

INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO



ANÁLISIS DE CONGLOMERADOS ESFÉRICO PARA
COMPROBAR AUTOCORRELACIÓN ESPACIAL POSITIVA:
UNA APLICACIÓN AL ÍNDICE DE MARGINACIÓN EN
MÉXICO

TESIS
QUE PARA OBTENER EL TÍTULO DE
LICENCIADO EN MATEMÁTICAS APLICADAS
PRESENTA
CARLOS ESPINO GARCÍA

ASESOR: DR. JUAN JOSÉ FERNÁNDEZ DURÁN

MÉXICO, D.F.

2015

Autorización

Con fundamento en el artículo 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada “Análisis de Conglomerados Esférico para Comprobar Autocorrelación Espacial Positiva: Una Aplicación al Índice de Marginación en México”, otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Baillères Jr. autorización para que fijen la obra en cualquier medio, incluido el electrónico y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por la divulgación una contraprestación.

Carlos Espino García

Fecha

Firma

*Para mis papás, Ida y Carlos
y mi hermana, Ida.*

Agradecimientos

Esta tesis representa la conclusión de una de las mejores etapas de mi vida. Quiero expresar mi agradecimiento a todas las personas que han formado parte de esta gran experiencia como estudiante en el ITAM.

Antes que nada, gracias a Dios por todas las bendiciones que he recibido en la vida, por ser mi fortaleza en momentos de debilidad y por brindarme una vida llena de aprendizajes, experiencias y felicidad.

Gracias a mi familia por estar siempre a mi lado. A mis papás, Ida y Carlos, por todo el apoyo, el cariño, los valores que me han inculcado y por haberme dado la oportunidad de tener una excelente educación. A mi hermana, Ida, por ser siempre un gran ejemplo y por inspirarme a estudiar tan increíble carrera. A mi cuñado Mau, por ser como un hermano para mí.

Gracias a mi asesor, Juan José Fernández Durán, por todas las horas dedicadas asesorando esta tesis, por sus enseñanzas, su paciencia y sus consejos.

Gracias a mis sinodales Rubén Hernández, Alberto Tubilla y Fernando Espón por sus comentarios y revisiones que ayudaron a enriquecer este trabajo.

Muchas gracias a todos mis profesores del ITAM por todas sus enseñanzas que me trajeron hasta aquí. Especialmente quiero agradecer a mis maestros de matemáticas y estadística Guillermo Grabinsky, Ramón Espinosa, Gustavo Preciado, Juan Carlos Aguilar, Victor Guerrero, Manuel Mendoza, Juan Jose Fernández Durán, Rubén Hernández, Luis Felipe González y Luis García Naranjo por reforzar mi gusto por las matemáticas.

A mis profesores de computación Fernando Esponda y Silvia Guardati por hacerme disfrutar la programación. A mis profesores de estudios generales Alfredo Villafranca y Margarita Aguilera por enseñarme que no sirve de nada lo que haga en la vida si no está basado en mejorar nuestro entorno y en ayudar a resolver los problemas de nuestra sociedad.

Gracias a mis amigos de toda la vida, Oso, Rodrigo, Gordo, Jeringa, Mañon, Gorgi y Andrés por todos los momentos que hemos vividos y porque sé que pase lo que pase siempre puedo contar con ustedes.

Gracias a mis amigos del ITAM por confiar y creer en mí y haber hecho de mi etapa universitaria una experiencia que jamás olvidaré.

Gracias a Sofía, Julián, Celina, Pau y Chonki por ser parte significativa en mi vida, sé que estos años apenas son el principio de una larga amistad.

Gracias Ame, Raúl, Nico, Hans, Oscarín, Jimmy, Juanpi, Linda y Andrea, sin ustedes tantas clases y tantas horas de estudio no hubieran sido tan divertidas. Sé que su amistad me la llevo para toda la vida.

Gracias a Guillermo Garduño por ofrecerme mi primer trabajo en Sinnia. A Elmer Garduño y Rodrigo Fortes por ser mis maestros fuera del aula de clases, y a mis amigos Tania, Sergio, Andrea, Maru, Yuriko, Areli y Sonia.

Índice general

Página de Título	I
Autorización	I
Dedicatoria	II
Agradecimientos	III
Índice	V
I Introducción y Marco Teórico	1
1. Introducción	2
2. Análisis de Conglomerados	5
2.1. ¿Qué es el análisis de conglomerados?	5
2.2. Enfoques	7
2.2.1. Particional	7
2.2.2. Jerárquico	7
2.3. Matrices de proximidad	8
2.4. Medidas de disimilitud basada en atributos	9
2.5. K-medias	9
2.5.1. K-medias esféricas	11
K-medias esféricas como densidad Langevin	12
2.5.2. Estadístico Gap para determinar K	14
3. Análisis Estadístico Espacial	17
3.1. ¿Qué es el análisis espacial?	17
3.2. Estadística Espacial	19

4. Autocorrelación Espacial	22
4.1. Matriz de Pesos	23
4.2. Pruebas de Autocorrelación Espacial Global	25
4.2.1. Variables continuas u ordinales	26
Índice \mathcal{I} de Moran	26
Índice \mathcal{C} de Geary	26
4.2.2. Variables Discretas: Estadístico join-count (Conteo de fronteras)	30
4.2.3. Pruebas de hipótesis	34
Bajo el supuesto de normalidad	35
Simulaciones de Monte Carlo	36
4.2.4. Diagrama de dispersión de Moran	38
II Resultados y Conclusión	39
5. Resultados	40
5.1. Descripción de la base	40
5.2. Análisis Exploratorio	42
5.3. Pruebas de autocorrelación espacial para índice de marginación	48
5.3.1. Índice \mathcal{I}	48
Bajo el supuesto de normalidad	49
Bajo el supuesto de aleatorización	49
Simulaciones de Monte Carlo	50
5.3.2. Índice \mathcal{C}	50
Bajo supuesto de normalidad	51
Bajo supuesto de aleatorización	52
Simulaciones de Monte Carlo	52
5.4. Análisis de conglomerados	54
5.4.1. Determinación de K^* utilizando el estadístico Gap .	54
5.4.2. Resultado de K -medias esféricas	55
5.5. Prueba de conteo de fronteras para conglomerados	64
Bajo supuesto de normalidad	65
Simulaciones de Monte Carlo	66
6. Conclusiones	69
A. Software y Reproducibilidad	72

Parte I

Introducción y Marco Teórico

Capítulo 1

Introducción

Este trabajo tiene cuatro objetivos: presentar una introducción al análisis de conglomerados y comparar dos algoritmos, dando enfoque al algoritmo de k -medias esférico; dar una introducción del análisis estadístico espacial y presentar algunas pruebas de autocorrelación espacial; hacer una análisis exploratorio espacial del índice de marginación en México; y por último, presentar una aplicación, agrupando municipios de México utilizando variables de marginación y probando autocorrelación espacial positiva entre los grupos formados. Existen algunos algoritmos de conglomerados que utilizan la estructura espacial de los datos para formar los grupos. Sin embargo, este trabajando pretende detectar una estructura espacial latente entre los municipios de México utilizando solamente sus medidas de marginación, sin información espacial, para generar los grupos.

El análisis de conglomerados tiene como objetivo agrupar objetos, tomando como base solamente la información que encontramos en los datos

que describen al objeto y a sus relaciones. Para dicho análisis se presentan los algoritmos de k -medias tradicional y k -medias esférico; este último utiliza la distancia de cosenos entre observaciones para medir la similitud entre observaciones. Así, si normalizamos las observaciones, éstas quedan sobre una esfera de radio 1 y basta con calcular el producto punto para medir el coseno del ángulo formado entre 2 observaciones. Para estimar el número óptimo de grupos, se presenta el estadístico Gap, propuesto por Tibshirani et al. (2001).

Adicionalmente, se da una introducción al análisis estadístico espacial y al problema de la autocorrelación espacial. El término “autocorrelación” se refiere a la correlación de una variable consigo misma, en este caso, sobre el espacio. El estudio de la estadística espacial toma diferentes formas de acuerdo al tipo de datos utilizados. En consecuencia, se presentan los estadísticos I de Moran y C de Geary para datos numéricos, y los estadísticos de conteo de fronteras para datos nominales. También se muestran algunas formas para hacer pruebas de significancia estadística sobre dichos estadísticos, ya sea utilizando el supuesto de normalidad o utilizando simulaciones de Monte Carlo.

Por último, se presenta una aplicación sobre la base de datos de CONAPO “Índice de Marginación por Entidad Federativa y Municipio 2010”. Primero, se utilizan los índices I y C para comprobar autocorrelación espacial positiva sobre el índice de marginación. En segundo lugar, se utiliza el algoritmo de k -medias esféricas sobre las variables indicadoras de marginación y escogiendo el número de grupos a través del estadístico Gap. Por último, se utilizan los estadísticos de conteo de fronteras para corroborar

la autocorrelación espacial positiva de los grupos formados, comprobando así, la homogeneidad de éstos en el espacio.

El presente trabajo se divide en 5 capítulos además de esta introducción. El segundo capítulo introduce los fundamentos del análisis de conglomerados y presenta los algoritmos de k -medias esféricas. El tercer capítulo da una introducción del análisis estadístico espacial. El cuarto capítulo define el término de autocorrelación espacial presentando los índices y pruebas utilizados. En quinto capítulo se exponen los resultados obtenidos al hacer el análisis de conglomerados y las pruebas de autocorrelación espacial. Por último, el sexto capítulo contiene las conclusiones y posibles aplicaciones.

Capítulo 2

Análisis de Conglomerados

Este análisis divide los objetos (individuos) de un conjunto de datos, en grupos (conglomerados o *clusters*) que sean significativos y/o útiles. Si el objetivo es obtener grupos significativos, los conglomerados deben capturar la estructura natural de los datos. Sin embargo, en algunos casos el análisis de conglomerados es un punto de partida útil para otros propósitos, como hacer un resumen de los datos. Ya sea por comprensión o por utilidad, este análisis ha jugado un papel importante en una amplia variedad de campos: aprendizaje de máquina, minería de datos, estadística, ciencias sociales y naturales, y reconocimiento de patrones.

2.1. ¿Qué es el análisis de conglomerados?

El análisis de conglomerados es una técnica de estadística multivariada que consiste en agrupar objetos, tomando como base solamente 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.

Un objeto puede ser descrito por un conjunto de mediciones, o por su relación con otros objetos. La meta también puede ser organizar los grupos en una jerarquía natural. Esto involucra agrupar sucesivamente de tal manera que a cierto nivel de jerarquía, los conglomerados que estén en el mismo grupo sean más similares entre sí que entre conglomerados de otros grupos.

El análisis de *clusters* también es utilizado para formar estadísticas descriptivas para verificar si los datos consisten o no en un conjunto de grupos, donde cada grupo representa objetos con características diferentes a los objetos en otros grupos.

Un factor común de los objetivos del análisis de conglomerados es la noción de grado de similitud entre dos objetos agrupados. Cualquier algoritmo utilizado para hacer conglomerados busca agrupar objetos basándose en su grado de similaridad.

Notación:

- $X \in \mathbb{R}^{n \times p}$ denota el conjunto de observaciones con 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 j como x_j , donde $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})^T \in \mathbb{R}^p$ para $j = 1, 2, \dots, n$. Así $X = (x_1, x_2, \dots, x_n)$.
- C_k denota el k -ésimo grupo.

- K denota el número total grupos.

2.2. Enfoques

El modo más común para distinguir entre diferentes tipos de conglomerados es si el conjunto de grupos está anidado o no anidado, en términos tradicionales, si es jerárquico o particional.

2.2.1. 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 $i, j = 1, 2, \dots, K, i \neq j$

2.2.2. Jerárquico

Si los grupos tienen subgrupos, entonces obtenemos un conglomerado jerárquico, que es un conjunto de conglomerados anidados, $H = \{H_1, H_2, \dots, H_Q\}$ ($Q \leq n$) tal que si $C_i \in H_m, C_j \in H_l$ con $m > l$ entonces, $C_i \subset C_j$ o $C_i \cap C_j = \emptyset$ para todo $i \neq j$ y $m, l = 1, 2, \dots, Q$ organizados como árbol.

Cada nodo (grupo) en el árbol (excepto por los nodos de las hojas) es la unión de sus hijos (subgrupos) y la raíz del árbol es el grupo que contiene a todos los objetos.

Hay dos enfoques para construir una jerarquía de grupos:

- Agrupamiento aglomerativo : construye una jerarquía partiendo de grupos pequeños que sucesivamente se van juntando en nodos padre.
- Agrupamiento divisivo : construye una jerarquía de arriba para abajo dividiendo grandes grupos en pequeños, empezando por un grupo que contiene todos los datos.

Un conglomerado jerárquico puede ser visto como una secuencia de conglomerados particionales y un conglomerado particional puede ser obtenido tomando cualquier miembro de esa secuencia, es decir, cortando el árbol jerárquico en un nivel en particular.

2.3. 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 . La matriz se introduce en los algoritmos de conglomerados.

La mayoría de los algoritmos utilizan una matriz de disimilitudes con entradas no negativas y elementos en la diagonal $d_{ii} = 0$, $i = 1, 2, \dots, n$.

2.4. Medidas de disimilitud basada en atributos

Los algoritmos de conglomerados más comunes utilizan como entrada la matriz de disimilitud, así que es necesario construir primero la disimilitud entre pares de observaciones. En el caso más común, definimos una disimilitud $d_j(x_{ij}, x_{i'j})$ entre valores del j -ésimo atributo, y después definimos

$$D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_{ij}, x_{i'j}) \quad (2.1)$$

como la disimilitud entre los objetos i y i' (Hastie et al., 2009). La elección más común es la distancia cuadrática. Sin embargo, hay más elecciones posibles que pueden conducir a resultados muy diferentes. Para atributos no cuantitativos, la distancia cuadrática podría ser poco apropiada. Además, a veces es deseable ponderar los atributos de manera individual.

2.5. K-medias

El algoritmo de K-medias es uno de los métodos iterativos de conglomerados de descenso más populares. Se utiliza para variables de tipo cuantitativo y la medida de similitud entre dos objetos utilizada es la distancia Euclídea al cuadrado:

$$d(x_i, x_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2. \quad (2.2)$$

El objetivo es minimizar :

$$E = \sum_{i=1}^n \|x_i - m_{k(x_i)}\|^2, \quad (2.3)$$

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 .

El algoritmo de descenso iterativo está dado por:

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

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

- 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.

Cada uno de los pasos 1 y 2 reducen el valor de la ecuación 2.3, asegurando convergencia. Sin embargo, el resultado puede corresponder a un mínimo local. Una forma de solucionar esto es empezar el algoritmo con distintas opciones aleatorias y escoger la solución cuyo valor de la función objetivo sea menor.

EL algoritmo de K -medias está relacionado con el algoritmo EM para estimar ciero modelo de mezclas Gaussianas (Hastie et al., 2009).

2.5.1. K-medias esféricas

Cuando se cuenta con datos de dimensiones altas como documentos de texto y canastas de mercado, se ha mostrado que la similitud de cosenos es una métrica superior a la distancia Euclideana. Esta implicación se sigue de que la dirección del vector de un documento, es más importante que su magnitud. La medida de distancia utilizada, que se busca minimizar, es la de disimilitud de cosenos:

$$\begin{aligned} d(x_i, x_{i'}) &= 1 - \cos(x_i, x_{i'}) \\ &= 1 - \frac{\langle x_i, x_{i'} \rangle}{\|x_i\| \|x_{i'}\|} \\ &= 1 - \frac{x_i^T x_{i'}}{\|x_i\| \|x_{i'}\|} \end{aligned} \tag{2.4}$$

Pero minimizar $1 - \cos(x_i, x_{i'})$ es equivalente a maximizar $\cos(x_i, x_{i'})$. Ahora bien, 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, $\mathcal{S}^p = \{x \in \Re^p : x^T x = 1\}$ entonces la ecuación 2.4 se convierte en $d(x_i, x_{i'}) = 1 - x_i^T x_{i'}$. 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

$$L = \sum_{i=1}^n x_i^T \mu_{k(x_i)} \tag{2.5}$$

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 2: Algoritmo K-medias esféricas

Input: Conjunto de n vectores de individuos unitarios

$X = (x_1, x_2, \dots, x_n)$ en \Re^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:

$$\|x - \mu\|^2 = \|x\|^2 + \|\mu\|^2 - 2x^T \mu = 2 - 2x^T \mu = 2(1 - x^T \mu). \quad (2.6)$$

Así pues, en una hiperesfera, maximizar la ecuación 2.5 es equivalente a minimizar la ecuación 2.3. De esta manera, convertimos el problema de optimización cuadrático en uno lineal.

Observación: El problema de optimización

K-medias esféricas como densidad Langevin

El algoritmo de k-medias clásico regresa los estimadores de máxima verosimilitud para las medias de k distribuciones normales con matrices

identidad de covarianza utilizando el algoritmo EM. El algoritmo EM garantiza que dichos estimadores de máxima verosimilitud correspondan a un óptimo local de la función de verosimilitud.

Sean X_1, X_2, \dots, X_n variables aleatorias independientemente distribuidas de una densidad Langevin p -variada $\mathcal{L}_p(\mu; \kappa)$ dada por

$$f(x; \mu, \kappa) = \frac{1}{c_p(\kappa)} \exp(\kappa x^T \mu) \quad (2.7)$$

donde κ es el parámetro de concentración y el coeficiente de normalización c_p está dado por

$$c_p(\kappa) = \frac{2\pi^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)}{\kappa^{\frac{p}{2}-1}} \quad (2.8)$$

donde X_i 's están sobre \mathcal{S}^p y $I_v(\cdot)$ denota la función de Bessel modificada de primer tipo y orden v . La constante de integración $c_p(\kappa)$ se modifica de manera apropiada cuando las X_i 's se encuentran sobre $\mathcal{S}_{\perp 1}^p$. Asumamos que hay k distribuciones Langevin f_h , $h = 1, \dots, k$, de tal forma que cada punto proviene de exactamente una de éstas. Así, lo que se busca es estimar los parámetros de dichas distribuciones de tal manera que la verosimilitud de los datos observados es maximizada. Sea z_{ih} el índice que indica si x_i fue generada por f_h , dado que las X_i 's son independientes, la función de verosimilitud es

$$L(\Theta; \mathcal{X}) = \prod_{i=1}^n \sum_{h=1}^k z_{ih} f(x_i; \mu_h) \quad (2.9)$$

donde μ_h es la media de la h -ésima distribución Langevin, $\Theta = (\kappa, \mu_1, \mu_2, \dots, \mu_k)$ y $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$. Para un valor dado de los parámetros la distribución

más verosímil de la cual proviene x_i (en términos de la log-verosimilitud) está dada por

$$\begin{aligned} h^* &= \operatorname{argmax}_h \log f(x_i | \mu_h) \\ &= \operatorname{argmax}_h \kappa x_i^T \mu_h - \log(c_p(\kappa)) \\ &= \operatorname{argmax}_h x_i^T \mu_h. \end{aligned} \quad (2.10)$$

Utilizando 2.10 para asignar cada x_i a una distribución, es posible reestimar las medias de las distribuciones correspondientes como sigue:

$$\mu_h = \frac{\sum_{x_i \in f_h} x_i}{\|\sum_{x_i \in f_h} x_i\|}, \quad h = 1, \dots, k. \quad (2.11)$$

Repetiendo los pasos dados en las ecuaciones 2.10 y 2.11 resulta en un esquema de ascenso gradiente cuya convergencia garantiza un mínimo local de la función de verosimilitud 2.9. A este algoritmo se le conoce como k -medias esféricas pues las observaciones se encuentran sobre la superficie de la esfera unitaria. El desempeño de dicho algoritmo se puede evaluar a través de la log-verosimilitud normalizada dada por:

$$\mathcal{J} = \frac{1}{n} \sum_{h=1}^k \sum_{x \in f_h} x^T \mu_h. \quad (2.12)$$

2.5.2. Estadístico Gap para determinar K

Para poder aplicar el algoritmo de k -medias se debe seleccionar primero el número de grupos K^* y un punto inicial. El número de grupos a escoger

depende de la aplicación. Si se desea segmentar datos, K está definida como parte del problema. Sin embargo, el análisis de conglomerados se utiliza con frecuencia para proporcionar un estadístico descriptivo para determinar si las observaciones cuentan con una agrupación natural. En este caso, el número de grupos K^* es desconocido y es necesario estimarlo a través de los datos.

Supongamos que se han agrupado las observaciones en k grupos C_1, C_2, \dots, C_K donde C_h denota el índice de las observaciones en el conglomerado h y $n_r = |C_r|$. Sea

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \quad (2.13)$$

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

$$W_k = \sum_{r=1}^k \frac{1}{2n_r} D_r. \quad (2.14)$$

La idea de este enfoque es estandarizar la gráfica de $\log W_k$ comparándola con su valor esperado bajo una distribución de referencia nula de los datos. La importancia de escoger una distribución nula apropiada es demostrada en Gordon (2004). Así, 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

$$\text{Gap}_n(k) = E_n^* [\log(W_k)] - \log(W_k), \quad (2.15)$$

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 E_n^* . Se elige la distribución uniforme como distribución de referencia, distribuyendo los datos sobre un rectángulo que contiene los datos. Tibshirani et al. (2001) resume la implementación computacional del estadístico Gap como sigue:

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 simulados usando la distribución uniforme. Agrupar cada conjunto y calcular $W_{Kb}^*, b = 1, 2, \dots, B, K = 1, 2, \dots, M$ y estimar la estadística Gap de la siguiente manera

$$\text{Gap} = \frac{1}{B} \sum_{b=1}^B \log W_{Kb}^* - \log W_k. \quad (2.16)$$

3. Sea $\bar{l} = \frac{1}{B} \sum_{b=1}^B \log W_{Kb}^*$, calcular la desviación estándar

$$sd_K = \left[\frac{1}{B} \sum_{b=1}^B (\log W_{Kb}^* - \bar{l})^2 \right]^{\frac{1}{2}} \quad (2.17)$$

y definir

$$s_K = sd_K \sqrt{1 + \frac{1}{B}}. \quad (2.18)$$

Finalmente, escoger \hat{K} donde \hat{K} es la menor K tal que

$$\text{Gap}(K) \geq \text{Gap}(K+1) - s_{K+1}. \quad (2.19)$$

Capítulo 3

Análisis Estadístico Espacial

Este capítulo pretende introducir algunos métodos para analizar datos espaciales cuantitativos. *Espacial* significa que cada elemento de los datos tiene referencia geográfica de tal manera que podamos saber en qué parte del mapa ocurre cada caso. Dicha referencia espacial es importante porque acarrea información relevante para el análisis de los datos.

3.1. ¿Qué es el análisis espacial?

El término “análisis espacial” es muy utilizado en la literatura de los Sistemas de Información Geográfica (GIS por sus siglas en inglés) y de Ciencia de Información Geográfica (GISc por sus siglas en inglés). El análisis espacial es un conjunto de técnicas y modelos que usan la referencia espacial asociada a cada individuo de los datos. Los métodos de análisis espacial necesitan hacer supuestos para describir la asociación o interacción espacial entre los casos. Los resultados de cualquier análisis espacial no son los

mismos bajo re-arreglos de la distribución espacial o reconfiguraciones de la estructura espacial del sistema bajo investigación.

El análisis espacial tiene tres elementos principales:

1. *Modelado cartográfico.* El conjunto de datos es representado como un mapa y las operaciones basadas en mapas generan nuevos mapas.
2. *Modelado matemático.* Las salidas del modelo dependen de la forma de modelar la interacción espacial entre los objetos, de las relaciones espaciales o del posicionamiento geográfico de los objetos dentro del modelo.
3. *Análisis estadístico espacial.* Consiste en el desarrollo y la aplicación de técnicas estadísticas para analizar los datos espaciales, por consiguiente, hace uso de las referencias espaciales en el conjunto de datos.

Se debe de tener cuidado al hacer el análisis estadístico. A pesar de que el análisis de dependencia espacial es una pieza clave del análisis estadístico espacial, si se pone demasiada atención a las características espaciales de los datos, se pueden llegar a ignorar otras características importantes de estos. El análisis estadístico espacial es sólo un subcampo del análisis estadístico. Para hacer un análisis correcto de los datos espaciales, hay un rol muy importante de las otras áreas de la teoría estadística donde no se involucra el uso de datos espaciales.

3.2. Estadística Espacial

Los datos espaciales se distinguen por observaciones obtenidas en ubicaciones espaciales s_1, s_2, \dots, s_n donde s_i 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 (e.g. países, regiones o lagos).

Los modelos espaciales intentan modelar la correlación entre observaciones en diferentes posiciones en el espacio. En estadística espacial, se cuenta principalmente con tres tipos de análisis de datos:

- **Geoestadística.** Se caracteriza por datos que involucran la medición de la variable de interés sobre una región espacial (e.g. país, municipio, lago) donde la medición de los datos puede variar de manera continua.

Sea D la región de interés, cada punto $s = (x, y)$ en D puede ser descrito por un par coordenado x y y en el plano.

- **Patrón de Puntos Espaciales.** Consideremos la región D , en ésta estamos interesados en la ubicación de ciertos “eventos”. Nos preguntamos si los eventos de interés ocurren de manera aleatoria a lo largo del área o si los eventos tienden a aglomerarse.

El problema con tratar de determinar la presencia de aglomeraciones en un conjunto de datos espaciales, puede ser muy complicado. Puntos generados de manera completamente aleatoria, pueden parecer aglomerados.

- **Datos en retícula.** Determinamos la región D como una colección finita de sitios espaciales donde se llevan a cabo las observaciones. A

la colección de puntos en D se le conoce como celosía o retícula (en inglés, “lattice”). Dichos sitios espaciales en la retícula, son identificados normalmente utilizando su longitud y latitud.

Hay tres características de los datos en celosía a tomar en cuenta:

1. ¿La retícula es regular o irregular? Por ejemplo, los estados de México son irregulares; pero, por otro lado, podemos tener mediciones sobre un campo agrícola particionado en bloques regulares.
2. ¿Las ubicaciones en la celosía hacen referencia a “puntos” o “regiones”? En los estados de México, cada estado es una región.
3. ¿La variable de interés es categórica o numérica?

Una forma sencilla de ilustrar gráficamente los datos en retícula es utilizando un *mapa coroplético*. Un mapa coroplético es un mapa que muestra las regiones o áreas con características similares. Por ejemplo, si queremos ver el índice de marginación en México dividido por municipio, como se ve en la Figura 3.1 los municipios con índices similares tienen colores similares.

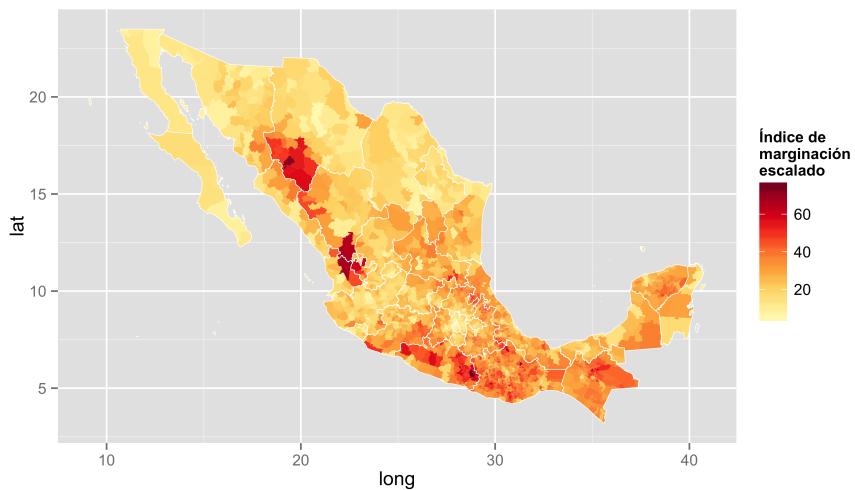


Figura 3.1: Índice de marginación por municipio.

Capítulo 4

Autocorrelación Espacial

Las observaciones realizadas en diferentes puntos espaciales pueden no ser independientes. Por ejemplo, medidas obtenidas en lugares cercanos pueden ser más parecidas que en lugares lejanos. 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 (Cliff y Ord, 1973, 1981).

Las pruebas de autocorrelación espacial examinan si el valor observado de una variable en algún lugar es independiente de valores de la misma variable en lugares cercanos o ontiguos.

Un índice de *autocorrelación espacial global* resume el nivel de similitud espacial observada entre las observaciones vecinas sobre el área entera estudiada.

- **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.

La mayoría de los índices de autocorrelación comparten una estructura común. En dicha estructura, se calcula la similitud entre valores en las localidades i y j , luego se pondera dicha similitud por la proximidad entre éstos. Altas similitudes con mucho peso provocan un valor alto del índice, mientras que bajas similitudes con mucho peso provocan un valor bajo del índice.

4.1. Matriz de Pesos

Para evaluar la autocorrelación espacial, se debe definir primero qué significa que dos observaciones sean cercanas, es decir, se debe definir una métrica de distancia. Dichas distancias son presentadas en una matriz de pesos en la cual se define la relación entre los diferentes lugares donde se obtuvieron las observaciones. Si los datos fueron obtenidos en n lugares distintos, entonces la matriz de pesos W es de $n \times n$, donde cada entrada w_{ij} , $i, j = 1, 2, \dots, n$ representa la dependencia espacial o peso entre los lugares i y j definiendo así la estructura de los vecinos sobre el área entera. Haining (2004) presenta las siguientes maneras para construir los pesos:

- **Contigüidad binaria:** Es la definición de pesos más sencilla, se de-

fine como

$$w_{ij} = \begin{cases} 1 & \text{si la región } i \text{ comparte frontera con } j \\ 0 & \text{e.o.c.} \end{cases} \quad (4.1)$$

Nótese que con dicha elección de medida de proximidad W es necesariamente simétrica , pues $w_{ij} = w_{ji}$.

- **Distancia:** Se definen los pesos como $w_{ij} = d_{ij}^{-\delta}$ donde d_{ij} denota la distancia entre i y j y el parámetro $\delta \geq 0$. La distancia puede definirse de distintas maneras (e.g. distancia Euclídea).
- **Función exponencial de distancia:** Se define como $w_{ij} = e^{d_{ij}^{-\delta}}$.
- **Frontera en común:** Sean l_{ij} la longitud de la frontera entre i y j y l_i la longitud de la frontera de la región i , se pueden definir los pesos como $w_{ij} = \left(\frac{l_{ij}}{l_i}\right)^\tau$ con $\tau \geq 0$.
- **Combinación de frontera y distancia:** Es una combinación entre los pesos por distancia y por tamaño de frontera, se define como $w_{ij} = \left(\frac{l_{ij}}{l_i}\right)^\tau d_{ij}^{-\delta}$ donde $\tau, \delta \geq 0$.

Nota: Para todos los casos $w_{ii} = 0$ y $w_{ij} \geq 0$ para $i, j = 1, 2, \dots, n$,

obteniendo así una matriz de la siguiente forma,

$$W = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1n} \\ w_{21} & 0 & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \cdots & 0 \end{pmatrix}. \quad (4.2)$$

Podemos ajustar W de tal forma que la suma de los pesos por renglón sea igual a 1 utilizando una matriz estandarizada por filas donde dividimos cada elemento w_{ij} por la suma de los pesos de los vecinos de la región i obteniendo una matriz W_{std} donde

$$w_{std,ij} = \frac{w_{ij}}{\sum_{j=1}^n w_{ij}}. \quad (4.3)$$

En el caso de los pesos binarios, dicha estandarización penaliza los pesos de las regiones con muchos vecinos.

4.2. Pruebas de Autocorrelación Espacial Global

Consideremos 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 . Para cada par de regiones i y j , si las observaciones y_i y y_j no están correlacionadas, entonces decimos que no hay autocorrelación espacial en el área de estudio para la variable Y . De manera inversa, decimos que existe autocorrelación espacial si las observaciones están correlacionadas por pares. Las pruebas de autocorrelación espacial propuestas en la literatura,

dependen del tipo de la variable de estudio (discreta, ordinal o continua).

4.2.1. Variables continuas u ordinales

Si Y es de escala continua u ordinal, utilizamos dos coeficientes que miden el grado de autocorrelación entre las y_i 's en regiones unidas, donde y_i es una observación en la región i .

Índice \mathcal{I} de Moran

El índice \mathcal{I} de Moran propuesto por Moran (1950) sirve para probar la autocorrelación espacial global para variables continuas. Se basa en los productos cruzados de las desviaciones de la media y se calcula para n observaciones de una variable y en lugares i, j como sigue:

$$\mathcal{I} = \left(\frac{n}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2} \right). \quad (4.4)$$

y_i es el valor de la variable en el lugar i y \bar{y} es la media muestral.

\mathcal{I} no es como un coeficiente de correlación común, pues no pertenece necesariamente al intervalo $(-1, 1)$. Usualmente $\mathcal{I} \in (-1, 1)$, a menos que se cuente con regiones con valores extremos de $y_i - \bar{y}$ con pesos muy altos.

Índice \mathcal{C} de Geary

El índice \mathcal{C} de Geary utiliza la suma de diferencias al cuadrado entre pares de observaciones como medida de variación. Fue sugerido por Geary

(1954) y está dado por

$$\mathcal{C} = \left(\frac{n-1}{2 \sum_{i=1}^n \sum_{j=1}^n w_{ij}} \right) \left(\frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij}(y_i - y_j)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \right). \quad (4.5)$$

El índice \mathcal{C} de Geary toma valores en el intervalo $[0, 2]$ donde 0 indica correlación espacial positiva perfecta (i.e. $y_i = y_j$ para cualquier par de regiones donde $w_{ij} > 0$) y 2 indica correlación espacial negativa perfecta.

\mathcal{C} no es propiamente un coeficiente de correlación, en cambio, corresponde a una estadístico d de Durbin-Watson (Cliff y Ord, 1981), utilizada para probar autocorrelación serial en análisis de regresión y en análisis de series de tiempo.

En contraste con \mathcal{I} , valores bajos de \mathcal{C} denotan autocorrelación espacial positiva y valores altos indican autocorrelación espacial negativa.

Ambos índices \mathcal{I} y \mathcal{C} tienen la forma clásica de un coeficiente de autocorrelación: el numerador en cada uno es una medida de covarianza entre las y_i 's y el denominador es una medida de varianza. Es evidente también que \mathcal{I} está basado en productos cruzados de las desviaciones entre y_i y \bar{y} , opuesto a las diferencias cuadráticas entre las y_i 's del coeficiente de Geary.

Se puede mostrar (Cliff y Ord, 1973, Capítulo 2) que \mathcal{I} y \mathcal{C} asintóticamente, 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.** Bajo este supuesto, asumimos que las observaciones y_i son resultado de n realizaciones de una población normal.

2. Aleatorización. Independientemente de la distribución subyacente de la población, 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. Hay $n!$ valores posibles.

Usando los subíndices N y R para denotar los supuestos de normalidad y aleatorización respectivamente. Si utilizamos pesos w_{ij} simétricos (i.e. $w_{ij} = w_{ji}$) se puede probar que los momentos de los índices quedan como sigue.

Coeficiente \mathcal{I}

$$E_N [\mathcal{I}] = E_R [\mathcal{I}] = -\frac{1}{n-1}, \quad (4.6)$$

$$E_N [\mathcal{I}^2] = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{S_0^2 (n^2 - 1)}, \quad (4.7)$$

$$E_R [\mathcal{I}^2] = \frac{n [(n^2 - 3n + 3)S_1 - n S_2 + 3 S_0^2] - b_2 [(n^2 - n)S_1 - 2n S_2 + 6 S_0^2]}{(n-1)(n-2)(n-3)S_0^2}. \quad (4.8)$$

Coefficiente \mathcal{C}

$$E_N[\mathcal{C}] = [\mathcal{C}] = 1, \quad (4.9)$$

$$\text{Var}_N(\mathcal{C}) = \frac{(2S_1 + S_2)(n - 1) - 4S_0^2}{2(n + 1)S_0^2}, \quad (4.10)$$

$$\begin{aligned} \text{Var}_R(\mathcal{C}) = & \left((n - 1)S_1 [n^2 - 3n + 3 - (n - 1)b_2] \right. \\ & - \frac{1}{4}(n - 1)S_2 [n^2 + 3n - 6 - (n^2 - n + 2)b_2] \\ & \left. + S_0^2 [n^2 - 3 - (n - 1)^2 b_2] \right) \left(\frac{1}{n(n - 2)(n - 1)S_0^2} \right). \end{aligned} \quad (4.11)$$

Donde

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}, \quad (4.12)$$

$$S_1 = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (w_{ij} + w_{ji})^2, \quad (4.13)$$

$$S_2 = \sum_{i=1}^n (w_{i\cdot} + w_{\cdot i})^2, \quad (4.14)$$

$$w_{i\cdot} = \sum_{j=1}^n w_{ij} \quad y \quad w_{\cdot i} = \sum_{j=1}^n w_{ij}. \quad (4.15)$$

Observación : Hoeffding (1952) demostró que las distribuciones asintóticas bajo N y R son las mismas bajo condiciones generales razonables.

4.2.2. Variables Discretas: Estadístico join-count (Conteo de fronteras)

Cuando la variable de interés y es categórica se puede utilizar la estadística join-count para medir el grado de dispersión o conglomeración entre las distintas clases.

Supongamos primero que y cuenta con 2 clases, es decir, es binaria y $y_i \in \{0, 1\}$. Mapeando y en 2 colores, W (blanco por su inicial en inglés) si $y_i = 0$ y B (negro por su inicial en inglés) 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.

El método de análisis es como sigue (Moran, 1948). El conteo de uniones BB ponderado por w_{ij} , está dado por

$$BB = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} y_i y_j, \quad (4.16)$$

el número de uniones BW ponderado es

$$BW = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - y_j)^2, \quad (4.17)$$

y el número de uniones WW ponderado es

$$WW = W - (BB + BW), \quad (4.18)$$

donde

$$W = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_{ij}. \quad (4.19)$$

Nótese que utilizando pesos w_{ij} binarios (ver (4.1)) BB , BW y WW son el conteo observado de uniones de cada tipo y el factor $\frac{1}{2}$ en (4.16), (4.17), (4.19) elimina el conteo duplicado, al contar las uniones ij y ji por la propiedad de simetría de los pesos binarios.

El método usual para determinar si BB , BW y WW distan significativamente del conteo esperado por sorteo es utilizando el hecho de que dichos estadísticos de conteo de fronteras, asintóticamente se distribuyen $\mathcal{N}(\mu, \sigma^2)$ (Cliff y Ord, 1973, Capítulo 2). Bajo este enfoque, los parámetros μ y σ^2 de los coeficientes pueden ser evaluados bajo uno de las dos supuestos:

1. **Muestreo con reemplazo**, donde suponemos que 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**, donde suponemos que 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 $n_a + n_b = n$.

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:

$$n_1 + n_2 + \dots + n_k = n. \quad (4.20)$$

Sean N_{rr} el número de uniones entre regiones del tipo rr , N_{rs} el número de uniones del tipo rs , con $r, s \in \{1, 2, \dots, k\}$ y J_{tot} el número de fronteras entre todas las regiones de distintas clases.

Ahora, el análisis procede haciendo el conteo ponderado de fronteras entre regiones de la misma categoría, dos categorías diferentes y todas las regiones de color diferente. Cada una de las distribuciones es evaluada de la manera planteada anteriormente.

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

Sea p_r la probabilidad de que una región sea de color r , los parámetros μ y σ^2 están dados por Moran (1948) como sigue.

Muestreo con reemplazo

Uniones entre regiones del mismo color N_{rr} (equivalente a BB para $k = 2$)

$$\mu = \frac{1}{2} S_0 p_r^2, \quad (4.21)$$

$$\sigma^2 = \frac{1}{4} [S_1 p_r^2 + (S_2 - 2S_1)p_r^3 + (S_1 - S_2)p_r^4]. \quad (4.22)$$

Uniones entre regiones de dos colores diferentes N_{rs} (equivalente a BW para $k = 2$)

$$\mu = S_0 p_r p_s, \quad (4.23)$$

$$\sigma^2 = \frac{1}{4} [2S_1 p_r p_s + (S_2 - 2S_1)p_r p_s(p_r + p_s) + 4(S_1 - S_2)p_r^2 p_s^2]. \quad (4.24)$$

Número total de uniones entre regiones de diferentes colores J_{tot} ($k \geq 3$; cuando $k = 2$, este caso es igual al número de fronteras BW)

$$\mu = S_0 \sum_{r=1}^{k-1} \sum_{s=r+1}^k p_r p_s, \quad (4.25)$$

$$\begin{aligned} \sigma^2 &= \frac{S_2}{4} \sum_{r=1}^{k-1} \sum_{s=r+1}^k p_r p_s (2S_1 - 5S_2) \sum_{r=1}^{k-2} \sum_{s=r+1}^{k-1} \sum_{t=s+1}^k p_r p_s p_t \\ &\quad + (S_1 - S_2) \left(\sum_{r=1}^{k-1} \sum_{s=r+1}^k p_r^2 p_s^2 - 2 \sum_{r=1}^{k-3} \sum_{s=r+1}^{k-3} \sum_{t=s+1}^{k-1} \sum_{u=t+1}^k p_r p_s p_t p_u \right). \end{aligned} \quad (4.26)$$

Muestreo sin reemplazo

Uniones entre regiones del mismo color.

$$\mu = \frac{S_0 n_r (n_r - 1)}{2n(n-1)}, \quad (4.27)$$

$$\sigma^2 = \frac{S_1 n_r^{(2)}}{4n^{(2)}} + \frac{(S_2 - 2S_1) n_r^{(3)}}{4n^{(3)}} + \frac{(S_0^2 + S_1 - S_2) n_r^{(4)}}{4n^{(4)}} - \mu^2. \quad (4.28)$$

Uniones entre regiones de dos colores diferentes.

$$\mu = \frac{S_0 n_r n_s}{n(n-1)}, \quad (4.29)$$

$$\begin{aligned} \sigma^2 &= \frac{S_1 n_r n_s}{2n^{(2)}} + \frac{(S_2 - 2S_1) n_r n_s (n_r + n_s - 2)}{4n^{(3)}} \\ &\quad + \frac{(S_0^2 + S_1 - S_2) n_r^{(2)} n_s^{(2)}}{n^{(4)}} - \mu^2. \end{aligned} \quad (4.30)$$

Número total de uniones entre regiones de diferentes colores

$$\mu = S_0 \sum_{r=1}^{k-1} \sum_{s=r+1}^k \frac{n_r n_s}{n^{(2)}}, \quad (4.31)$$

$$\begin{aligned} \sigma^2 &= \left[\frac{S_2}{4n^{(2)}} - \frac{(S_0^2 + S_1 - S_2)(n-1)}{4n^{(4)}} \right] \sum_{r=1}^{k-1} \sum_{s=r+1}^k n_r n_s \\ &+ \left[\frac{(S_1 - S_2)}{n^{(4)}} + \frac{S_0^2(2n-3)}{n^{(2)}n^{(4)}} \right] \sum_{r=1}^{k-1} \sum_{s=r+1}^k n_r^2 n_s^2 \\ &+ \left[\frac{2S_1 - 5S_2}{2n^{(3)}} + \frac{3(S_0^2 + S_1 - S_2)}{n^{(4)}} + \frac{2S_0^2}{n^{(3)}(n-1)} \right] \sum_{r=1}^{k-2} \sum_{s=r+1}^{k-1} \sum_{t=s+1}^k n_r n_s n_t \\ &- 2 \left[\frac{S_1 - S_2}{n^{(4)}} + \frac{2S_0^2(2n-3)}{n^{(2)}n^{(4)}} \right] \sum_{r=1}^{k-3} \sum_{s=r+1}^{k-2} \sum_{t=s+1}^{k-1} \sum_{u=t+1}^k n_r n_s n_t n_u. \end{aligned} \quad (4.32)$$

4.2.3. Pruebas de hipótesis

La hipótesis nula H_0 es de no autocorrelación espacial, es decir, de aleatoriedad espacial:

- Los valores observados en cierta región, no dependen de los valores observados en regiones vecinas.
- El patrón espacial observado es igual de probable que cualquier otro patrón espacial.
- La ubicación de los valores puede ser alterada sin alterar el contenido de información de los datos.

Hay dos procedimientos a seguir de acuerdo al supuesto bajo el que estamos trabajando.

Bajo el supuesto de normalidad

▪ Conteo de fronteras

Utilizando el estadístico de conteos del mismo color N_{rr} , $r \in \{1, 2, \dots, k\}$ podemos hacer una prueba cada una de las categorías de manera independiente, obteniendo k pruebas distintas. El estadístico de prueba es

$$z = \frac{N_{rr} - \mu}{\sqrt{\sigma^2}} \sim \mathcal{N}(0, 1). \quad (4.33)$$

Se calculan μ y σ^2 de acuerdo al supuesto bajo el que estemos trabajando. Utilizamos el supuesto de muestreo con reemplazo si las p_i 's son conocidas a priori. Si dichas probabilidades son estimadas a partir de los datos por $\frac{n_i}{n}$ ($i = 1, 2, \dots, k$) debemos utilizar muestreo sin reemplazo.

También podemos probar sobre N_{rs} ó J_{tot} .

▪ Estadísticos \mathcal{I} y \mathcal{C}

De igual manera que el punto anterior, utilizamos un estadístico z .

Para \mathcal{I} calculamos

$$z = \frac{\mathcal{I} - \text{E}[\mathcal{I}]}{\sqrt{\text{Var}(\mathcal{I})}}, \quad (4.34)$$

mientras que para \mathcal{C}

$$z = \frac{\text{E}[\mathcal{C}] - \mathcal{C}}{\sqrt{\text{Var}(\mathcal{C})}}. \quad (4.35)$$

Recordemos que el coeficiente de Geary está construido de tal forma que, bajo H_0 , $c = 1$; valores de $\mathcal{C} < 1$ indican autocorrelación espacial

positiva; y valores de $\mathcal{C} < 1$ indican autocorrelación espacial negativa.

Entonces, calculamos $E[\mathcal{C}] - \mathcal{C}$ en vez de $\mathcal{C} - E[\mathcal{C}]$ de tal forma que valores positivos del estadístico correspondan a autocorrelación espacial positiva, y valores negativos a autocorrelación espacial negativa.

Nótese que podemos utilizar dicho estadístico de prueba para cualquiera de los dos supuestos de normalidad y aleatorización ya que, asintoticamente y bajo condiciones regulares tienen la misma distribución (ver 4.2.1).

Simulaciones de Monte Carlo

Si dudamos del supuesto de normalidad, podemos utilizar simulaciones de Monte Carlo para examinar la forma de la función de densidad de los coeficientes de autocorrelación espacial bajo la hipótesis nula.

Es recomendable hacer esta prueba ya que la función de densidad del estadístico es sensible a los siguientes factores (Cliff y Ord, 1981):

1. La forma de las regiones en el área de estudio y el número promedio de fronteras por región.
2. Los pesos w_{ij} utilizados.
3. La distribución de la variable Y .
4. El tamaño de la muestra n .

El proceso de muestreo es el siguiente:

1. Permutamos aleatoriamente las etiquetas y_1, y_2, \dots, y_n a través de las regiones. Por ejemplo, el mapa 4.1 es una permutación aleatoria del mapa 3.1.

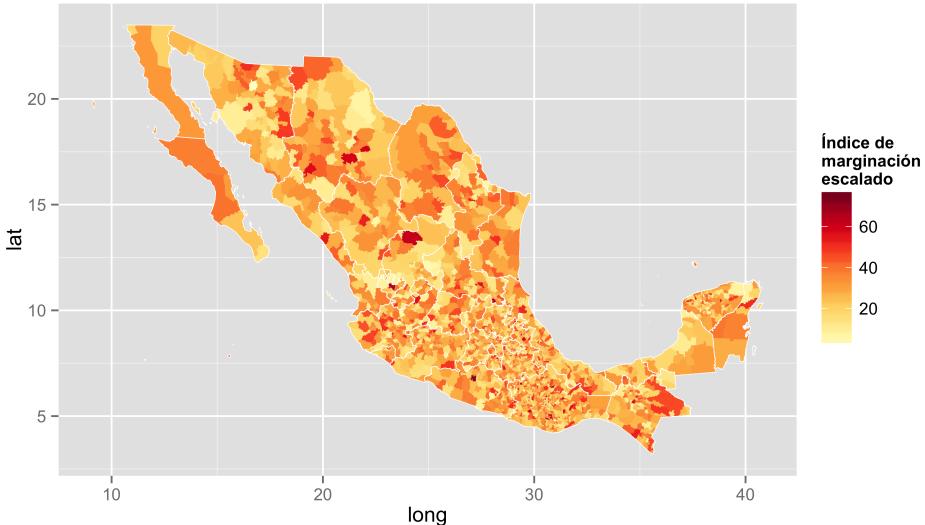


Figura 4.1: Permutación aleatoria del índice de marginación por municipio.

2. Calculamos el estadístico de interés, digamos \mathcal{I} , con las etiquetas permutadas. Si la variable Y es continua hay $n!$ permutaciones posibles; si es discreta, $\binom{n}{n_1 n_2 \dots n_k}$.
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. Si la hipótesis alternativa. Si la \mathcal{I} observada cae en un área mayor a $(1 - \alpha) \%$ o menor a $\alpha \%$, entonces \mathcal{I} es significativamente (positivo o negativo) a un nivel de significancia α .

4.2.4. Diagrama de dispersión de Moran

Un enfoque para visualizar asociación espacial se basa en el concepto del diagrama de dispersión de Moran propuesto por Anselin (1993) que compara la variable de interés y_i contra su retraso espacial (promedio ponderado de Y en las regiones vecinas de i), para $i = 1, 2, \dots, n$.

Si restamos \bar{y} a ambos valores, la nube queda partida por los ejes en cuatro cuadrantes. Puntos en los cuadrantes Alto-Alto y Bajo-Bajo indican autocorrelación espacial positiva; mientras que puntos los cuadrantes Alto-Bajo y Bajo-Alto, autocorrelación espacial negativa.

Si colocamos el retraso espacial en el eje vertical y el valor en cada región en el eje de horizontal, el índice I corresponde a la pendiente de la línea de regresión ajustada.

Parte II

Resultados y Conclusión

Capítulo 5

Resultados

5.1. Descripción de la base

La información sobre la cual se trabajó corresponde a la base de datos de CONAPO (Consejo Nacional de Población): “Índice de Marginación por Entidad Federativa y Municipio 2010”. La base cuenta con 2,456 observaciones y cada una tiene 15 atributos descritos en la tabla 5.1 que corresponden a variables para medir el grado de marginación de un municipio.

Para cada observación, se cuenta con un polígono geolocalizado que corresponde a la forma del municipio o delegación. Entonces, nuestra área de estudio espacial es la República Mexicana, que corresponde a una retícula irregular, donde cada región corresponde a un municipio.

Las variables analf, sprim, sdren, selec, sagua, hacina, pisot, l5khab y bingreso son indicadores de marginación definidos por CONAPO (2004)

Cuadro 5.1: Leyenda variables de marginación 2010.

Variable	Descripción
clave_ent	Clave de la entidad federativa
clave_mun	Clave de municipio
nom_mun	Nombre del municipio
poblac	Población total
analf	% de Población de 15 años o más analfabeta
sprim	% de Población de 15 años o más sin primaria completa
sdren	% Ocupantes en viviendas sin drenaje ni excusado
selec	% Ocupantes en viviendas sin energía eléctrica
sagua	% Ocupantes en viviendas sin agua entubada
hacina	% Viviendas con algún nivel de hacinamiento
pisot	% Ocupantes en viviendas con piso de tierra
pl5khab	% Población en localidades con menos de 5 000 habitantes
bingreso	% Población ocupada con ingreso de hasta 2 salarios mínimos
imarg	Índice de marginación
gmarg	Grado de marginación
imarges	Índice de marginación escalado de 0 a 100
lugar	Lugar que ocupa a nivel nacional

(Consejo Nacional de Población) que miden el nivel de marginación en cuatro dimensiones: educación, vivienda, distribución de la población e ingresos monetarios.

El valor del índice de marginación es la primera componente del método de componentes principales, aplicado a los nueve indicadores mencionados; una vez determinados los valores para cada área, se clasifican en cinco grupos diferenciados y delimitados mediante la técnica de estratificación óptima de Dalenius y Hodges (CONAPO, 2011).

5.2. Análisis Exploratorio

Empezamos observando el diagrama de dispersión de Moran 5.1 que compara el índice de marginación de un municipio contra el promedio de índice de marginación de sus vecinos, podemos observar que hay una correlación positiva muy alta. Esto es un síntoma de autocorrelación espacial positiva, pues esperamos valores parecidos de índice de marginación entre vecinos.

Los puntos que están por arriba de la nube de puntos tienen menor índice de marginación con respecto a sus vecinos; en contraste, puntos por debajo de la nube, tienen mayor índice de marginación con respecto a sus vecinos.

Se realizó una regresión lineal simple para encontrar aquellas observaciones cuyos residuales son altos.

Algunos casos interesantes son los siguientes:

- Podemos ver que San Mateo del Mar está muy lejos y por debajo de

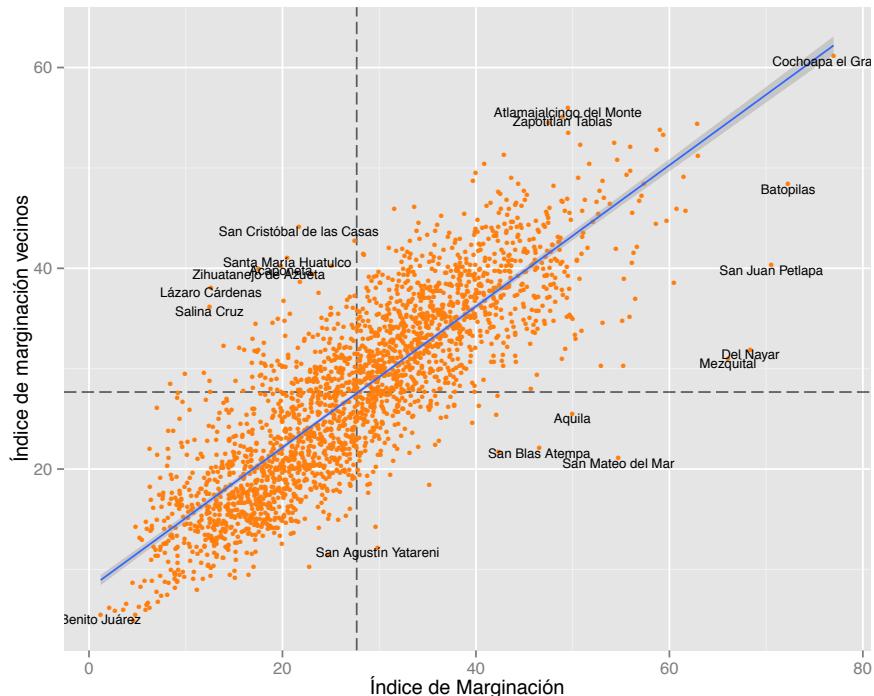


Figura 5.1: Gráfica de Moran para índice de marginación.

la nube de puntos, su índice de marginación es muy alto en comparación con el de sus vecinos. En contraste, vemos en la parte de arriba a la izquierda al municipio de Salina Cruz, que tiene un índice de marginación más bajo. De hecho, ambos municipios son vecinos en el estado de Oaxaca. Salina Cruz es un importante centro industrial, debido a la presencia de la Refinería Ing. Antonio Dovalí Jaime de PEMEX.

- Otros municipios que se encuentran muy por debajo de la curva, son Mezquital, en Durango y Del Nayar, en Nayarit. Estos municipios

ocupan el quinto y el tercer lugar de marginación a nivel nacional respectivamente. Aunque estén en estados diferentes, ambos municipios comparten frontera y por lo tanto, tienen características similares. Los dos son municipios que se mantuvieron al margen de la evolución ocurrida en el resto de su entidad correspondiente, por la concentración de población indígena. Ambos ocupan el primer lugar de marginación dentro de su entidad. Se pueden apreciar en el mapa 5.2 de color rojo intenso, al sur de Durango y norte de Nayarit.

- En el caso de San Cristobal de las Casas, está por arriba de la nube de puntos. Es el municipio de Chiapas con menor índice de marginación. Esto se puede deber al turismo y a las inversiones en la región.

El mapa coroplético 5.2 muestra de manera clara que hay una alta autocorrelación espacial positiva en el índice de marginación. Se puede observar que en general hay una transición suave entre colores oscuros y claros, es decir, entre municipios vecinos hay colores muy parecidos, lo que indica valores de marginación cercanos.

Encontramos puntos rojos oscuros de alta marginación en lugares como la Sierra Tarahumara al suroeste de Chihuahua, sur de Durango y norte de Nayarit (Mezquital y Del Nayar), límites entre Puebla y Veracruz, los estados de Oaxaca, Guerrero y Chiapas, etc.

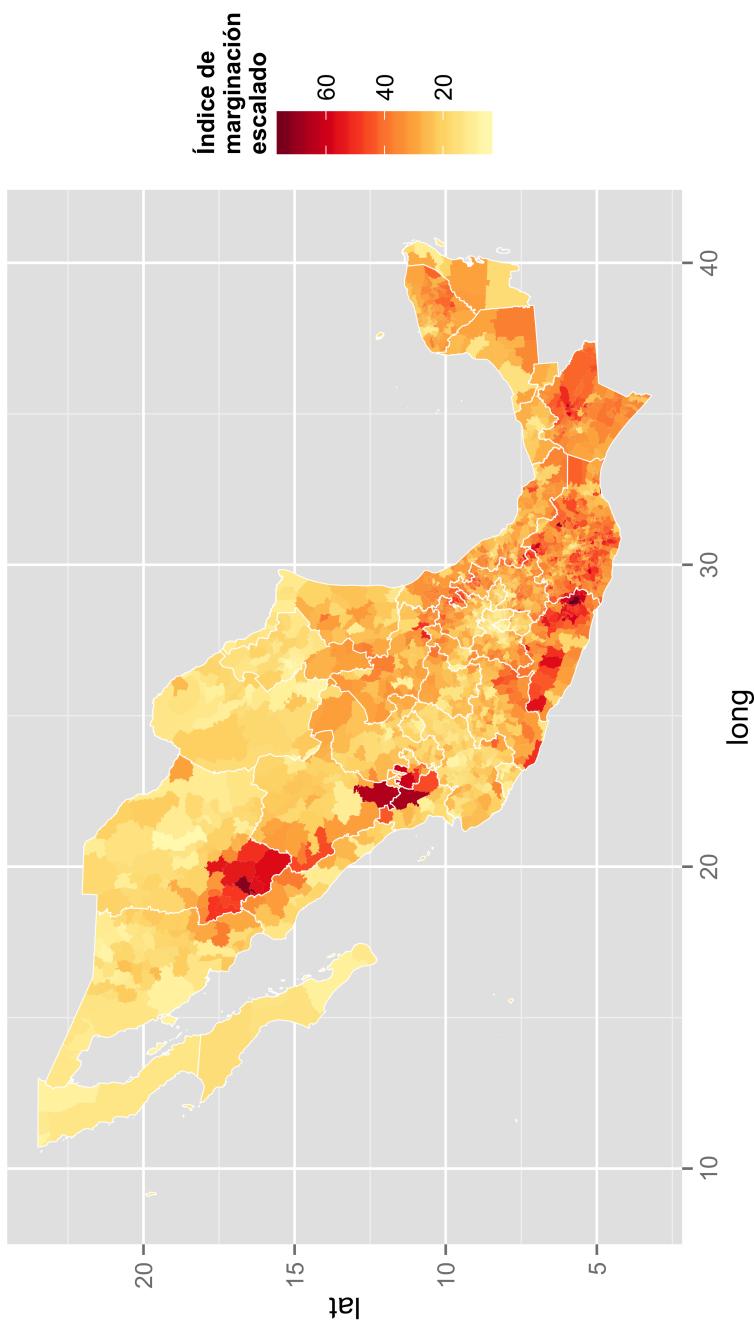


Figura 5.2: Mapa de México coloreando cada municipio por su índice de marginación.

Las tabla 5.2 muestra los municipios más marginados. Aparecen los casos de Mezquital y Del Nayar mencionados anteriormente. Batopilas es un municipio que se encuentra en la Sierra Tarahumara en Chihuahua; San Juan Petlapa, en Oaxaca y Cochoapa el Grande, el municipio más marginado, en Guerrero.

nom_mun	gmarg	imarges	lugar
Cochoapa el Grande	Muy alto	76.98	1
Batopilas	Muy alto	72.27	2
San Juan Petlapa	Muy alto	70.55	3
Del Nayar	Muy alto	68.38	4
Mezquital	Muy alto	66.06	5

Cuadro 5.2: Tabla de municipios más marginados,

La lista de los municipios menos marginados se encuentra en la tabla 5.3. Destacan tres delegaciones del Distrito Federal (Coyoacán, Miguel Hidalgo y Benito Juárez) y dos municipios que forman parte de la zona metropolitana de la Ciudad de Monterrey en Nuevo León (San Pedro Garza García y San Nicolás de los Garza).

nom_mun	gmarg	imarges	lugar
Coyoacán	Muy bajo	3.86	2,452
Miguel Hidalgo	Muy bajo	3.56	2,453
San Nicolás de los Garza	Muy bajo	2.69	2,454
San Pedro Garza García	Muy bajo	2.09	2,455
Benito Juárez	Muy bajo	1.21	2,456

Cuadro 5.3: Tabla de municipios menos marginados.

En la gráfica 5.3 se presenta el conteo de municipios por nivel de grado de marginación por estado.



Figura 5.3: Conteo grados de marginación.

Observamos que Guerrero, Chiapas y Oaxaca destacan por tener gran proporción de municipios con grados de marginación “Alto” y “Muy Alto”, corroborando lo que habíamos notado anteriormente con base en el mapa 5.2.

5.3. Pruebas de autocorrelación espacial para índice de marginación

Habiendo observado los datos, se harán las pruebas de autocorrelación espacial de Moran y de Geary para el índice de marginación.

Los pesos w_{ij} a utilizar son los pesos binarios estandarizados por fila.

5.3.1. Índice \mathcal{I}

El valor del estimador \mathcal{I} , tiene el valor de

$$\hat{\mathcal{I}} = 0.703. \quad (5.1)$$

Sea \mathcal{I}_0 el valor esperado de \mathcal{I} bajo la hipótesis nula, como esperamos autocorrelación espacial positiva para el índice de marginación, usamos la hipótesis alternativa $H_0 : \mathcal{I} > \mathcal{I}_0$.

Utilizando los distintos supuestos, obtenemos los siguientes resultados.

Bajo el supuesto de normalidad

Se obtuvieron los siguientes resultados

$$\widehat{E}_N[\widehat{\mathcal{I}}] = -0.000407332,$$

$$\widehat{Var}_N(\widehat{\mathcal{I}}) = 0.000148633,$$

$$z = 57.6945,$$

$$\text{valor-p} < 2.2 \times 10^{-16}.$$

El valor del estadístico z está muy lejos de la región de no rechazo de H_0 , es claro que rechazamos la hipótesis nula de no autocorrelación, y por lo tanto, $\widehat{\mathcal{I}}$ es significativamente mayor que \mathcal{I}_0 .

Bajo el supuesto de aleatorización

Bajo el supuesto de aleatorización, se obtuvieron resultados bastante similares

$$\widehat{E}_R[\widehat{\mathcal{I}}] = -0.0004073320,$$

$$\widehat{Var}_N(\widehat{\mathcal{I}}) = 0.0001486393,$$

$$z = 57.6933,$$

$$\text{valor-p} < 2.2 \times 10^{-16}.$$

El valor del estadístico z bajo este supuesto es prácticamente el mismo que bajo el supuesto de normalidad, de la misma manera rechazamos la hipótesis nula de no autocorrelación espacial.

Simulaciones de Monte Carlo

Como el supuesto de normalidad es sensible es sensible a varios factores, es recomendable hacer simulaciones de Monte Carlo (vease 4.2.3) para muestrear de la distribución de \mathcal{I} .

Usando $n_{sim} = 9999$, se obtuvieron los siguientes resultados

orden observado = 10000,

valor-p = 1×10^{-4} .

Es decir, las 9999 observaciones del muestreo cayeron por debajo de $\hat{\mathcal{I}}$. Por lo tanto, esta prueba también nos indica claramente autocorrelación espacial positiva.

En la gráfica 5.4 podemos ver la densidad obtenida a partir de la muestra de Monte Carlo.

- Viendo la forma de la densidad, el supuesto de normalidad parece muy razonable.
- Se observa como $\hat{\mathcal{I}}$ cae muy lejos de la densidad de I bajo el supuesto de no autocorrelación espacial.

5.3.2. Índice \mathcal{C}

El valor del estimador $\hat{\mathcal{C}}$, tiene el valor de

$$\hat{\mathcal{C}} = 0.2943260693. \quad (5.2)$$

Sea, \mathcal{C}_0 el valor esperado de \mathcal{C} bajo la hipótesis nula, como esperamos autocorrelación espacial positiva para el índice de marginación, usamos la

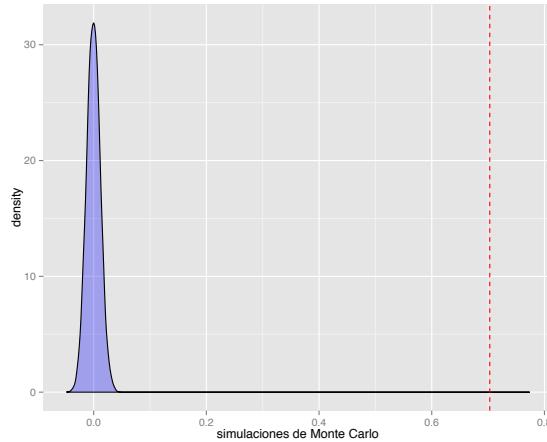


Figura 5.4: Densidad de la muestra de Monte Carlo de \mathcal{I} . La línea punteada indica donde se encuentra $\hat{\mathcal{I}}$.

hipótesis alternativa $\mathcal{C} < \mathcal{C}_0$, sin embargo, como construimos el estadístico z en la ecuación 4.35, esperamos que el estadístico z observado sea significativamente mayor a z bajo la hipótesis nula.

Así, utilizando los distintos supuestos, obtenemos los siguientes resultados.

Bajo supuesto de normalidad

Se obtuvieron los siguientes resultados

$$\text{E}_N[\hat{\mathcal{C}}] = 1,$$

$$\text{Var}_N(\hat{\mathcal{C}}) = 0.0001911584,$$

$$z = 51.0396,$$

$$\text{valor-p} < 2.2 \times 10^{-16}.$$

El valor del estadístico z está muy lejos de la región de no rechazo de H_0 , es claro que rechazamos la hipótesis nula de no autocorrelación, y por lo tanto, $\hat{\mathcal{C}}$ es significativamente menor que \mathcal{C}_0 .

Bajo supuesto de aleatorización

Bajo el supuesto de aleatorización, se obtuvieron resultados bastante similares

$$\widehat{E_N[\hat{\mathcal{C}}]} = 1,$$

$$\widehat{Var_N(\hat{\mathcal{C}})} = 0.0001889559,$$

$$z = 51.3362,$$

$$\text{valor-p} < 2.2 \times 10^{-16}.$$

El valor del estadístico z bajo este supuesto es prácticamente el mismo que bajo el supuesto de normalidad, de la misma manera rechazamos la hipótesis nula de no autocorrelación espacial.

Simulaciones de Monte Carlo

Al igual que hicimos con el coeficiente de Moran, ahora muestreamos de la distribución de \mathcal{C} .

Usando $n_{sim} = 9999$, se obtuvieron los siguientes resultados

$$\text{orden observado} = 1,$$

$$\text{valor-p} = 1 \times 10^{-4}.$$

Es decir, las 9999 observaciones del muestreo cayeron por encima de \hat{C} . Por lo tanto, esta prueba también nos indica claramente autocorrelación espacial positiva.

En la gráfica 5.5 podemos ver la densidad obtenida a partir de la muestra de Monte Carlo.

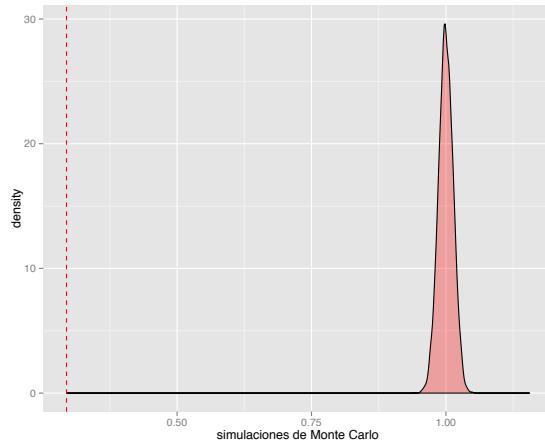


Figura 5.5: Densidad de la muestra de Monte Carlo de C . La línea punteada indica donde se encuentra \hat{C} .

- Viendo la forma de la densidad, el supuesto de normalidad parece muy razonable.
- Se observa como \hat{C} cae muy por debajo de la densidad de C bajo el supuesto de no autocorrelación espacial.

5.4. Análisis de conglomerados

Ahora, procedemos a agrupar los municipios por similitud. Para comparar los municipios, utilizaremos las variables que definen las cinco dimensiones de marginación propuesto por CONAPO (2004): analf, sprim, sdren, selec, sagua, hacina, pisot, l5kha y bingreso.

Se utilizará el algoritmo de k -medias esférico que utiliza como medida de disimilitud entre observaciones, la disimilitud de cosenos.

5.4.1. Determinación de K^* utilizando el estadístico Gap

Empezamos escogiendo el número de grupos K^* a través del estadístico Gap. Escogiendo el número máximo de valores K a probar $M = 10$, y $B = 100$ para el tamaño de la muestra de Monte Carlo, podemos ver los resultados en la tabla 5.4.

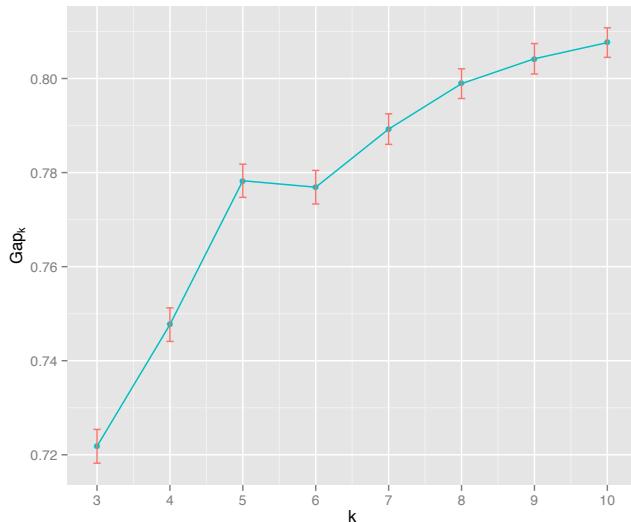
	$\log W_K$	$E^*\{\log W_K\}$	Gap_K	s_K
1	5.5489	6.1259	0.5770	0.00389
2	5.3049	5.9731	0.6682	0.00318
3	5.1981	5.9205	0.7225	0.00298
4	5.1272	5.8754	0.7481	0.00305
5	5.0644	5.8431	0.7787	0.00283
6	5.0364	5.8140	0.7776	0.00307
7	5.0024	5.7921	0.7897	0.00307
8	4.9711	5.7707	0.7995	0.00290
9	4.9501	5.7543	0.8042	0.00293
10	4.9317	5.7392	0.8075	0.00300

Cuadro 5.4: Valor de Gap para K , $K = 1, 2, \dots, 10$.

Observando la gráfica 5.6 y utilizando el criterio de Hastie et al. (2009)

(véase la ecuación 2.19) encontramos que $K^* = 5$.

Figura 5.6: Gráfica del estadístico Gap.



5.4.2. Resultado de K -medias esféricas

Habiendo obtenido que $K^* = 5$, corremos el algoritmo de k -medias esféricas sobre las variables mencionadas anteriormente.

Podemos ver la distribución de los 5 grupos en la tabla 5.5 y en la gráfica 5.7.

Vemos que en el grupo 3 es donde cayeron más observaciones (33 %), mientras que en el 4 cayó el menor número de observaciones (11 %). Los demás grupos tienen una distribución similar.

Ahora, en la tabla 5.6 vemos los centroides de cada grupo. Éstos nos dan una caracterización de cada uno de los grupos.

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

Cuadro 5.5: Tabla del tamaño de los 5 Grupos.

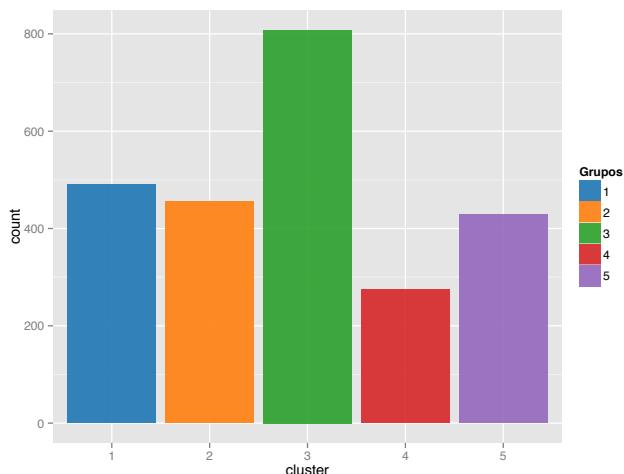


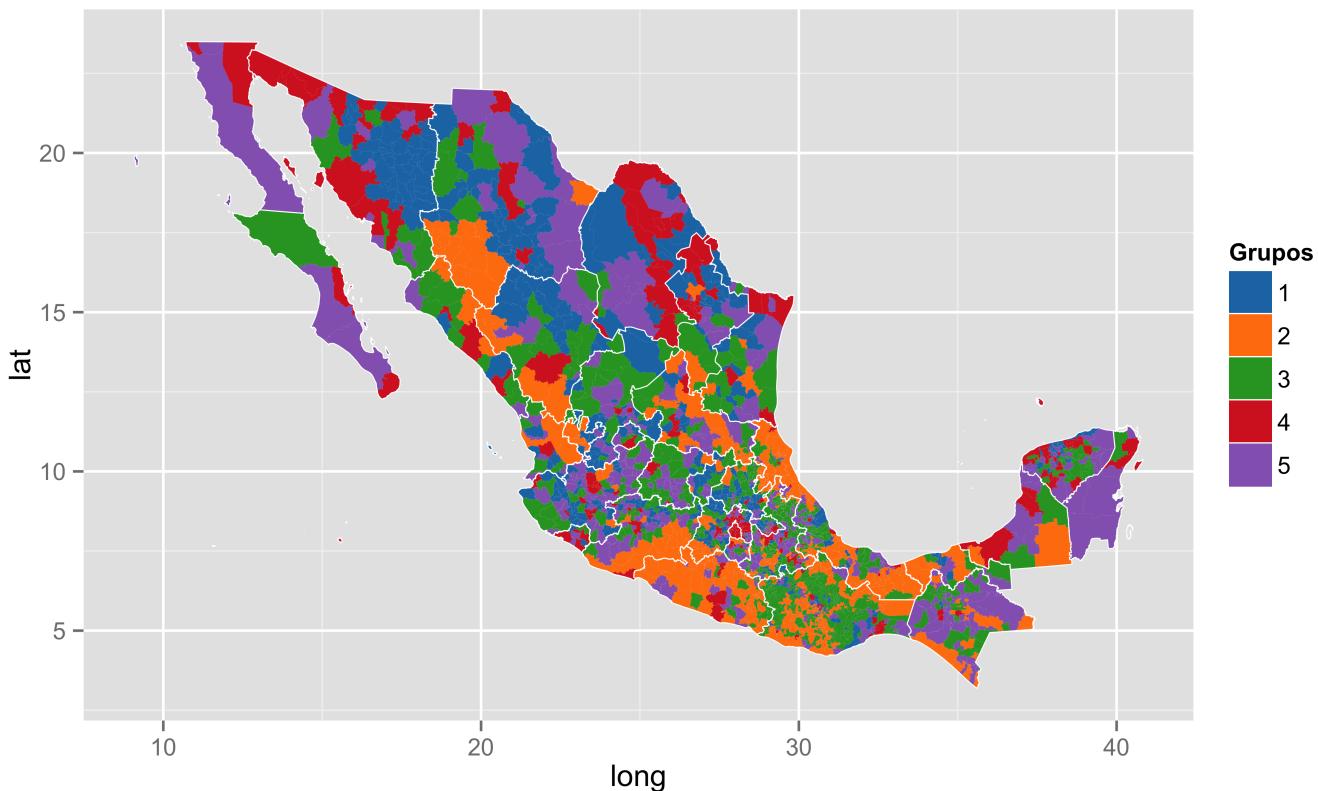
Figura 5.7: Distribución de los grupos.

En el mapa coroplético 5.8 se colorea cada municipio de acuerdo al grupo en el que cayó.

	analf	sprim	sdren	selec	sagua	hacina	pisot	pl5khab	bingreso
1	0.07	0.24	0.05	0.02	0.05	0.30	0.05	0.81	0.44
2	0.14	0.29	0.08	0.06	0.32	0.37	0.15	0.60	0.51
3	0.13	0.30	0.05	0.03	0.07	0.37	0.12	0.66	0.54
4	0.09	0.30	0.03	0.02	0.09	0.67	0.08	0.12	0.65
5	0.12	0.34	0.05	0.02	0.09	0.53	0.10	0.43	0.62

Cuadro 5.6: Centroides de los 5 conglomerados.

Figura 5.8: Mapa de México con los municipios coloreados por grupo.



En la gráfica 5.9 vemos la distribución por grupo de las 9 variables utilizadas en el algoritmo de conglomerados esféricos y en la gráfica 5.10 vemos la distribución de los grupos de marginación por grupo.

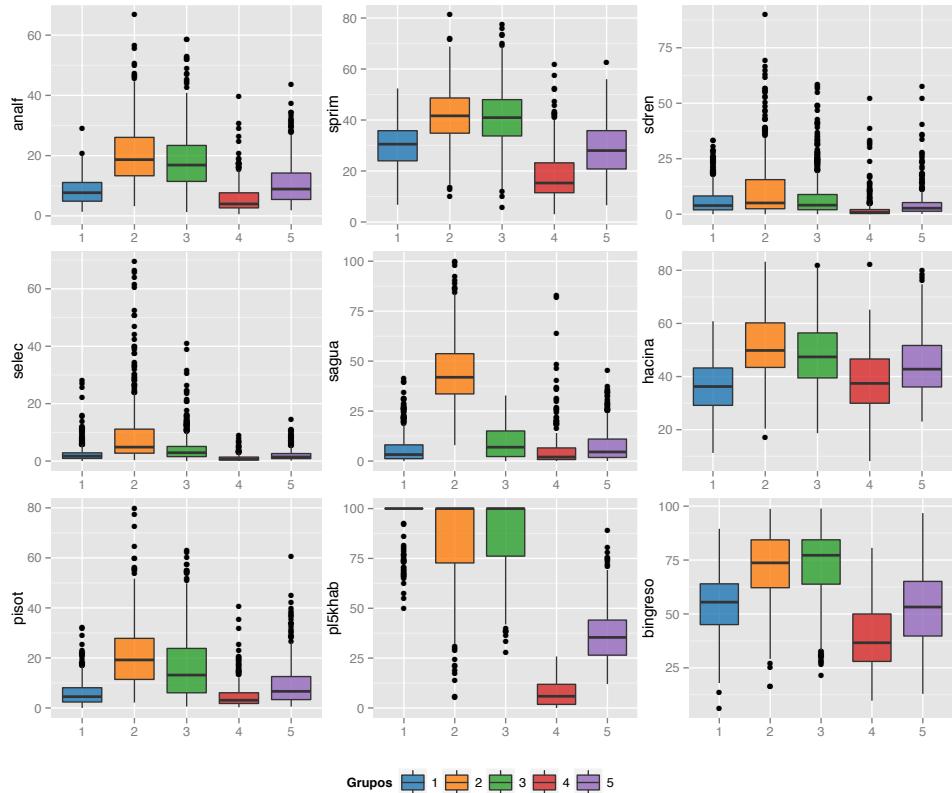


Figura 5.9: Distribución de variables de marginación por grupo.

A partir de la tabla 5.6, las gráficas 5.10 y 5.9 y observando el mapa 5.8 podemos recalcar lo más característico de cada uno de los cinco grupos.

- **Grupo 1:** Se caracteriza por tener municipios con un alto porcentaje de población en localidades con menos de 5,000 habitantes y

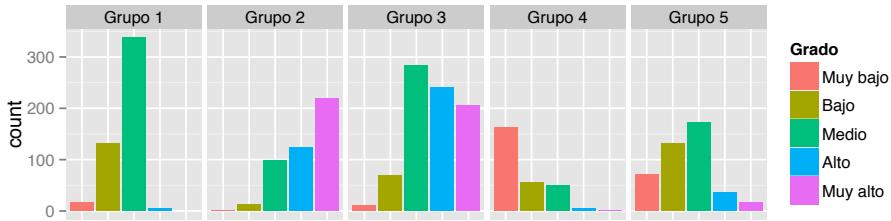


Figura 5.10: Grado de marginación por grupo.

poco marginados. De hecho podemos observar en la gráfica 5.10 que principalmente hay municipios con grado de marginación “Medio” y “Medio Bajo”. Los municipios más cercanos al centroide los podemos ver en la tabla 5.7.

	nom_mun	estado	imarges	gmarg	disim
1	Viesca	Coah	21.74	Medio	0.00153
2	San Gabriel	Jal	22.51	Medio	0.00154
3	Villa de la Paz	SLP	20.62	Medio	0.00160
4	Santa María del Oro	Nay	23.24	Medio	0.00176
5	Nazas	Dgo	22.21	Medio	0.00206

Cuadro 5.7: Municipios más representativos del grupo 1. La columna disim corresponde a la disimilitud de cosenos entre la observación i y el centroide del grupo 1.

- **Grupo 2:** En este grupo se encuentran los municipios más marginados mencionados en la tabla 5.2. El grado de marginación de dichos municipios es “Muy Alto”, “Alto” o “Medio” (ver gráfica 5.10). Se caracteriza por tener porcentajes altos en: población analfabeta y ocupantes en viviendas sin drenaje ni escusado, sin energía eléctrica, con

piso de tierra y sin agua entubada. De hecho, es el grupo que tiene mayor porcentaje de viviendas sin agua entubada, siendo la principal característica que lo diferencia del grupo 3. Podemos ver en la tabla 5.8 los municipios más representativos.

	nom.mun	estado	imarges	gmarg	disim
1	Xochiatipan	Hgo	45.17	Muy alto	0.00216
2	Coxquihui	Ver	43.21	Muy alto	0.00356
3	Playa Vicente	Ver	32.72	Alto	0.00357
4	Zozocolco de Hidalgo	Ver	47.71	Muy alto	0.00372
5	San Pedro Ixcatlán	Oax	42.32	Muy alto	0.00416

Cuadro 5.8: Municipios más representativos del grupo 2. La columna disim corresponde a la disimilitud de cosenos entre la observación i y el centroide del grupo 2.

- **Grupo 3:** Como en el grupo 1, los municipios del grupo 3 tienen alto porcentaje de localidades pequeñas (menos de 5,000 habitantes), pero los municipios tienen grado de marginación “Muy Alto”, “Alto” o “Medio”. Además los municipios en este grupo, tienen porcetajes altos de población ocupada con ingreso de hasta 2 salarios mínimos. En la tabla 5.9 podemos ver las observaciones más representativas. Es el grupo más grande con 808 municipios, representando un 33 % del total.
- **Grupo 4:** En este grupo podemos encontrar muchos municipios con baja marginación. Es el grupo más pequeño con tan solo el 11 % de las observaciones. Todas las delegaciones del Distrito Federal y los principales municipios de la Ciudad de Monterrey, en Nuevo León,

	nom_mun	estado	imarges	gmarg	disim
1	Amatenango de la Frontera	Chis	37.35	Alto	0.00102
2	Coetzala	Ver	35.01	Alto	0.00119
3	Mártires de Tacubaya	Oax	34.38	Alto	0.00160
4	Totutla	Ver	33.94	Alto	0.00187
5	Juan N. Méndez	Pue	36.20	Alto	0.00205

Cuadro 5.9: Municipios más representativos del grupo 3. La columna disim corresponde a la disimilitud de cosenos entre la observación i y el centroide del grupo 3.

están en este grupo. En este grupo, los municipios tienen un porcentaje relativamente bajo de viviendas con algún nivel de hacinamiento, bajo porcentaje de población ocupada con ingreso de hasta 2 salarios mínimos y bajo porcentaje de población de 15 años o más sin primaria completa. Los cinco municipios más representativos los podemos ver en la tabla 5.10.

	nom_mun	estado	imarges	gmarg	disim
1	Jiutepec	Mor	8.60	Muy bajo	0.00143
2	Cuernavaca	Mor	7.26	Muy bajo	0.00187
3	Coatzacoalcos	Ver	11.22	Muy bajo	0.00188
4	Matamoros	Tam	10.82	Muy bajo	0.00382
5	Río Bravo	Tam	13.51	Muy bajo	0.00392

Cuadro 5.10: Municipios más representativos del grupo 4. La columna disim corresponde a la disimilitud de cosenos entre la observación i y el centroide del grupo 4.

Podemos encontrar también algunos lugares turísticos como San Cristóbal de las Casas, Acapulco, Benito Juárez en Quintana Roo (aquí se en-

cuenta Cancún), Los Cabos, Puerto Vallarta, Veracruz, etc.

- **Grupo 5:** Este grupo es muy parecido al grupo 1, pero los municipios dentro del grupo 5 tienen mayor porcentaje de poblaciones con más de 5,000 habitantes, es decir, tienen poblaciones más grandes. Los cinco municipios más cerca del centroide 5 se encuentran en la tabla 5.11.

	nom_mun	estado	imarges	gmarg	disim
1	Villagrán	Gto	16.83	Bajo	0.00137
2	Cortazar	Gto	16.38	Bajo	0.00186
3	Angel Albino Corzo	Chis	35.21	Alto	0.00232
4	San Juan Bautista Tuxtepec	Oax	19.51	Bajo	0.00277
5	Celaya	Gto	11.76	Muy bajo	0.00294

Cuadro 5.11: Municipios más representativos del grupo 5. La columna disim corresponde a la disimilitud de cosenos entre la observación i y el centroide del grupo 5.

Otro punto que podemos destacar observando el mapa 5.8, es que existe cierta uniformidad espacial entre miembros del mismo grupo. Es decir, vemos que regiones que pertenecen al mismo grupo, comparten frontera.

En la próxima sección se harán las pruebas de hipótesis para detectar la presencia de autocorrelación espacial positiva.

5.5. Prueba de conteo de fronteras para conglomerados

Sean \hat{N}_{rs} los conteo observado, $N_{rs,0}$ los conteo esperado de fronteras del tipo rs , $r, s = 1, 2, 3, 4, 5$.

Dado que no contamos con las p_i 's (probabilidad de que una región sea de tipo i) a priori, las estimamos a partir de los datos. Por lo tanto, en el siguiente análisis de autocorrelación espacial se utilizará el supuesto de muestreo sin reemplazo. También se utilizarán los pesos binarios estandarizados por renglón.

La tabla 5.12 donde se muestran los conteos de fronteras observados conteos esperados y las varianzas bajo muestreo sin reemplazo y el valor z calculado.

Los conteos de fronteras entre municipios del mismo color son mucho mayores que lo esperado. Viendo el valor z entre las fronteras del tipo rr , podemos resaltar que, para $1 : 1$, es el más alto con 28.43; mientras que para $5 : 5$, es el más bajo. Esto indica que las regiones del grupo 1 están más aglomeradas que las regiones del grupo 5 (ver mapa 5.8).

Otro punto a mencionar es el caso del conteo N_{54} , es el único conteo de fronteras de tipo rs con $r \neq s$, cuyo valor z es positivo. Es decir, el número de fronteras entre regiones del grupo 4 y 5 es mayor a esperado. Podemos corroborarlo en el mapa 5.8, muchos municipios rojos tienen frontera con morados.

Como ejemplo de no autocorrelación espacial, se presenta el mapa 5.11, donde las etiquetas de los grupos están permutadas de manera aleatoria.

	\hat{N}_{rs}	$N_{rs,0}$	Var	z
1:1	98.61	48.80	6.53	19.49
2:2	110.35	42.07	5.77	28.43
3:3	194.99	132.80	14.00	16.62
4:4	50.36	15.23	2.36	22.88
5:5	63.59	37.40	5.22	11.46
2:1	35.95	90.81	12.56	-15.48
3:1	103.48	161.27	20.00	-12.92
3:2	138.83	149.75	18.74	-2.52
4:1	48.76	54.69	7.96	-2.10
4:2	18.10	50.78	7.48	-11.95
4:3	42.67	90.18	11.81	-13.82
5:1	82.05	85.63	11.93	-1.03
5:2	55.61	79.51	11.19	-7.14
5:3	122.55	141.19	17.79	-4.42
5:4	62.10	47.88	7.11	5.33
Jtot	710.10	951.70	37.24	-39.59

Cuadro 5.12: Conteo de fronteras.

Ahora, corroboramos lo dicho haciendo pruebas de hipótesis sobre N_{rr} , $r = 1, 2, 3, 4, 5$.

Bajo supuesto de normalidad

Suponiendo que $N_{11}, N_{22}, \dots, N_{55}$ se distribuyen normal, la tabla 5.13 muestra la misma información que la tabla 5.12 para conteo de bordes del mismo tipo, pero se añade el valor p porque ya estamos suponiendo una distribución.

Ahora, interpretando el valor z bajo el supuesto de normalidad, los conteos \hat{N}_{rs} 's caen muy lejos de la región de no rechazo. Por lo tanto, rechazamos la hipótesis nula de no autocorrelación espacial.

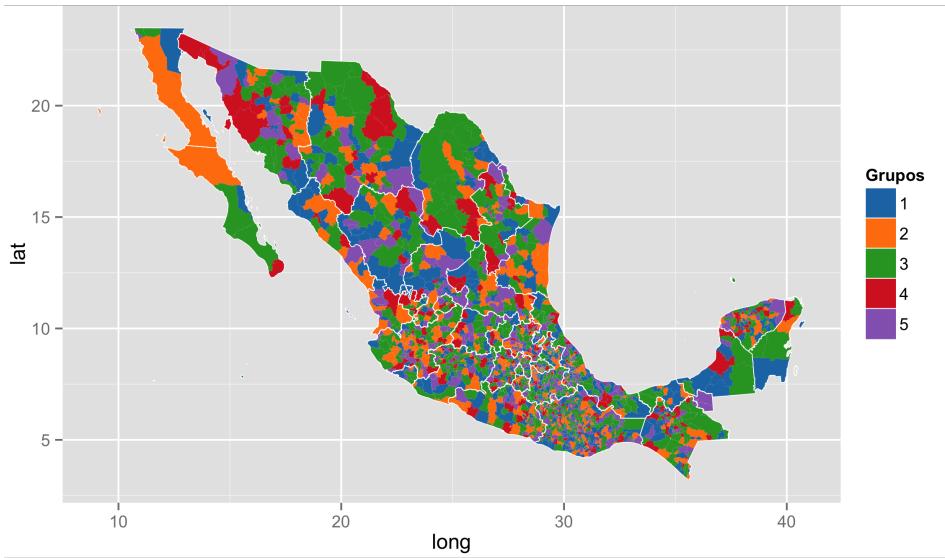


Figura 5.11: Permutación aleatoria de los grupos obtenidos mediante el algoritmo de k -medias esféricas.

Simulaciones de Monte Carlo

Ahora, realizamos la prueba omitiendo el supuesto de normalidad. Sean $N_{rs,0}^*$ y Var^* el valor esperado y la varianza esperada, ambos calculados a partir de la muestra de Monte Carlo. Usando una muestra de tamaño $n_{sim} = 9999$, podemos ver los resultados en la tabla 5.14.

El conteo \hat{N}_{rr} es mayor que las 9,999 simulaciones de Monte Carlo , para $r = 1, 2, 3, 4, 5$. Es decir, para todos los grupos, se rechaza la hipótesis

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	$< 2.2 \times 10^{-16}$

Cuadro 5.13: Pruebas de hipótesis para N_{rr} , $r = 1, 2, 3, 4, 5$.

grupo	\hat{N}_{rs}	$N_{rs,0}^*$	Var*	orden	valor-p
1	98.61	48.82	6.44	10000	1×10^{-4}
2	110.35	42.11	5.92	10000	1×10^{-4}
3	194.99	132.86	13.68	10000	1×10^{-4}
4	50.36	15.23	2.34	10000	1×10^{-4}
5	63.59	37.37	5.26	10000	1×10^{-4}

Cuadro 5.14: Resultados de simulaciones de Monte Carlo.

nula de no autocorrelación. Podemos ver la distribución de las muestras simuladas y donde caen los valores observados en la gráfica 5.12.

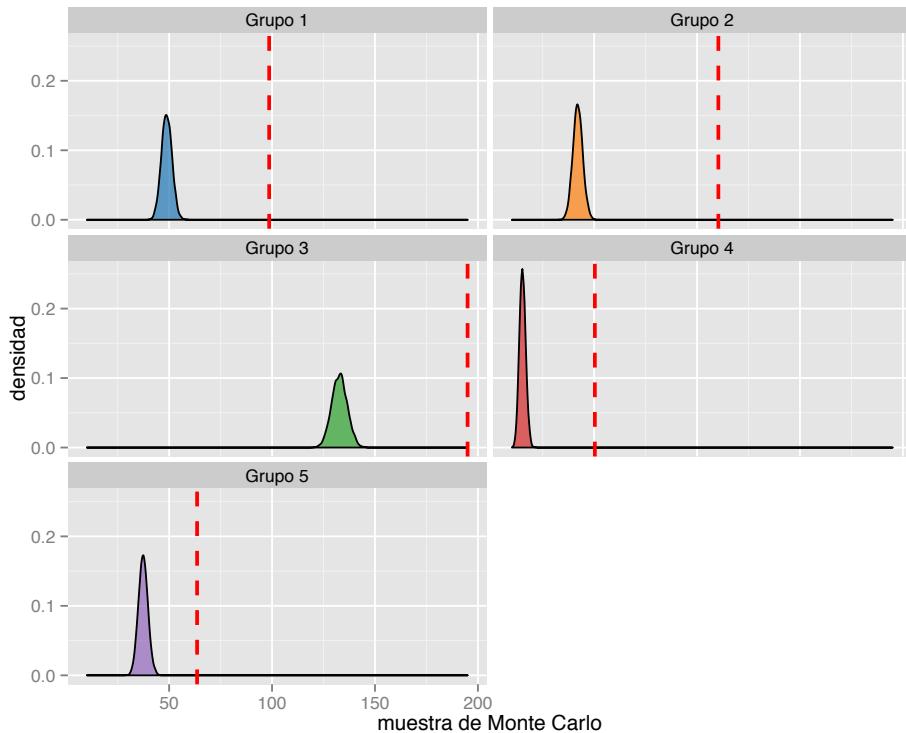


Figura 5.12: Densidad de la muestra de Monte Carlo de N_{rr} . La línea punteada indica donde se encuentra \hat{N}_{rr} .

Capítulo 6

Conclusiones

Después de haber realizado el análisis de conglomerados y las pruebas de autocorrelación espacial, se obtuvieron los siguientes resultados:

- Las pruebas sobre los índices \mathcal{I} y \mathcal{C} mostraron alta autocorrelación espacial positiva en el índice de marginación. Cómo se observó en el mapa 5.2 hay aglomeraciones muy marcadas. Las zonas más marginadas corresponden principalmente a municipios con población indígena y de difícil acceso.
- A partir del análisis de conglomerados esférico sobre las variables de marginación, se encontró una estructura espacial latente entre los municipios:
 - 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. En el primer

grupo cayeron municipios con localidades de pocos habitantes y con grado marginación de media a bajo; en el segundo, cayeron los municipios más marginados cuyo principal rasgo es la carencia de agua entubada; 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; en el cuarto, se agruparon los municipios con menor grado de marginación; y en el quinto, cayeron municipios con características similares al del primer grupo pero se diferencia en que cuenta con localidades más grandes.

- 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.

La importancia de este estudio está en que 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.

Apéndice A

Software y Reproducibilidad

Se puede encontrar una copia de este trabajo en el siguiente URL:
<https://github.com/carlosespino11/tesis>.

Ahí se encuentran todos los datos y archivos necesarios para generar todas las figuras, el documento escrito y el código para replicar todos los resultados de este trabajo.

Software utilizado:

- R
- L^AT_EX

Bibliografía

- Agresti, A. (2007), *An Introduction to Categorical Data Analysis*, John Wiley & Sons, 2 edición.
- Anselin, L. (1993), The moran scatterplot as an esda tool to assess local instability in spatial association, *GISDATA Specialist Meeting on GIS and Spatial Analysis*, pp. 645–678.
- Banerjee, A. y Ghosh, J. (2004), Frequency sensitive competitive learning for clustering on high-dimensional hyperspheres, *IEEE Transactions on Neural Networks*, 16(3); pp. 645–678.
- Cliff, A. D. y Ord, J. K. (1973), *Spatial Autocorrelation*, Pion Limited.
- Cliff, A. D. y Ord, J. K. (1981), *Spatial Processes: Models & Applications*, Pion Limited.
- Cliff, A. D., Martin, R. L., y Ord, J. (1975), A test for spatial autocorrelation in choropleth maps based upon a modified χ^2 statistic, *Transactions of the Institute of British Geographers*, (65); pp. 109–129.
- CONAPO, Índice de marginación 2005 (2004), URL <http://www.conapo.gob.mx>.
- CONAPO, Índice de marginación por entidad federativa y municipio 2010 (2011), URL <http://www.conapo.gob.mx>.
- Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Wiley, revised edición.
- Geary, R. (1954), The contiguity ratio and statistical mapping, *The Incorporated Statistician*, (5); pp. 115–145.

- Gordon, A. (2004), Null models in cluster validation, En *From Data To Knowledge*, pp. 32–44, Springer, New York.
- Haining, R. (2004), *Spatial Data Analysis: Theory and Practice*, Cambridge University Press, 1a edición.
- Hastie, T., Tibshirani, R., y Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer, 2da edición.
- Hoeffding, W. (1952), The large-sample power of tests based on permutations of observations, *Annals of Mathematical Statistics*, (23); pp. 169–192.
- Longley, P. A., Goodchild, M. F., Maguire, D. J., y W., R. D. (2005), *Geographical Information Systems: Principles, Techniques, Applications and Management*, Abridged, 2 edición.
- Maitra, R. y Ramler, I. P. (2010), A k -mean-directions algorithm for fast clustering of data on the sphere, *Journal of Computational and Graphical Statistics*, 19(2); pp. 377–396.
- Moran, P. (1948), The interpretation of statistical maps, *Royal Statistical Society*, 10(2); pp. 243–251.
- Moran, P. (1950), Notes on continuous stochastic phenomena, *Biometrika*, (37); pp. 17–23.
- Pearson, K. (1900), On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *Philosophical Magazine Series 5*, (50); pp. 157–175.
- Prematilake, C. C. (2011), Applications of spatial autocorrelation, Master's thesis, Texas Tech University.
- Tibshirani, R., Walther, G., y Hastie, T. (2001), Estimating the number of clusters in a data set via the gap statistic, *Journal of the Royal Statistical Society*, (63); pp. 411–423.
- Xu, R. y Wunsch, D. C. (2005), Survey of clustering algorithms, *IEEE Transactions on Neural Networks*, 5(3); pp. 3180–3185.

Xu, R. y Wunsch, D. C. I. (2008), *Clustering*, IEEE Press Series on Computational Intelligence, Wiley-IEEE Press, 2da edición.

Yan, M. (2005), *Methods of Determining the Number of Clusters in a Data Set and a New Clustering Criterion*, PhD thesis, Virginia Polytechnic Institute and State University.

Zhong, S. (2005), Efficient online spherical k-means clustering, *IEEE Transactions on Neural Networks*, 5(3); pp. 3180–3185.