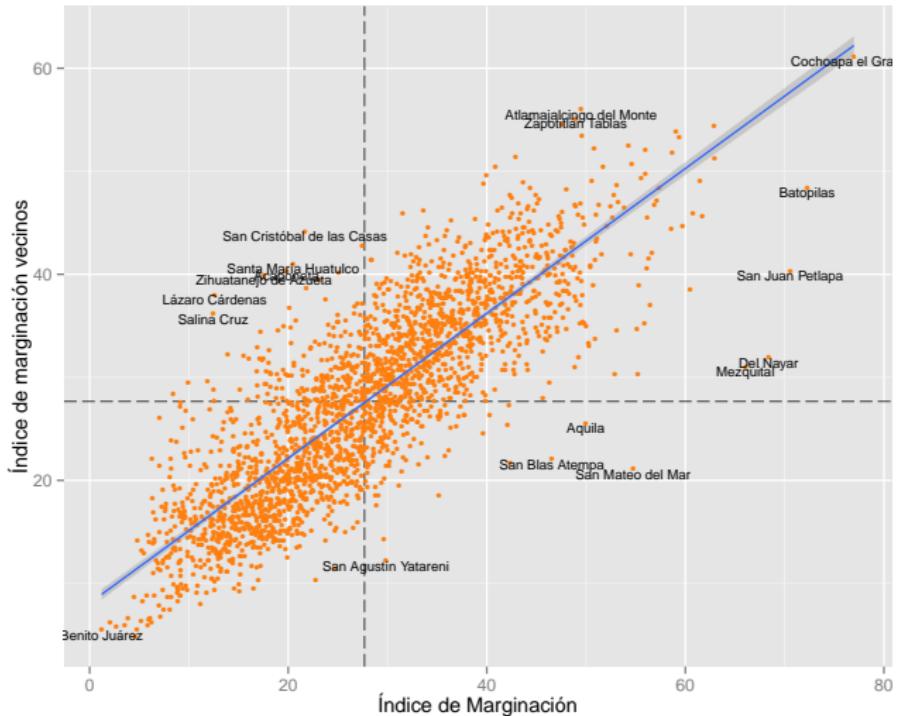
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

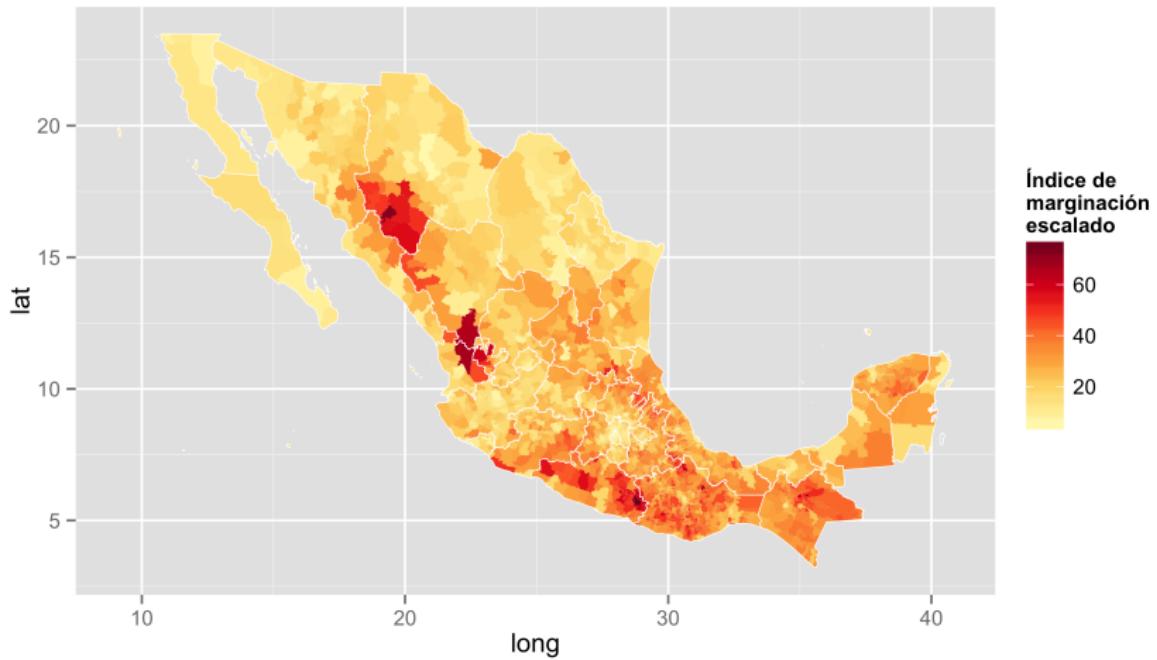
4 Conclusiones

Base de Datos

- Base de datos: “Índice de Marginación por Entidad Federativa y Municipio 2010”.
- 2,456 municipios y cada una tiene 12 atributos.
- 9 de estos atributos son indicadores de marginación.
- Los atributos restantes son el índice de marginación, índice de marginación escalado de 0 a 100 y grado de marginación.
- El índice de marginación es un resumen de estos atributos y el grado de marginación es una categorización de éste.

Análisis Exploratorio: Índice de Marginación





Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

Pruebas de Autocorrelación Espacial para el índice de marginación

- Se escogen pesos binarios estandarizados.
- Hipótesis alternativa, autocorrelación espacial positiva
- Obtenemos: $\hat{\mathcal{I}} = 0.703$ y $\hat{\mathcal{C}} = 0.2943260693$
- Bajo el supuesto de normalidad: $z(\mathcal{I}) = 57.6933$ y $z(\mathcal{C}) = 51.0396$
- Utilizando simulaciones de Monte Carlo con $n_{sim} = 9999$:

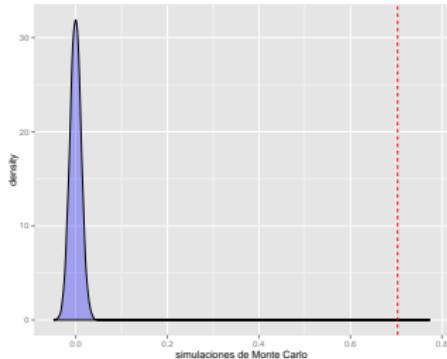


Figure: Simulaciones de \mathcal{I}

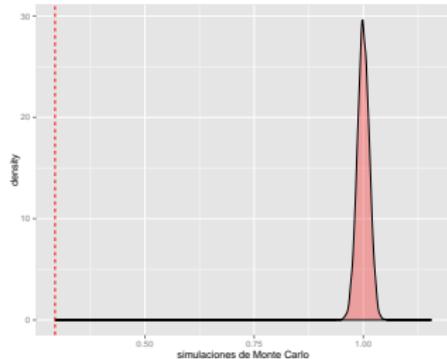


Figure: Simulaciones de \mathcal{C}

Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

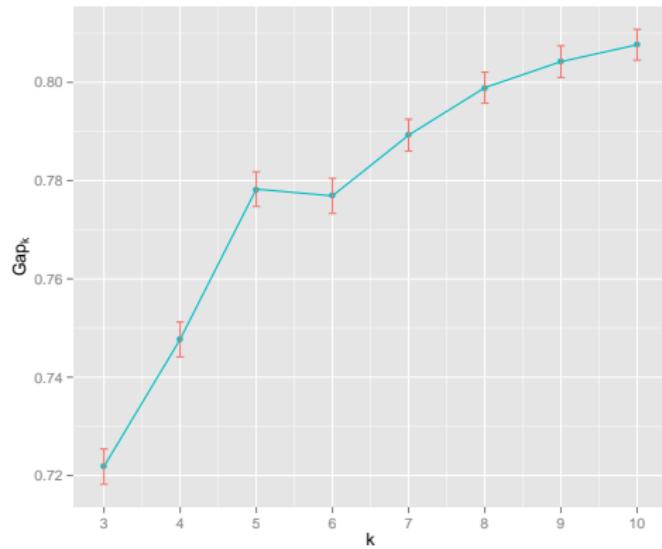
3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- **Análisis de Conglomerados esférico**
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

K-medias esféricas: escoger K^*

- Escogemos K^* utilizando el estadístico Gap.
- Utilizando $M = 10$ y $B = 100$:

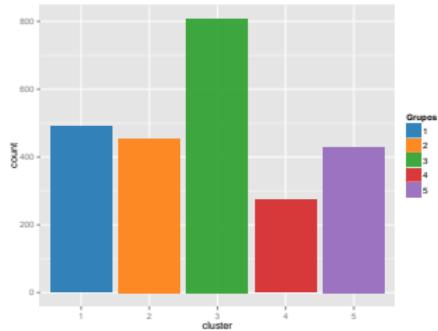


- Obtenemos $K^* = 5$

Reultados K-medias esféricas

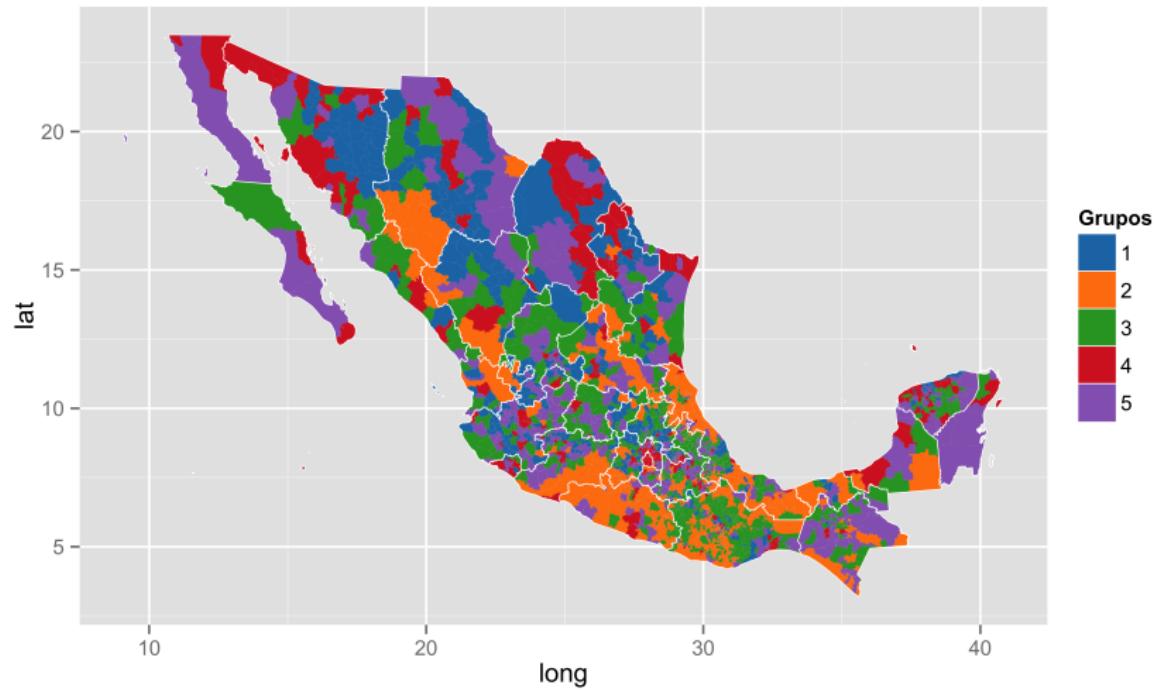
- 1 En el primer grupo cayeron municipios con localidades de pocos habitantes y con grado marginación de media a bajo.
- 2 En el segundo, cayeron los municipios más marginados cuyo principal rasgo es la carencia de agua entubada.
- 3 El tercer grupo tiene municipios con grado de marginación de medio a alto, lo que lo separa del grupo dos es que tiene mayor porcentaje de viviendas con agua entubada.
- 4 En el cuarto, se agruparon los municipios con menor grado de marginación.
- 5 En el quinto, cayeron municipios con características similares al del primer grupo pero se diferencia en que cuenta con localidades más grandes.

Distribución de los grupos



	conteo	%
1	490	20%
2	455	19%
3	808	33%
4	274	11%
5	429	17%

Municipios coloreados por grupo



Outline

1 Análisis de Conglomerados

- ¿Qué es?
- K-medias
- K-medias esféricas
- Determinar K

2 Análisis Estadístico Espacial

- ¿Qué es?
- Autocorrelación Espacial

3 Resultados

- Análisis Exploratorio
- Pruebas de Autocorrelación Espacial para el índice de marginación
- Análisis de Conglomerados esférico
- Pruebas de Autocorrelación Espacial para los clusters

4 Conclusiones

¿Existe autocorrelación espacial positiva?

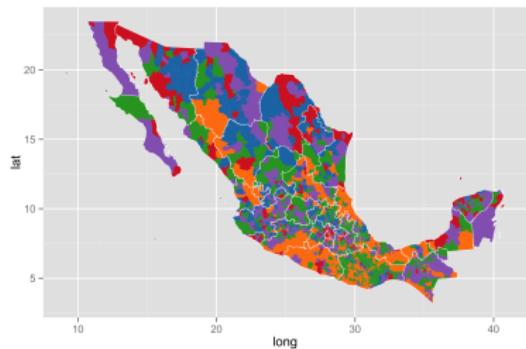


Figure: Obtenido

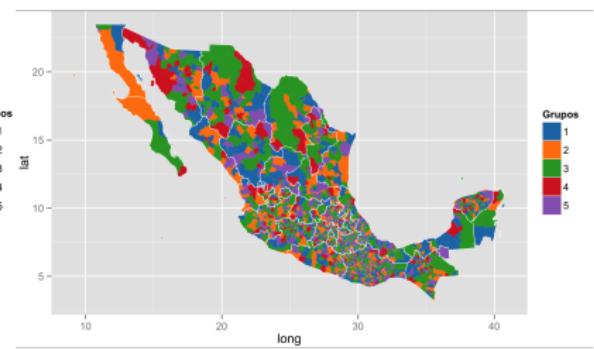


Figure: Aleatorizado

A ojo parece que sí...

Pruebas de Autocorrelación Espacial para los clusters

- Utilizando el estadístico de conteo de fronteras del mismo color N_{rr} , con $r = 1, 2, \dots, 5$, bajo muestreo sin reemplazo se obtiene:
- Bajo supuesto de normalidad:

grupo	\hat{N}_{rs}	$N_{rs,0}$	Var	z	valor-p
1	98.61	48.80	6.53	19.49	$< 2.2 \times 10^{-16}$
2	110.35	42.07	5.77	28.43	$< 2.2 \times 10^{-16}$
3	194.99	132.80	14.00	16.62	$< 2.2 \times 10^{-16}$
4	50.36	15.23	2.36	22.88	$< 2.2 \times 10^{-16}$
5	63.59	37.40	5.22	11.46	$< 2.2 \times 10^{-16}$

- En todos los casos rechazamos la hipótesis nula de no autocorrelación espacial.

Utilizando simulaciones de Monte Carlo con $n_{sim} = 9999$:

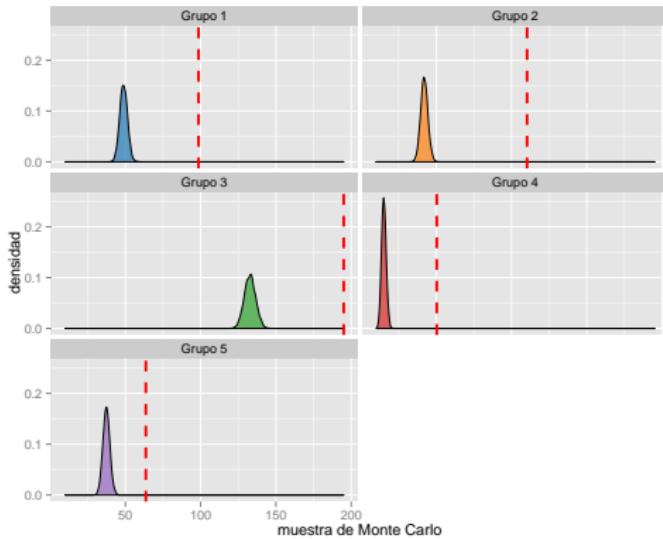


Figure: Densidad de la muestra de Monte Carlo de N_{rr} . La línea punteada indica donde se encuentra \hat{N}_{rr} .

Conclusiones

- Las pruebas sobre los índices \mathcal{I} y \mathcal{C} mostraron alta autocorrelación espacial positiva para el índice de marginación.
- Dentro del análisis de conglomerados, el estadístico Gap mostró que 5 es un número óptimo de grupos y se utilizó el algoritmo de k -medias esféricas para hacer los conglomerados.
- Para comprobar la autocorrelación espacial positiva de los grupos obtenidos, se utilizaron los estadísticos N_{ss} de conteo de fronteras. Los conteos entre fronteras del mismo grupo son significativamente mayores a los conteos esperados, indicando un grado de asociación espacial alto.
- A partir del análisis de conglomerados esférico, se encontró una estructura espacial latente para los datos de marginación.

Importancia del estudio

- Este estudio nos permite identificar aglomeraciones en el mapa y conocer las necesidades de éstas.
- Esto permite definir estrategias en materia de infraestructura para poder atender las carencias o necesidades de cada uno de los municipios.
- Por ejemplo, la instalación de centros de salud o de atención en una zona céntrica en la Sierra Tarahumara.
- Es importante señalar que la marginación de un municipio podría estar correlacionada con otras variables, como la dificultad de acceso o las condiciones geográficas del municipio.

Otros enfoques posibles

- Podría realizarse un análisis similar para identificar focos rojos de violencia, necesidades en cuestión de salud e incluso para identificar segmentos de mercado por región.
- Si quisiéramos hacer un estudio más puntual, podríamos realizar el mismo estudio a nivel AGEB (Área Geoestadística Básica) o por manzana, enfocándonos en una región específica.
- También es posible realizar estudios espacio-temporales para ver la evolución de la marginación de los municipios a través del tiempo.

¡Muchas gracias!