UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Diciembre 2019

Modelo Logístico Multinomial para datos de Áreas

Multinomial Logistic Regression to areal data

Luisa Fernanda Rodríguez Ortiz.[a]
luisarodriguezo@usantotomas.edu.co

Wilmer Pineda Ríos.[b]
wilmerpineda@usantotomas.edu.co

1. Introducción

De acuerdo a la constitución propuesta en 1991, el país se ha propuesto un proceso de descentralización, cuyo objetivo, es mejorar las condiciones y calidad de vida de la población. La idea original es que a medida que se van midiendo las competencias, se irán distribuyendo equitativamente los recursos dentro de los entes y entidades territoriales, de esta manera, se podrán identificar las necesidades de la población.

Para lograr lo anteriormente descrito, el principal objetivo es identificar las falencias de los entes territoriales para alcanzar un mayor nivel de bienestar. Es necesario entonces, determinar que tan robusto es el gobierno presente, así como determinar la mejor manera de ejecutar proyectos futuros que ayuden al incremento de la calidad de vida.

Medir las capacidades de los gobiernos es el principal logro, para eso, el Departamento Nacional de Planeación (DNP), creo un instrumento y/o herramienta, capaz de determinar el desempeño integral municipal y departamental para Colombia, se denomina "Índice de Desempeño Integral"(IDI), cuya funcionalidad es orientar las diversas políticas públicas para un mejor resultado de la administración de los recursos estatales. Principalmente, el índice evalúa la capacidad de las entidades territoriales, en cuanto a la eficacia y eficiencia en el desarrollo de sus planes de gobierno, así como en el cumplimiento de sus propuestas iniciales.

La primera metodología de medición permitió medir el desempeño para descubrir los retos y logros de los pequeños gobiernos; focalizar las actividades del gobierno central según su capacidad; y por último, definir una primera herramienta para llevar a cabo la descentralización territorial.

Este artículo tiene como objetivo general, identificar por medio de la regresión logística multinomial espacial la probabilidad de ocurrencia de un suceso en particular. Para este caso en particular, identificar la probabilidad de pertenecer a un nivel bajo, alto, sobresaliente o satisfactorio de desempeño integral.

Para el desarrollo del objetivo anterior, el presente informe estará dividido en 5 principales secciones, en primera instancia se presentará la introducción, en donde se mostrará de manera general la hipótesis, los objetivos y el desarrollo del trabajo, a continuación, se encontrará la explicación teórica a cerca de modelos mixtos y modelos lineales geenralizados orientado hacia una visión espacial, la tercera división englobará los métodos y herramientas de simulación que lograrán determinar los resultados y conclusiones. Se utilizarán datos de los 32 departamentos de Colombia, con 34 variables enfocadas hacia el pilar de educación.

[a]Economista y estudiante pregrado en Estadística, U. Santo Tomás, sede Bogotá
[b]Docente Facultad de Estadística, U. Santo Tomás, sede Bogotá

1.1. Metodología para la medición del IDI

El nuevo esquema de medición esta basado en 3 pilares específicos:

- Reducir las desigualdades de la población.
- Reducir la dispesión entre la gestión y los resultados de la capacidad política territorial de los pequeños gobiernos.
- Poder explicar las diferencias presentes entre la gestión política de las diversas entidades territoriales.

Para el análisis de dichos principios, se cuenta con la medición de 2 componentes principales:

- Componente de gestión: Basado en la manipulación, distribución y eficiencia de las entes territoriales.

1 Recursos:

$$U_{indicador(0-1)} = \frac{Ui - MIN_U}{MAX_U - MIN_U}$$

2 Económico:

$$DE_{indicador(0-1)} = \frac{DEi - MIN}{MAX - MIN}$$

2 Dimensión urbana:

$$PC_{indicador(0-1)} = \frac{PCi - MIN_{ajust}}{MAX_{ajust} - MIN_{ajust}}$$

- Componente de resultados: Basado en los resultados obtenidos en pruebas externas a la medición del IDI

1 Educación

2 Salud

3 Acceso a servicios

4 Seguridad

2. Modelos Lineales Generalizados

Los modelos lineales generalizados son considerados como una generalización de los modelos lineales simples. Este tipo de modelos permite el uso de variables respuesta diferentes a las correspondientes a la distribución normal, siempre y cuando pertenezcan a la familia exponencial, es decir, incumplan con los supuestos de normalidad en los errores y varianza constante.

Los GLM (por sus siglas en inglés), están compuestos por tres pilares principales (Marín, 2011):

- Componente aleatoria: "Identifica la variable respuesta y su distribución de probabilidad."
- Componente sistemática: "Especifica las variables explicativas (independientes o predictoras) utilizadas en la función predictora lineal."
- Función Link o enlace: "Relaciona el valor esperado de Y con el predictor lineal, la cual debe de cumplir características especiales como ser monótona y doblemente diferenciable"

2.1. Definición:

Sean $Y_1, \cdots, Y_n$ variables aleatorias independientes, cada una con una distribución perteneciente a la familia exponencial y con las siguientes propiedades:

- La distribución de cada variable Y_i, tiene forma canónica y depende de un solo parámetro θ_i. La distribución queda expresada de la siguiente manera:

$$f(y_i, \theta_i) = exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)] \tag{1}$$

- La distribución de cada variable Y_i es la misma para todas, por lo tanto el subíndice de b, c y d no son necesarios. De acuerdo a lo anterior, la función de densidad de probabilidad de $Y_i, \cdots, Y_N$, queda expuesta así:

$$f(Y_i, \cdots, Y_N; \theta_i, \cdots, \theta_N) = \prod_{i=1}^{N} exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)] = exp[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)] \tag{2}$$

Según la definición de los GLM, se determinan 3 componentes principales, la transformación o pertenencia de las distribuciones de probabilidad simples a la familia exponencial (equación 1) en donde se evidencia la componente aleatoria (Y_i), y la componente sistemática, la cual se describe así:

$$g(\mu_i) = \eta_i \tag{3}$$

donde η_i se conoce como el predictor lineal, también escrito $X_i^T \beta$, con $\beta = (\beta_i, \cdots, \beta_p)$ (p < N) como los parámetros a estimar, x_i como las variables independientes, y $g(\cdot)$ como la función de enlace (Paula, 2004).

2.2. Modelo Logístico Multinomial

Como se decía en el item anterior, los GLM son usados principalmente con distribuciones pertenecientes a la familia exponencial, como por ejemplo la distribución poisson, la distribución gamma, entre otras. Para este caso de estudio en particular, se hará uso del modelo logístico multinomial, el cual es una generalización del modelo logístico simple, en donde la variable respuesta (Y_i), está determinada por dos categorías de respuesta; para este caso, dentro de la base de datos de la que se hará uso, la variable de respuesta (níveles de calidad de agua) tendrá más de dos categorías de análisis.

Para los modelos de respuesta binaria, se tiene una variable Y que puede tener 2 respuesta, $Y = 1$(donde la variable posee alguna característica específica), o $Y = 0$ (donde la variable carece de esa característica). La ecuación orginal es:

$$P[Y = 1|X] = \frac{exp(b_0 + \sum_{s=1}^{n} b_s x_s)}{1 + exp(b_0 + \sum_{s=1}^{n} b_s x_s)} \tag{4}$$

donde $P[Y = 1|X]$ es a probabilidad de que el individuo posea la caracteística del estudio, es decir, la probabilidad de que $Y = 1$. De ahora en adelante se denotará como π_i.

Esta ecuación ya viene dada de tipo exponencial, lo que la hace útil para el presente documento, pero para mayor facilidad, se le calculará el logaritmo, lo que resulta en:

$$ln\left[\frac{\pi_i}{1-\pi_i}\right] = b_0 + \sum_{s=1}^{n} b_s x_s \tag{5}$$

Para el caso de la función logística multinomial, se modela mediante el uso de la suma de la función anterior (equación 5), dependiendo de cuantas categorías presenta la variable dependiente.

2.2.1. Definición

Se considera una variable de respuesta politómica Y, con más de 2 niveles de respuesta, denotados $Y_i = Y_1, \cdots, Y_k$, con una categoría como referencia. Se pretende explicar la probabilidad de cada nivel en función de una variables explicativas X_i. Para su desarrollo, como se denotaba en el item anterior, es necesaria la suma de las probabilidades $(\pi_{i1}, \pi_{i2}, \cdots, \pi_{ik})$ de cada una de las categorías respuesta para cada observación, por lo cual:

$$ln\left[\frac{\pi_{ij}}{\pi_{ik}}\right] = X_i^T \beta_j \tag{6}$$

donde i denota los individuos, k denota la categoría de referencia, y j denota las demás categoías de la variable respuesta. La función anterior, es la función de enlace del modelo logístico multinomial. Aplicando exponencial a la equación 6, se obtiene:

$$\frac{\pi_{ij}}{\pi_{ik}} = exp^{X_i^T \beta_j} \tag{7}$$

pasando a multiplicar π_{ik}, se obtiene:

$$\pi_{ij} = \pi_{ik} exp^{X_i^T \beta_j} \tag{8}$$

Teniendo en cuenta que la suma de probabilidades debe de ser igual a 1, y teniendo en consideración que el objetivo de este modelo es encontrar la probabilidad de ocurrencia de cada categoría, se obtiene la siguiente fórmula:

$$\begin{aligned} \sum_{j=1}^{k} \pi_{ij} = 1 \rightarrow \sum_{j=1}^{k-1} \pi_{ij} + \pi_{ik} &= 1 \\ \rightarrow \sum_{j=1}^{k-1} \pi_{ik} exp^{X_i^T \beta_j} + \pi_{ik} &= 1 \\ \rightarrow \pi_{ik} \left(1 + \sum_{j=1}^{k-1} exp^{X_i^T \beta_j}\right) &= 1 \\ \rightarrow \pi_{ik} = \frac{1}{1 + \sum_{j=1}^{k-1} exp^{X_i^T \beta_j}} & \end{aligned} \tag{9}$$

Finalmente, despejando la ecuación 9, se obtiene la fórmula con la que se desarrollará el resto del trabajo.

$$\pi_{ik} = \frac{exp^{X_i^T \beta_j}}{1 + \sum_{j=1}^{k-1} exp^{X_i^T \beta_j}} \tag{10}$$

3. Modelos Mixtos (MLM)

Los modelos lineales mixtos son presentados como una generalidad al procedimiento efectuado por los modelos lineales simples pero con una peculiaridad, estos modelos son esencialmente utilizados para trabajar con datos que tienen una variabilidad o aleatoriedad que debe ser captada en las estimaciones, por lo tanto, el modelo estima la media, la varianza y la covarianza de la información; para eso, el procedimiento que se efectúa es realizado en dos etapas (Uribe, 2016).

Etapa 1: Ajustar un modelo de regresión para cada sujeto por separado.

Sea Y_{ij} la variable respuesta para el i-ésimo sujeto medida en el tiempo X_{ij}, con i = 1,2,...,m y j = 1,2,...,n_i, se tiene que $\mathbf{Y_i} = (Y_{i1}, Y_{i2}, \cdots, Y_{in_i})^T$ es el vector de respuestas para el sujeto i.

Así las cosas, se determina que el modelo lineal que estima la variabilidad intra-sujeto es $\mathbf{Y_i} = \mathbf{Z_i}\beta_\mathbf{i} + \epsilon_\mathbf{i}$, donde $\mathbf{Z_i}$ es la matriz de covariables conocidas de tamaño n_i x p, $\beta_\mathbf{i}$ es un vector donde se visualizan los coeficientes de la regresión, y $\epsilon_\mathbf{i}$ son los errores de la regresión, distribuidos $N(0, \Sigma_i)$, donde Σ_i es una matriz de varianzas y covarianzas.

Etapa 2: Explicar la variabilidad en los coeficientes de regresión de cada modelo estimado en la etapa anterior, usando variables conocidas (efectos fijos) o desconocidas (efectos aleatorios).

Se busca estimar la variabilidad entre los sujetos, mediante el siguiente modelo: $\beta_\mathbf{i} = \mathbf{K_i}\beta + \mathbf{b_i}$, donde $\mathbf{K_i}$ es una matriz de covariables conocidas (de tamaño q x p), β es un vector de parámetros de regresión desconocida, y $\mathbf{b_i} \sim N(0, D)$ donde D es una matriz de varianzas y covarianzas (q x q).

3.1. Forma matricial del MLM

Sean

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_m \end{pmatrix}, X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{pmatrix}, e = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_m \end{pmatrix}, Z = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_m \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

Así, la forma matricial queda definida como:

$$Y = X\beta + Zb + e \tag{11}$$

donde $E(e) = 0$ y

$$\sum_e = \begin{pmatrix} e_1 & 0 & \cdots & 0 \\ 0 & e_2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & e_m \end{pmatrix}, \sum_b = \begin{pmatrix} D & 0 & \cdots & 0 \\ 0 & D & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & D \end{pmatrix} = I_m \otimes D$$

4. Modelos Lineales Generalizados Mixtos (MLMG)

Generalmente, los modelos mixtos son usados para el modelamiento de variables en su mayoría contínuas. Los modelos lineales generalizados mixtos, buscan modelar el comportamiento de respuestas no necesariamente contínuas.

Suponga que $\mathbf{Y}$ es el vector de observaciones, se asume que $\mathbf{Y}$ al estar condicionado a $\mathbf{b}$, toma una distribución de la familia exponencial, por lo tanto:

$$f_{y_i|b}(y_i|\mathbf{b}, \beta, \phi) = e\left\{\frac{y_i\eta_i + c(\eta_i)}{a(\phi)} + d(y_i, \phi)\right\} \tag{12}$$

donde $\mathbf{b} \sim f_b(\mathbf{b}|\mathbf{D})$. El Modelo Lineal Gneralizado Mixto (MLMG) se identifica por:

$$\eta_i = \mathbf{X}_i^{'}\beta + \mathbf{z}_i^{'}\mathbf{b} \tag{13}$$

con $\mathbf{X}_i^{'}$ como la i-ésima fila de $\mathbf{X}$, y $\mathbf{z}_i^{'}$ la i-ésima fila de $\mathbf{z}$.

4.1. Estimación

Se determina la función de verosimiitud para la ecuación 1 como:

$$L(\beta, \phi, \mathbf{D}|\mathbf{Y}) = \int \prod_{i=1}^{n} f_{y_i|b}\ (y_i|\mathbf{b}, \beta, \phi)\ f_b\,(\mathbf{b}|\mathbf{D})\ \ d\mathbf{b} \tag{14}$$

4.2. Modelo logístico multinomial mixto

Como se decía en items anteriores, dentro de la notación de un modelo logístico multinomial, el subíndice i denota los individuos o clusters dentro de la base de datos, mientras que el subíndice j denota las observaciones anidadas o las categorías de respuesta. Adiconal a lo anterior, se debe de tener en cuenta una categoría base, de donde se pueda partir para la realización del modelo. La ecuación inicial es:

$$\pi_{ij} = P(y_{ij} = c|\beta) = \frac{exp(z_{ij})}{1 + \sum_{h=1}^{C} exp(z_{ij})} \quad para \quad cada \quad c = 2, 3, \cdots, C \tag{15}$$

$$\pi_{ij} = P(y_{ij} = 1|\beta) = \frac{1}{1 + \sum_{h=1}^{C} exp(z_{ij})} \tag{16}$$

donde, la ecuación 16 determina la probabilidad de ocurrencia de la categoría de respuesta base, mientras la fórmula 15, la probabilidad de ocurrencia de las demás categorías.

Además de lo anterior, se tiene que $z_{ij} = W_{ij}^{'}\alpha_c + X_{ij}^{'}\beta_{ic}$ es el parámetro mixto del modelo, para este caso práctico, determina el parámetro espacial del modelo logístico multinomial. Aquí $W_{ij}^{'}$ es el vector de covariables de tamaño $s\ x\ 1$ y $X_{ij}^{'}$ es el vector diseño para los r efectos aleatorios. Cada uno de los vectores anteriores deben de estar determinado para todas las categorías respuesta, de acuerdo a todas los individuos presentes en la base de datos.

De igual manera, se define que α_c es un vector desconocido de parámetros de regresión fijos, de tamaño $s\ x\ 1$, y β_{ic} es un vector desconocido de efectos aleatorios de tamaño $r\ x\ 1$. La distribución de los efectos aleatorios es asumido distribuido de forma normal multivariado con un vector de media igual a 0 y una matriz de covarianza Σ_c

Para mayor facilidad, es necesario estandarizar los efectos aleatorios del modelo anterior. Para eso, $\beta_{ic} = T_c\theta_c$, donde $T_cT_c^{'} =_c$ es la descomposición de Cholesky de Σ_c. La reparametrización del modelo quedaría:

$$z_{ij} = W_{ij}^{'}\alpha_c + X_{ij}^{'}T_c\theta_c \tag{17}$$

El modelo escrito anteriormente, permite la estimación de cualquier par de comparaciones entre las categorías de respuesta. Ahora, es más beneficioso escribir el modelo nominal para cualquier $C - 1$ categorías. De acuerdo a lo anterior, la función de probabilidad de cada categoría queda escrito:

$$\pi_{ij} = \frac{exp(z_{ij})}{\sum_{j=1}^{C} exp(z_{ij})} \quad para \quad cada \quad c = 1, 2, 3, \cdots, C \tag{18}$$

donde ahora,

$$z_{ij} = W_{ij}^{'}\Gamma d_c + (X_{ij}^{'}T_c\theta_c) \tag{19}$$

Para nuestro caso práctico, en donde se determina el término mixto desde un ámbito espacial, quedaría:

$$z_{ij} = X_i^t \beta_i + V_{ij} \tag{20}$$

4.2.1. Estimación

Como se determinó en el item anterior, la estimación para este tipo de modelos se realiza por medio de verosimilitud, para eso, se debe de calcular la integral de la función de probabilidad para las categorías $C-1$, quedando de la siguiente manera:

$$Y_{ij} = (Y_{i1}, Y_{i2}, \cdots, Y_{ic})$$

$$\begin{aligned} l(y_{ij}|\theta) &= \prod_{h=1}^{C} \pi_{ij}^{y_{ij}} \\ &= \prod_{j=1}^{C} \left(\frac{exp(z_{ij})}{1 + \sum_{l=1}^{C-1} exp(z_{il})} \right)^{y_{ij}} \end{aligned} \tag{21}$$

$$l(y_{ij}) = \int_{\mathbb{R}^d} \left[\prod_{j=1}^{C} \left(\frac{exp(z_{ij})}{1 + \sum_{l=1}^{C-1} exp(z_{il})} \right)^{y_{ij}} \right] [N(0, \Sigma_i)] \; d\gamma \tag{22}$$

al sacar el logaritmo, queda:

$$L(l(\beta_j)) = \sum_{j=1}^{C} (y_{ij}(z_{ij})) - y_{ij} log(1 + \sum_{j=1}^{C-1} exp(z_{il})) + N(0, \Sigma_j) \tag{23}$$

Dado que la función anterior no puede desarrollarse de forma cerrada, se dispone a usar métodos numéricos para encontrar la solución más aproximada.

5. Métodos de Simulación

5.1. Fisher Scoring

Consiste en sustituir el procedimiento de segundas derivadas (evaluadas en β^T) de la función de máxima verosimilitud, por su valor esperado.

$$E\left[\frac{d^2 l}{d\beta_j d\beta_k} \right] = \sum_{i=1}^{n} = \frac{x_{ij} x_{ik}}{V[Y_i] g'(\mu_i)^2} \tag{24}$$

Equivale a resolver o a encontrar convergencia entre las funciones, de manera iterativa. La sucesión β^T debe de converger al máximo estimador de verosimilitud de β

5.2. Cuadratura Gauss - Hermite

Una de las mejores metodologías para el desarrollo de integrales es el procedimiento de cuadratura de Gauss, donde, su cálculo está basado en la regla del trapecio, lo que le permite tomar dos puntos estratégicos de la curva de integración, y por lo tanto, reducir el error presente en las estimaciones, logrando identificar el área bajo la curva. De dicho cálculo, se derivan varias ecuaciones de cuadraturas, una de ellas es la cuadratura de Gauss - Hermite, la cual permite estimar modelos con efectos mixtos, su procedimiento teórico se describe a continuación:

$$\int_{-\infty}^{\infty} f(x)e^{-x^2}\ dx \approx \sum_{i=1}^{n} w_i f(x_i) \tag{25}$$

cuyo objetivo es calcular el valor de las abscisas x_i y de los pesos w_i de la integral original. Las abscisas se entienden como los ceros presentes en los polinomios de Hermite, los cuales son:

$$H_n(x) = n! \sum_{i=0}^{[n/2]} \frac{(-)^i}{i!(n-2)!}(2x)^{n-2i} \tag{26}$$

y los pesos estan descritos como:

$$w_i = \frac{2^{n-1}n!\sqrt{\pi}}{[nH_{n-1}(x_i)]^2} = \frac{2^{n+1}n!\sqrt{\pi}}{[H_n'(x_i)]^2} \tag{27}$$

La iteración de Newton de primer orden para estabilizar las raíces de x_i es:

$$x_i \longrightarrow x_i - \frac{f(x_i)}{f'(x_i)}, \tag{28}$$

y el apoyo que termina la fracción contínua es:

$$\frac{H_n(x)}{H_n'(x)} = \frac{1}{2n}\frac{H_n(x)}{H_{n-1}(x)} = \frac{1}{2n}\left[2x - \frac{2(n-1)}{2x-}\frac{2(n-2)}{2x-}\cdots\frac{2}{2x}\right] \tag{29}$$

6. Correlación espacial

Para el análisis de la correlación espacial entre los departamentos de Colombia, se debe de tener en cuenta la herramienta *Joint Count Test*, la cual es comúnmente usada para manejar variables respuesta con más de 2 categorías. Al igual que el test de Moran, es usado para probar la hipótesis nula de la no existencia de autocorrelación espacial en los datos expuestos.

De acuerdo con su cálculo, se deben de tener en cuenta dos metodologías:

$$J_1 = \frac{1}{2}\sum_i\sum_j w_{ij}Y_iY_j$$

$$J_2 = \frac{1}{2}\sum_i\sum_j w_{ij}(Y_i - Y_j)^2$$

Para lograr el siguiente estadístico:

$$t = \frac{J_i - E(J_i)}{\sqrt{var(J_i)}/n}$$

7. Análisis de resultados

En la sección anterior se establecieron bases teóricas para la estimación del modelo y análisis del mismo, en el presente capítulo, se mostrarán los resultados obtenidos de las estimaciones y modelaciones realizadas con base en la información extraída de el tratamiento de aguas.

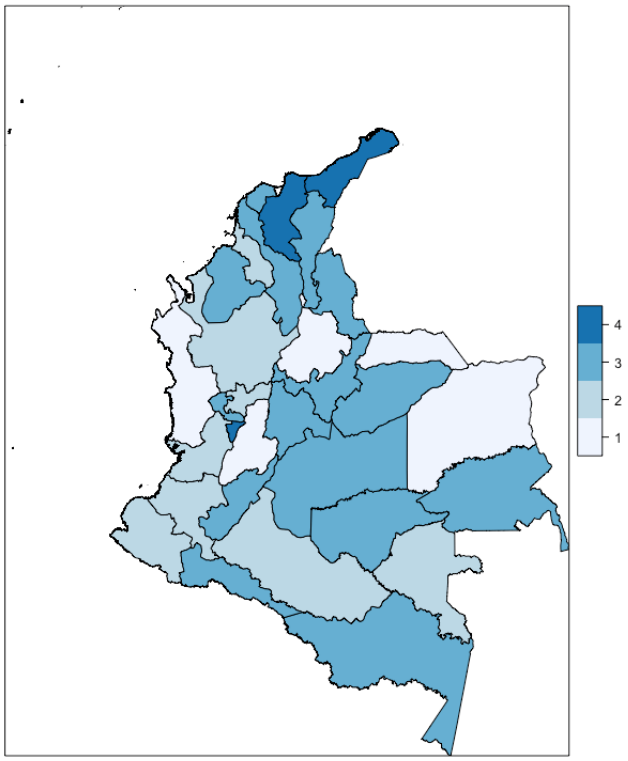

Figura 1: Departamentos de Colombia

Fuente: Elaboración propia. Datos Departamento Nacional de Planeación, 2017

En donde las categorías son: 4 - Sobresaliente, 3 - Satisfactorio, 2 - Alto, 1 - Bajo. De acuerdo a lo anterior, se puede observar que departamentos como La guajira y el magdalena se encuentran en un nivel sobresaliente de desempeño, departamentos como el amazonas, guanía, y varios ubicados en el centro del país, se ubican dentro de un nivel satisfactorio de desempeño; así mismo, departamentos como Nariño, Vaupés, y algunos ubicados al oriente del país, se encuentran en un nivel alto de desempeño; y por último, departamentos como vichada, chocó, entre otros, se encuentran dentro de un nivel bajo de desempeño.

A continuación se presenta un análisis previo y general de la información:

7.1. Análisis descriptivo

Los datos uilizados para el presente artículo, fueron extraídos del Departamento Nacional de Planeación (DNP), todo con propósitos acádemicos. Esta información puede ser utilizada por el público en general, por lo que se determina como información abierta.

Para la realización de los modelos, se tomó en cuenta la información presente en los departamentos

de Colombia. En total son 32 observaciones con 34 variables, de las cuales solo resultaron significativas 8, descritas a continuación:

- IDI: Índice de desempeño integral
- REPRO: Cantidad de reprobados reprobados
- DES_TRAN: Cantidad de deserción en transición
- DES_PRI: Cantidad de deserción en primaria
- DES_SEC:Cantidad de deserción en secundaria
- DES_MED: Cantidad de deserción en media
- CN_TRAN: Cantidad neta de estudiantes en Transición
- CN_PRI: Cantidad neta de estudiantes en Primaria

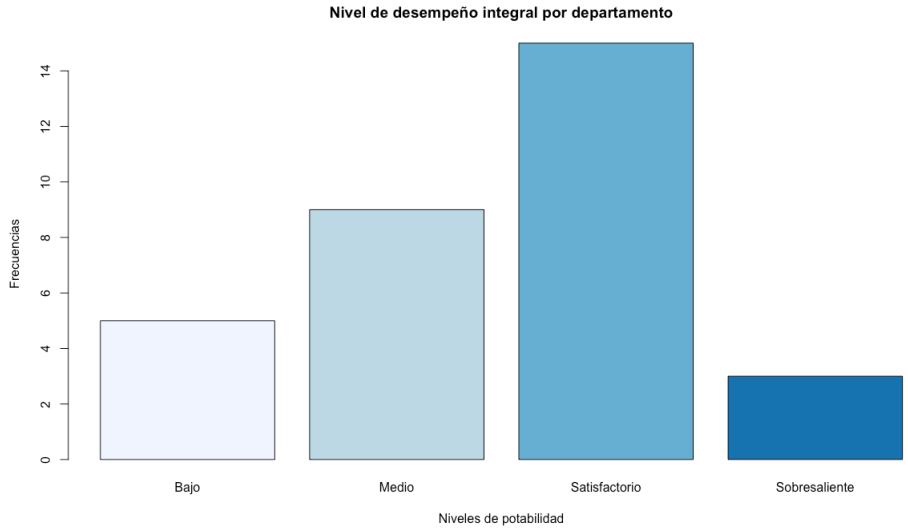

Figura 2: Frecuencias de las categorías de Nivel de desempeño integral (IDI)

Fuente: Elaboración propia, datos tomados del DNP, 2017

Según el gráfico anterior, se puede determinar que la mayor parte de los departamentos se encuentran en un nivel satisfactorio de desempeño integral, basado en un análisis de variables orientadas a la educación.

7.2. Análisis espacial

En primera instancia, se debe de determinar una matrix de vecindades o de pesos espaciales, que permite considerar las relaciones de forma espacial entre las unidades o individuos presentes en el análisis, de forma matemática. Para esto, se determina la distancia entre los centroides de cada territorio.

Como se presentó en capítulos anteriores, el proceso de autocorrelación espacial es primordial para comprender las dependencias espaciales entre las unidades o individuos que se quieren analizar, es decir, es aquella medida que determina que tan similar es un objeto con respecto a otro cercano, presente en

la base de información.

Para el caso práctico propuesto en este artículo, se hizo uso del código *joincount.test*, el cual arrojó los siguientes p-valores (cabe anotar que el código calculó un p-valor por categoría):

```
        Join count test under nonfree sampling

data:  x
weights: nb2listw(Vecindades, style = "B")

Std. deviate for Bajo = 0.11247, p-value = 0.4552
alternative hypothesis: greater
sample estimates:
Same colour statistic           Expectation              Variance
            1.0000000             0.9032258             0.7403249

        Join count test under nonfree sampling

data:  x
weights: nb2listw(Vecindades, style = "B")

Std. deviate for Medio = -0.71534, p-value = 0.7628
alternative hypothesis: greater
sample estimates:
Same colour statistic           Expectation              Variance
             4.000000              5.419355              3.936934

        Join count test under nonfree sampling

data:  x
weights: nb2listw(Vecindades, style = "B")

Std. deviate for Satisfactorio = 3.3681, p-value = 0.0003784
alternative hypothesis: greater
sample estimates:
Same colour statistic           Expectation              Variance
            26.000000             15.806452              9.159424

        Join count test under nonfree sampling

data:  x
weights: nb2listw(Vecindades, style = "B")

Std. deviate for Sobresaliente = 0.89297, p-value = 0.1859
alternative hypothesis: greater
sample estimates:
Same colour statistic           Expectation              Variance
            1.0000000             0.4516129             0.3771359
```

Figura 3: Autocorrelación espacial

Como se puede ver en lo arrojado por la prueba, se determinó que, bajo una confiabilidad del 95 %, la única categoría que no presenta autocorrelación espacial es el nivel de desempeño satisfactorio, teniendo en cuenta que es aquella con el p-valor menor a 0.05.

7.3. Modelo

Categoría	Estimador REPRO	Estimador DES_TRAN
1	-0.00197	0.00104
2	-0.00108	0.00360
3	-0.00359	0.00857
4	0.00669	-0.01314

Figura 4: Resultados modelo mixto

Fuente: Elaboración propia. Datos obtenidos del DNP, 2017

De acuerdo a la tabla anterior, se puede determinar en primera instancia que de las 8 variables iniciales utilizadas en la modelación, solo 2 resultaron ser significativas para evaluar el comportamiento del IDI. Esas variables son el valor de reprobados en el municipio (REPRO) y el valos de deserción en transición (DES_TRAN).

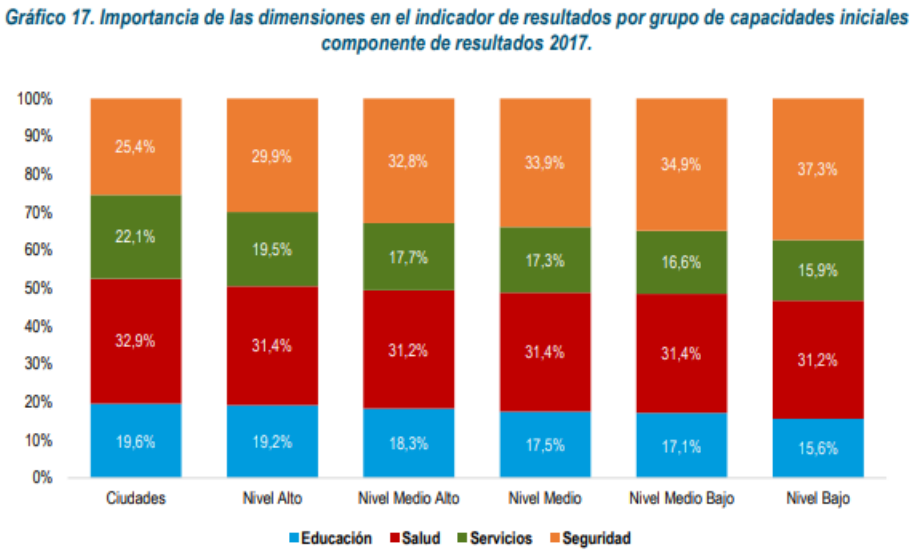

Figura 5: Resultados informe DNP

Fuente: Informe de resultados MDM 2017, DNP

Según un informe presentado por el DNP en el 2017, de acuerdo a la nueva metodología de medición del IDI, se concluyó que dentro de los items que conforman el componente de resultados, el componente educativo (utilizado para el presente artículo), es el que menos porcentaje de avances ha tenido en cuanto al mejoramiento de las condiciones de vida de la población dentro de un ente territorial, es decir, de acuerdo a la Figura 5, educación ha sido uno de los items con avances mínimos en comparación con los demás, esto debido a los resultados obtenidos en las pruebas saber por municipio.

8. Bibliografía

- ANIF Estudios Económicos. (01 de Junio de 2018). *Medición del Desempeño Municipal (MDM): un instrumento para una descentralización efectiva*. Obtenido de Anif - Centro de estudios económicos - Biblioteca Virtual: //www.anif.co/Biblioteca/politica-fiscal/medicion-del-desempeno-municipal-mdm-un-instrumento-para-una
- Dobson, A. J. (2002). *An Introduction to generalized linear models*. New York: CHAPMAN HALL/CRC.
- Treglia, M. L. (2016). *Join Count and Autocorrelation Analyses in R*. Material for Assignment 6 of Landscape Analysis and Modeling, 2 - 8.
- Rodríguez,C. et al. (2017). *El contagio en el fracaso empresarial como consecuencia de la proximidad geográfica: un análisis con los estadísticos join-count aplicado al sector servicios*. Cartagena, España: Revista de métodos cuantitativos para la economía y la empresa.
- Plant, R. E. (2012). *Spatial Data Analysis in Ecology and Agriculture Using R*. New York: CRC Press.
- Agresti, A. (2015). *Foundations of linear and generalized linear models*. Florida: John Wiley & Sons, Inc.
- Hedeker, D. (2003).*A mixed-effects multinomial logistic regression model*. Statistics in medicine, 1433 - 1446.
- Giraldo, R. (Diciembre de 2011). *Estadística espacial - Notas de clase*. Bogotá, Colombia.
- Departamento Nacional de Planeación, DNP. (2017). *Índice de desempeño integral por departamentos*. Bogotá: Departamento Nacional de Planeación, DNP.
- Uribe, J. C. (2016). *Introducción a los modelos mixtos*. Medellín: Universidad Nacional.