Construcción y medición de un indicador socioeconómico mediante el análisis de componentes principales no lineal y técnicas de remuestreo

Katherine Morales
katherine.morales@epn.edu.ec
Miguel Flores
miguel.flores@epn.edu.ec

Escuela Politécnica Nacional
Facultad de Ciencias

4 Julio 2018

Estratificación socioeconómica

Es una forma de clasificación social que busca distribuir a la población en segmentos, generalmente respecto a las características de los individuos, de la vivienda y del hogar.
Generalmente se construye indicadores que recopilen información de interés dentro de la población, mediante métodos estadísticos que permitan clasificar a individuos, hogares y viviendas.

Fuente de información

La Asociación Chilena de Investigación de Mercados concluye que las variables que tienen mayor poder discriminante en la medición del nivel socioeconómico son: la cantidad de bienes presentes en el hogar, la actividad principal y nivel de eduación del jefe de hogar y las características de la vivienda.
Además, el Instituto Nacional de Estadísticas de Chile (INE) basa la estratificación socioeconómica a hogares utilizando variables relacionadas con el hogar, la vivienda, la educación y ocupación del jefe de hogar.

Técnica utilizada

Estudios en el Ecuador, Jesús Tapia (2007) y Andrea Espinoza y Pamela Guevara (2013) presentan el Análisis de Componentes Principales No Lineal como una metodología adecuada para la construcción de un indicador socioeconómico a nivel nacional

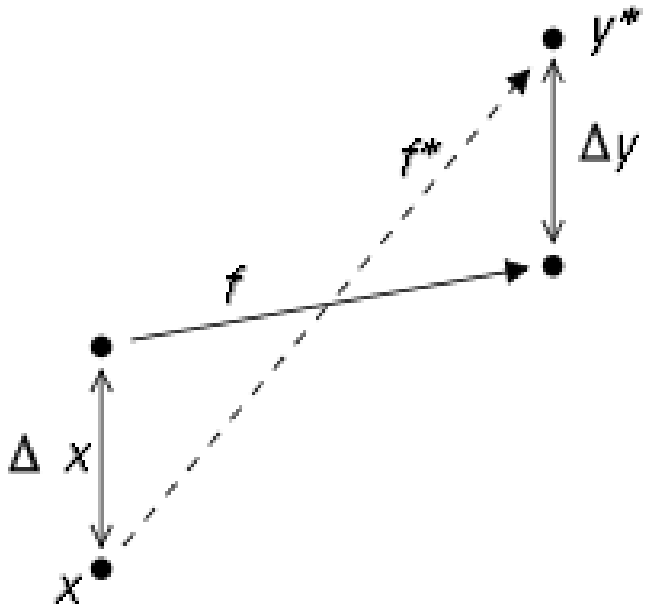
¿Por qué?

Permite incorporar variables numéricas y categóricas.

Cuantifica las variables mediante el proceso de escalamiento óptimo.

Reduce la información (reducción de dimensiones).

Estabilidad



Una solución es estable si *"un cambio pequeño y sin importancia en los datos, el modelo o la técnica conduce a un cambio pequeño y sin importancia en los resultados "*.

*Bootstrap

Análisis de Componentes Principales

Sea la matriz de datos $\mathbf{H} \in \mathbb{R}^{n \times m}$, el Análisis de Componentes Principales está formulado en términos de la función de pérdida:

$$\sigma(Z, A) = tr(\mathbf{H} - ZA^{\top})^{\top}(\mathbf{H} - ZA^{\top}) \tag{1}$$

Donde Z es una matriz de tamaño $n \times r$, n puntuaciones de r componentes, $r \in [1, m]$ y A es una matriz de pesos de tamaño $m \times r$, que contiene los coeficientes de las combinaciones lineales.
El valor mínimo de la función de pérdida (1) sobre A y Z se encuentra por la descomposición propia de S o por la descomposición en valores singulares de $\mathbf{H}$.

Escalamiento Óptimo

El escalamiento óptimo es una técnica de cuantificación que asigna valores numéricos a las categorías de las variables bajo las restricciones del nivel de análisis de la variable:

- El nivel nominal múltiple donde los valores de una variable representan categorías no ordenadas.
- El nivel nominal simple donde los valores representan categorías no ordenadas a la vez que presentan la característica de la dicotomía.
- El nivel ordinal donde los valores de una variable representan categorías ordenadas
- El nivel numérico en el cual los valores de una variable representan categorías ordenadas con una métrica de manera que se preserve la distancia original de las categorías.

- Aproximación low-rank

$$\sigma_L(Z, A, \mathbf{H}^*) = tr(\mathbf{H}^* - ZA^\top)^\top(\mathbf{H}^* - ZA^\top) \tag{2}$$

Bajo las restricciones

$$\mathbf{H}^{*\top} u = 0_m$$

$$diag\left[\frac{\mathbf{H}^{*\top}\mathbf{H}^*}{n}\right] = I_m$$

Funciones de pérdida

- Análisis de homogeneidad

$$\sigma_H(Z, W) = \sum_{j=1}^{m} tr[(Z - G_j W_j)^\top (Z - G_j W_j)] \quad (3)$$

Bajo las restricciones

$$\begin{aligned} Z^\top u &= 0_r \\ Z^\top Z &= n I_r \end{aligned}$$

donde W_j contiene las cuantificaciones de las categorías de la variable j y G_j es la matriz indicatriz de la variable j

Funciones de pérdida

Operando las funciones dada y considerando las restricciones de cada una se obtiene las siguientes funciones de pérdida:

- Aproximación Low Rank

$$\sigma_L(Z, A, \mathbf{H}^*) = nm - 2tr(\mathbf{H}^{*\top} ZA^\top) + ntr(AA^\top) \quad (4)$$

- Análisis de homogeneidad

$$\sigma_H(Z, W) = nrm + ntr[A^\top A] - 2tr[A^\top H^{*\top} Z] \quad (5)$$

Mínimos Cuadrados Alternantes

Dada $\sigma(x_1, x_2, x_3)$ una función de pérdida donde (x_1, x_2, x_3) son las matrices de parámetros de la función. Denotamos la t-ésima estimación de x como $x^{(t)}$. Para minimizar $\sigma(x_1, x_2)$ sobre x_1 y x_2, el algoritmo de ALS actualiza las estimaciones de x_1 y x_2 al resolver el problema de mínimos cuadrados para cada parámetro:

$$
\begin{aligned}
x_1^{(t+1)} &= \arg\min_{x_1} \quad \sigma(x_1, x_2^{(t)}, x_3^{(t)}) \\
x_2^{(t+1)} &= \arg\min_{x_2} \quad \sigma(x_1^{(t+1)}, x_2, x_3^{(t)}) \\
x_3^{(t+1)} &= \arg\min_{x_3} \quad \sigma(x_1^{(t+1)}, x_2^{(t+1)}, x_3)
\end{aligned}
$$

Bootstrap

- Obtener B muestras bootstrap de la muestra principal de tamaño $n \times m$, con n. Cada muestra bootstrap contiene observaciones de la muestra principal.
- Ejecutar el algoritmo en cada una de las muestras bootstrap.
- Los resultados son B valores para cada uno de los valores de interés, en este proyecto son las cuantificaciones de las categorías de las variables.
- Para cada valor de resultado, estos B valores bootstrap forman una distribución bootstrap a partir de la cual se puede calcular un intervalo de confianza.

Interpretación: Si presentan intervalos de confianza pequeños, las cuantificaciones de las muestras bootstraps no son muy diferentes, en este caso, la solución es estable

Cómo construir el indicador

El procedimiento utilizado para la construcción del indicador se sigue de García (2010), el cual se basa en la matriz de pesos[1] de tamaño $m \times r$ A, el vector λ de tamaño $r \times 1$ el vector de ponderación de las r componentes y la matriz de datos cuantificada $\mathbf{H}^*$. El indicador se construye de la siguiente manera:

$$I = \mathbf{H}^*(A\lambda)$$

Re-escalamiento:

$$I_r = \frac{I - \min(I)}{\max(I) - \min(I)} \tag{6}$$

Aplicación

Análisis de datos:
Análisis de las variables que describen el acceso a servicios, tenencia de bienes y condiciones de vida de los hogares del país.
Diseño muestral:

- Seleccionar una variable estratificadores a través de la etapa anterior: Agua sin tratar, Agua hervida, Comprar agua purificada, Agua tratada con cloro, Agua filtrada.
- Tomar una muestra representativa dentro de cada provincia utilizando la técnica de muestreo estratificado.

Aplicación

Muestreo Estratificado:

Consideramos cada población p_i heterogénea con N_i hogares para $i = 1, \ldots, 24$, recordemos que la subdividimos en 5 subpoblaciones denominados estratos lo más homogéneos posibles, a través de la siguiente fórmula:

$$n_i = \frac{N(z_{\alpha/2})^2 \sum_{j=1}^{K} N_{i_j} s_i^2}{E^2 N^2 + (z_{\alpha/2})^2 \sum_{j=1}^{K} N_{i_j} s_{i_j}^2}$$

Donde:

N_i: número de elementos de la población p_i, $i = 1, \ldots, 24$

K: número de estratos $K = 5$

N_{i_j}: El número de elementos (hogares) en el estrato j, $j = 1, \ldots, K$

$s_{i_j}^2$: la varianza en el estrato j, $j = 1, \ldots, K$

$z_{\alpha/2}$: el coeficiente asociado a la distribución normal estándar

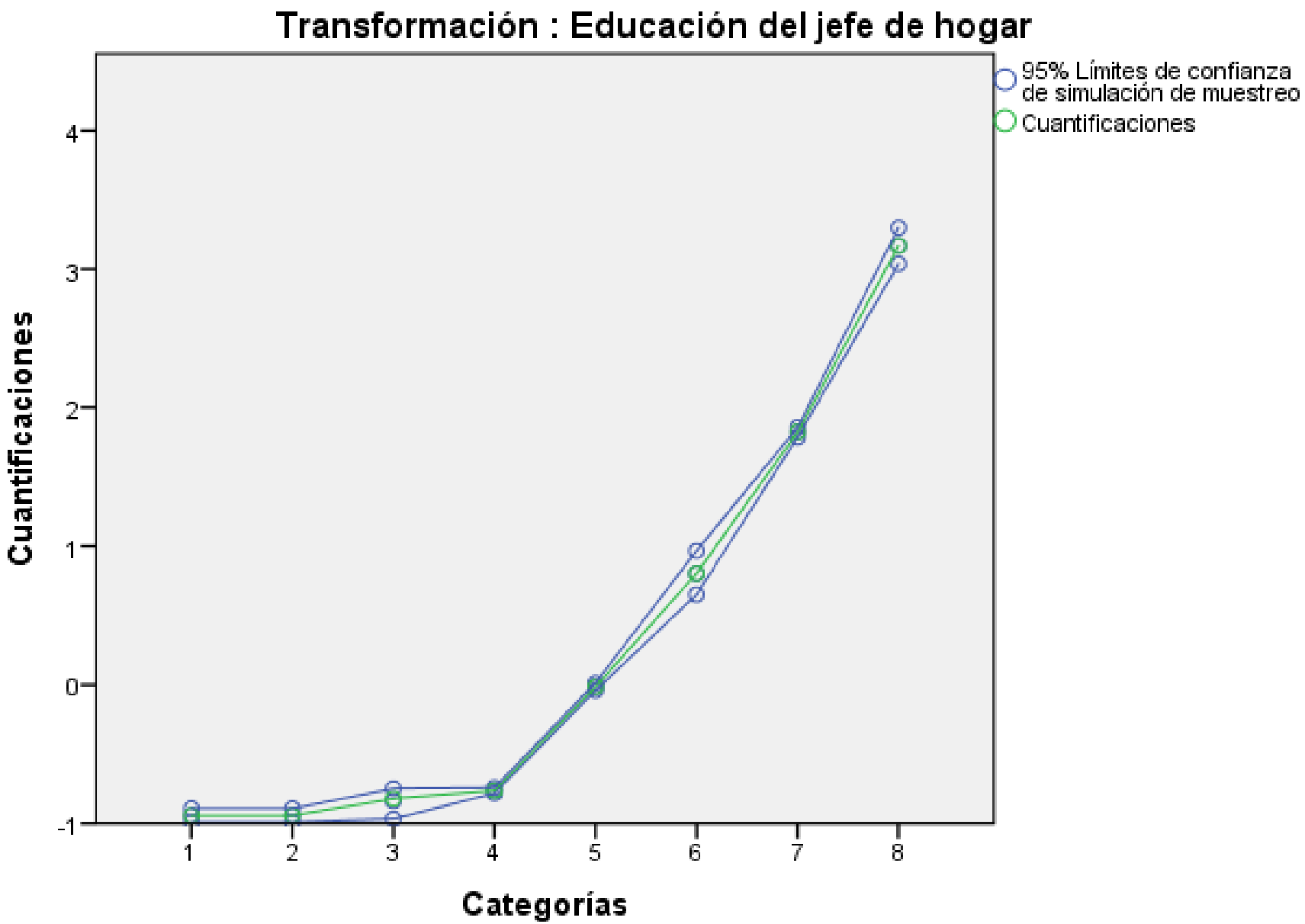
E: es el error de la estimación.

Aplicación

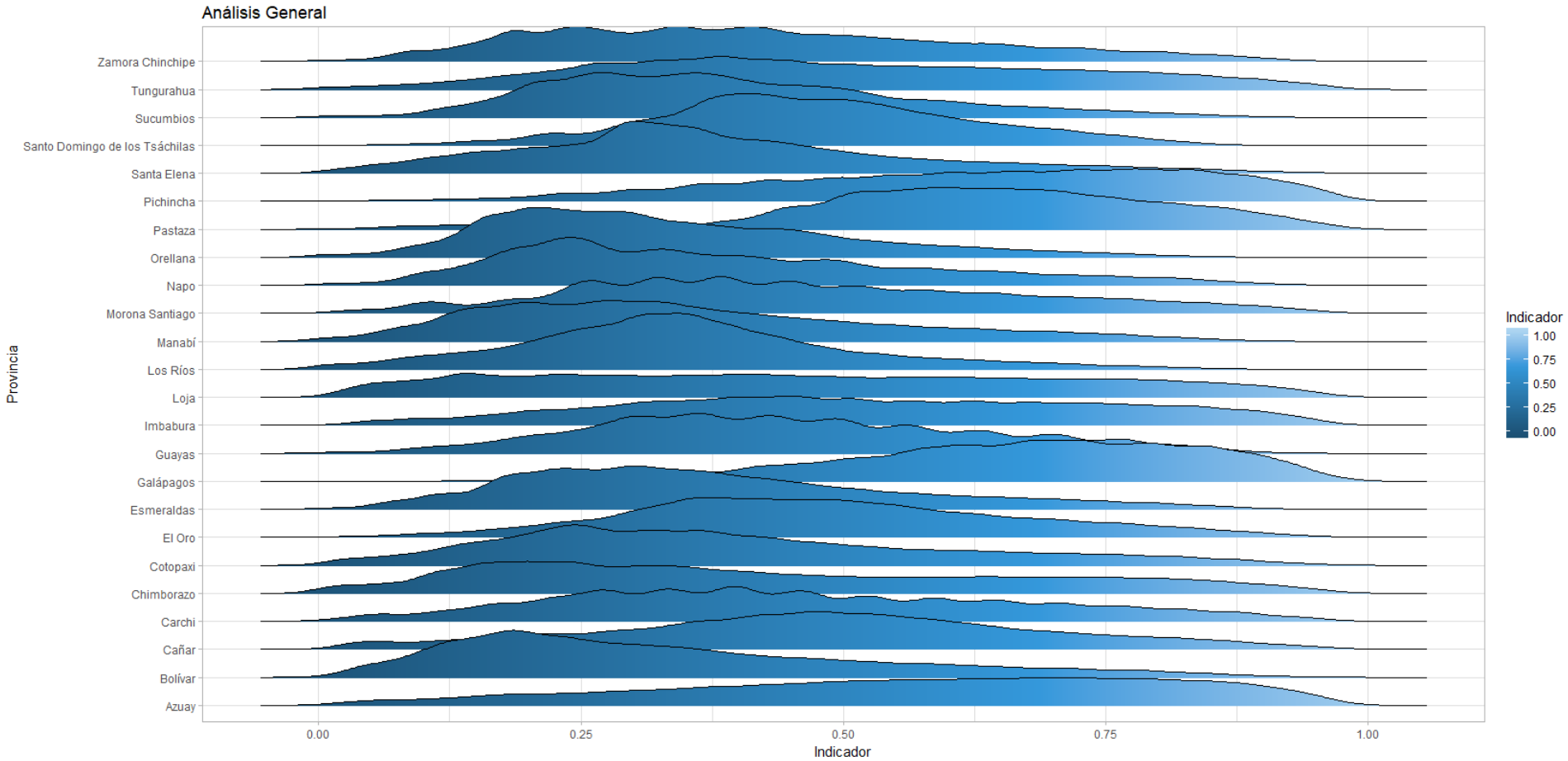
Resultados del algoritmo:

- Analizar las variable que mas aportan a la varianza de la primera componente. (primera estancia)
- Obtener las cuantificaciones de las categorías.
- Analizar la estabilidad de los resultados.

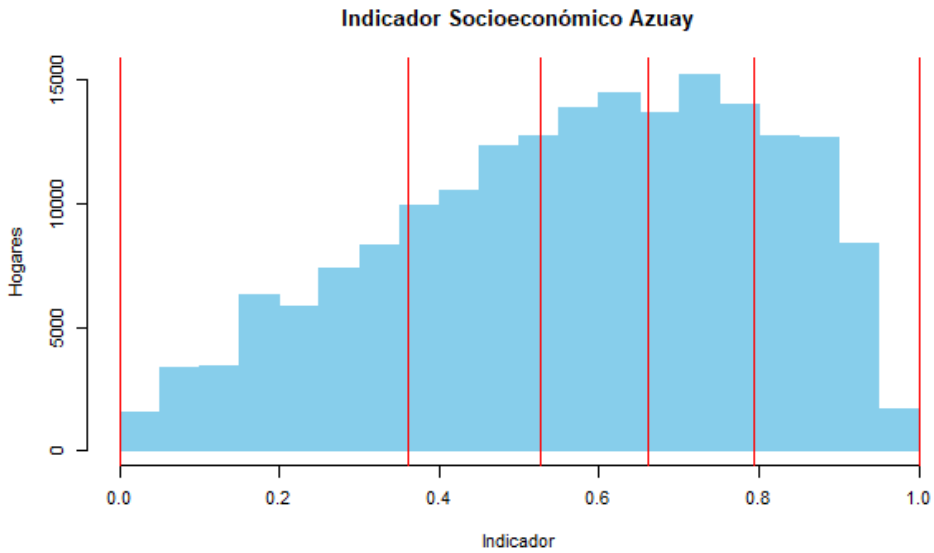
Aplicación

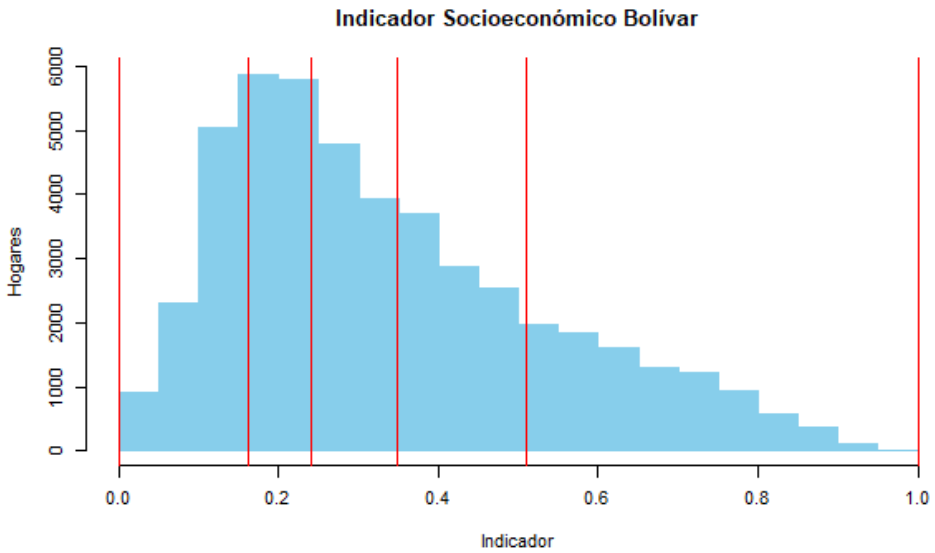

Transformación : Educación del jefe de hogar
95% Límites de confianza de simulación de muestreo
Cuantificaciones
Cuantificaciones
Categorías
4
3
2
1
0
-1
1
2
3
4
5
6
7
8

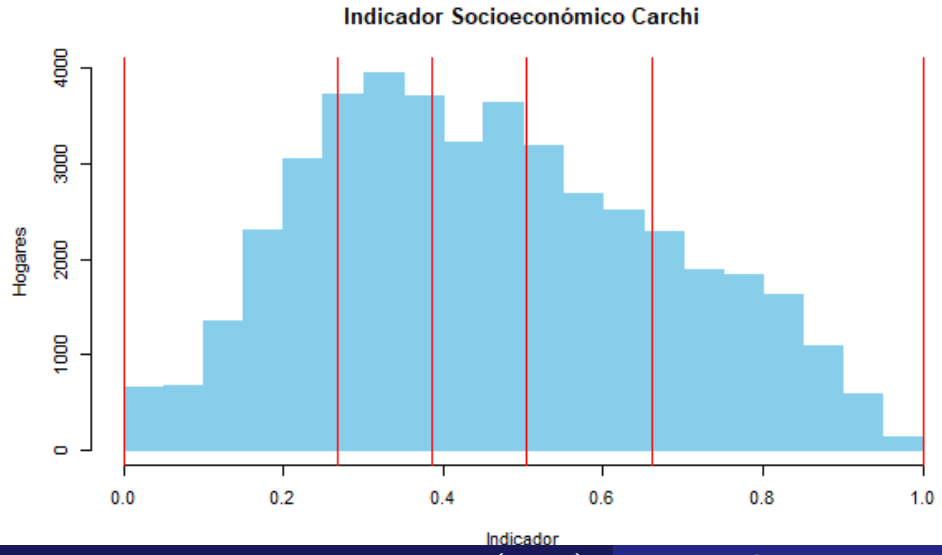
Creación del indicador

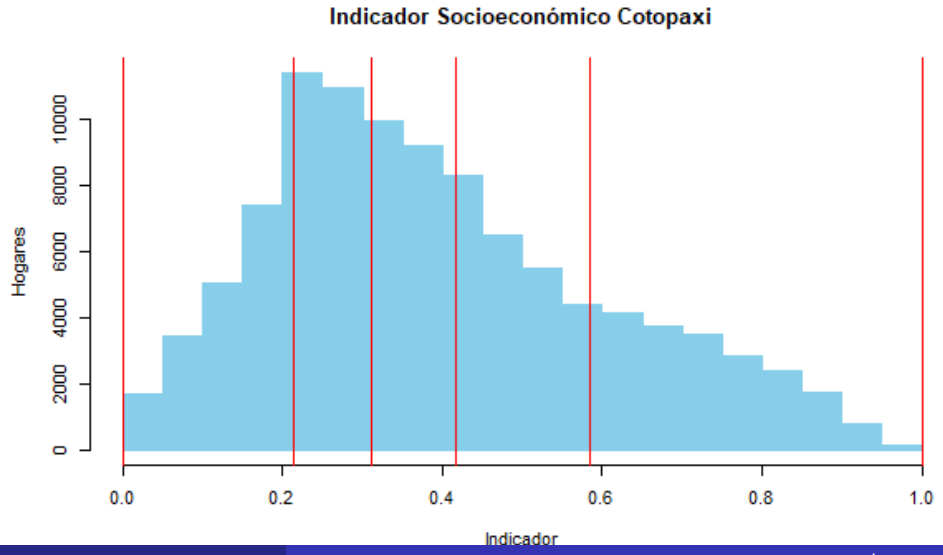


Análisis por quintiles

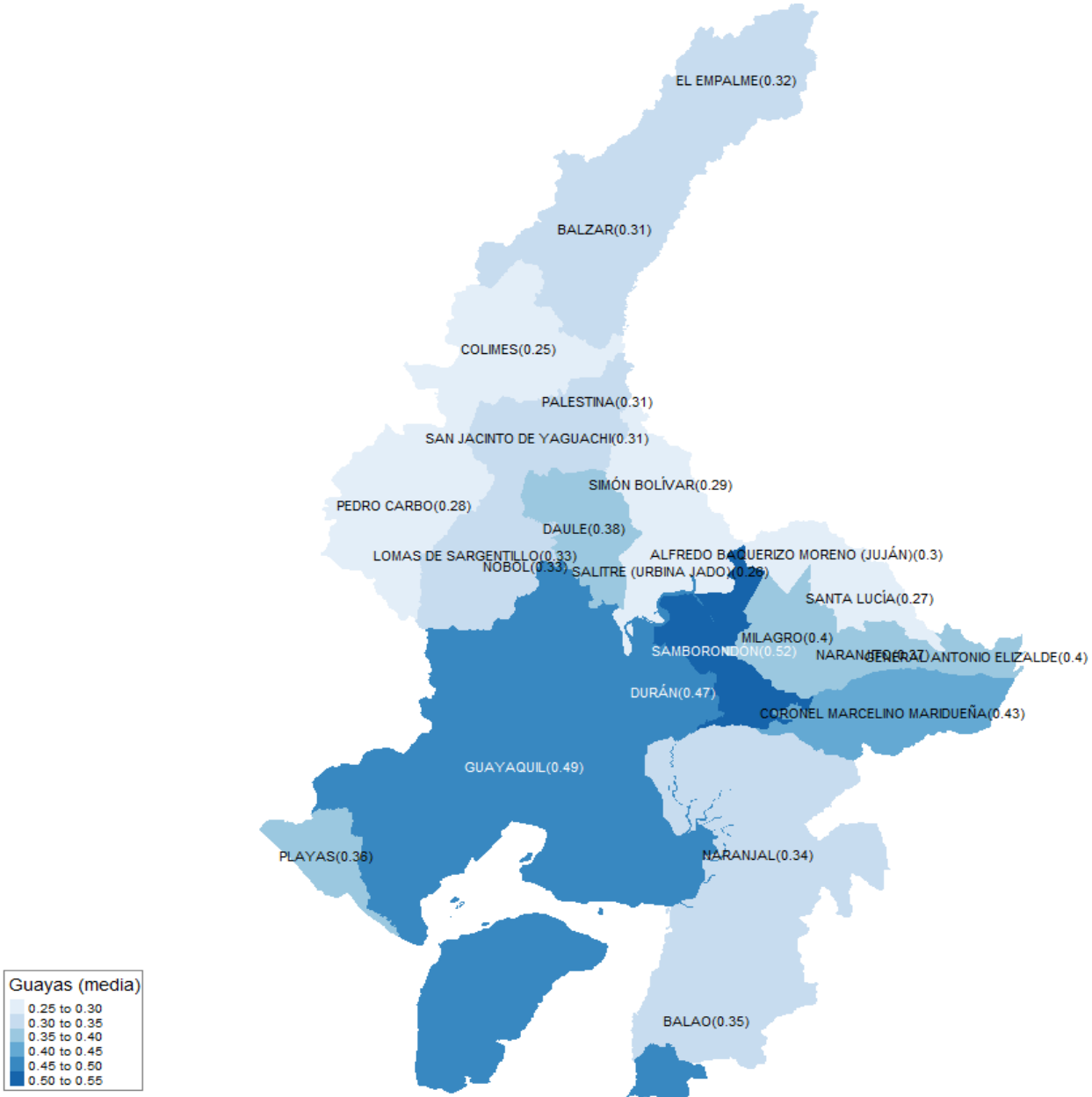






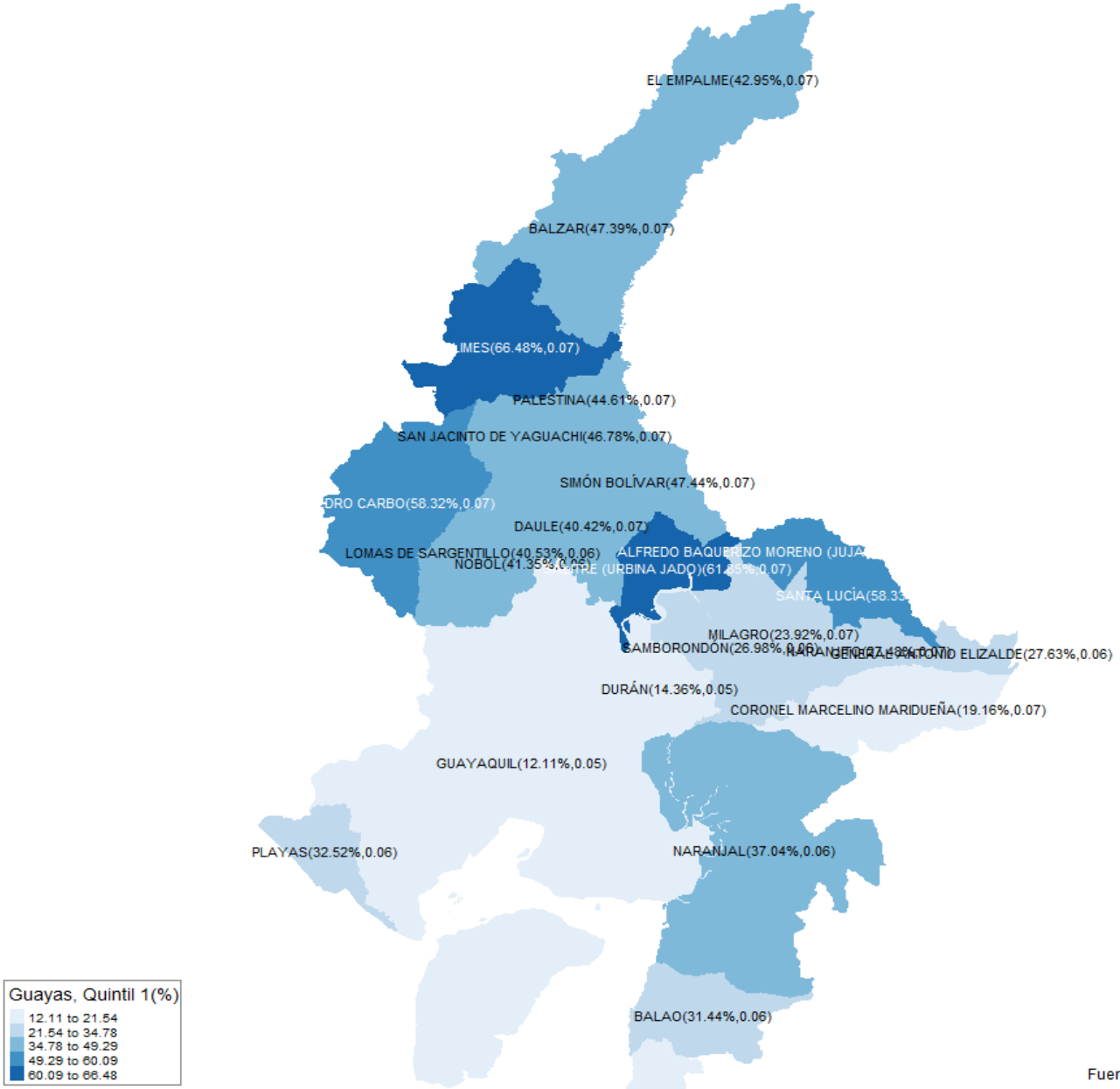


Caso: Guayas



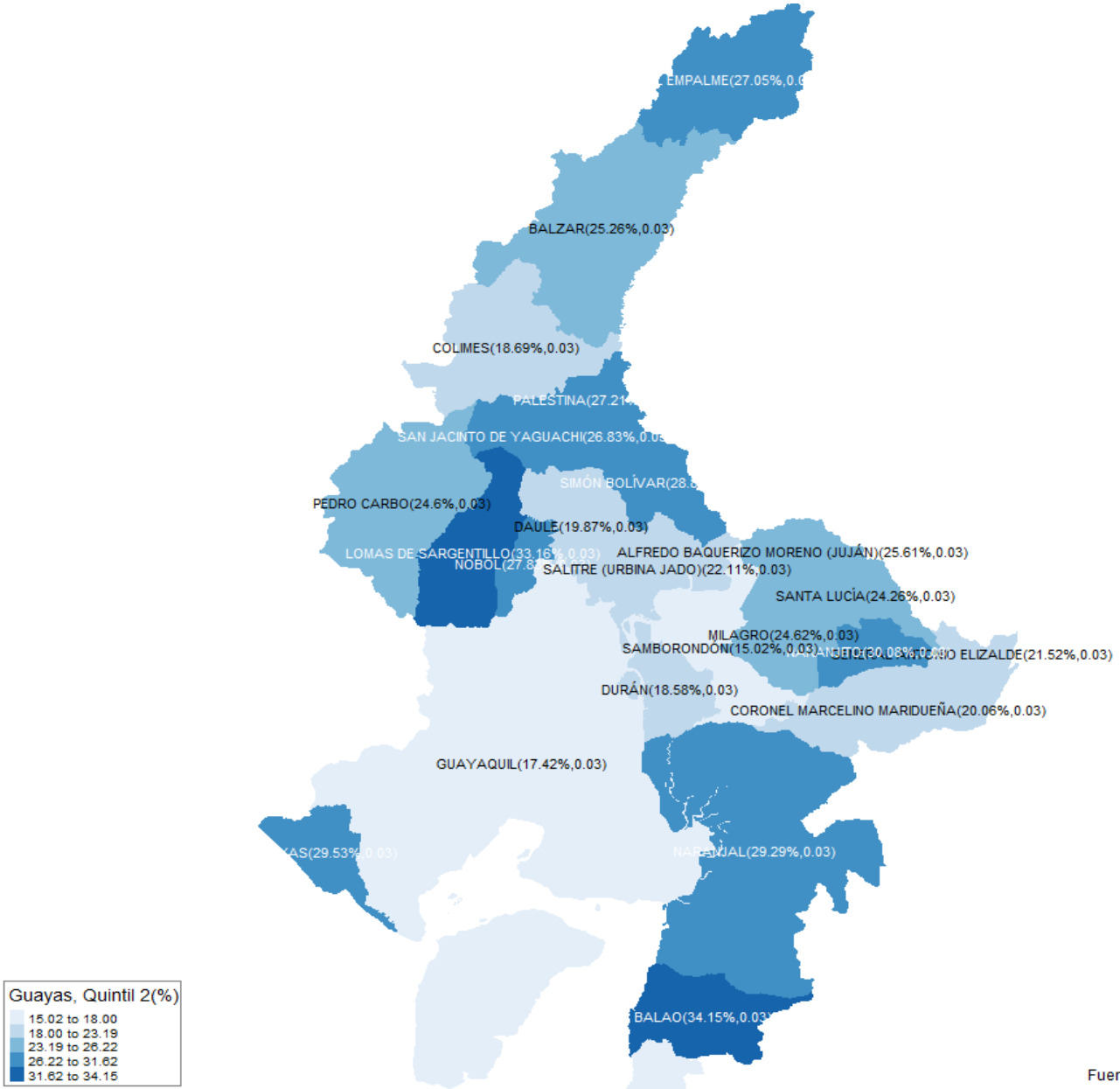
Fuente: Elaboración propia

Primer Quintil



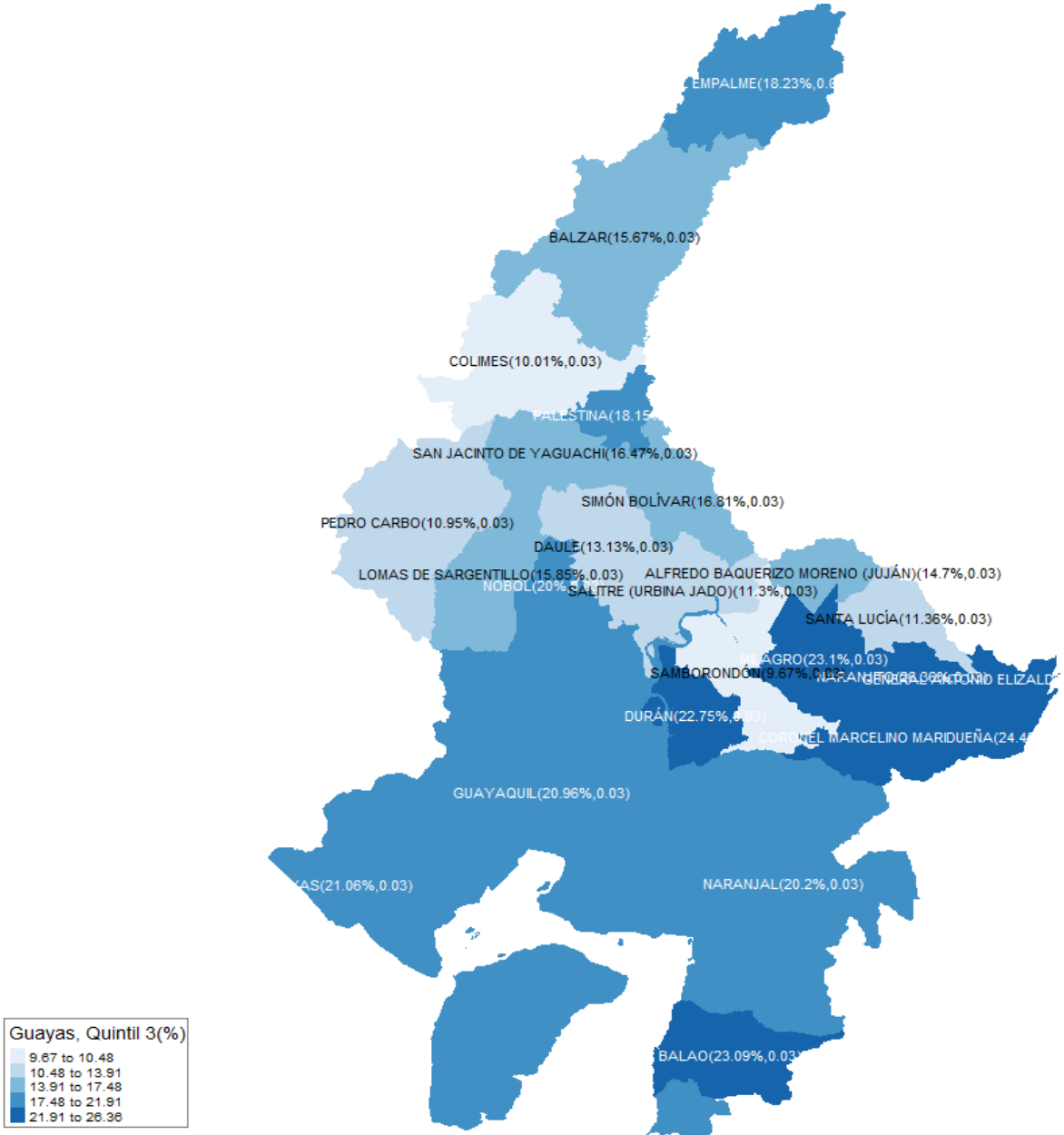
Fuente: Elaboración propia

Segundo Quintil



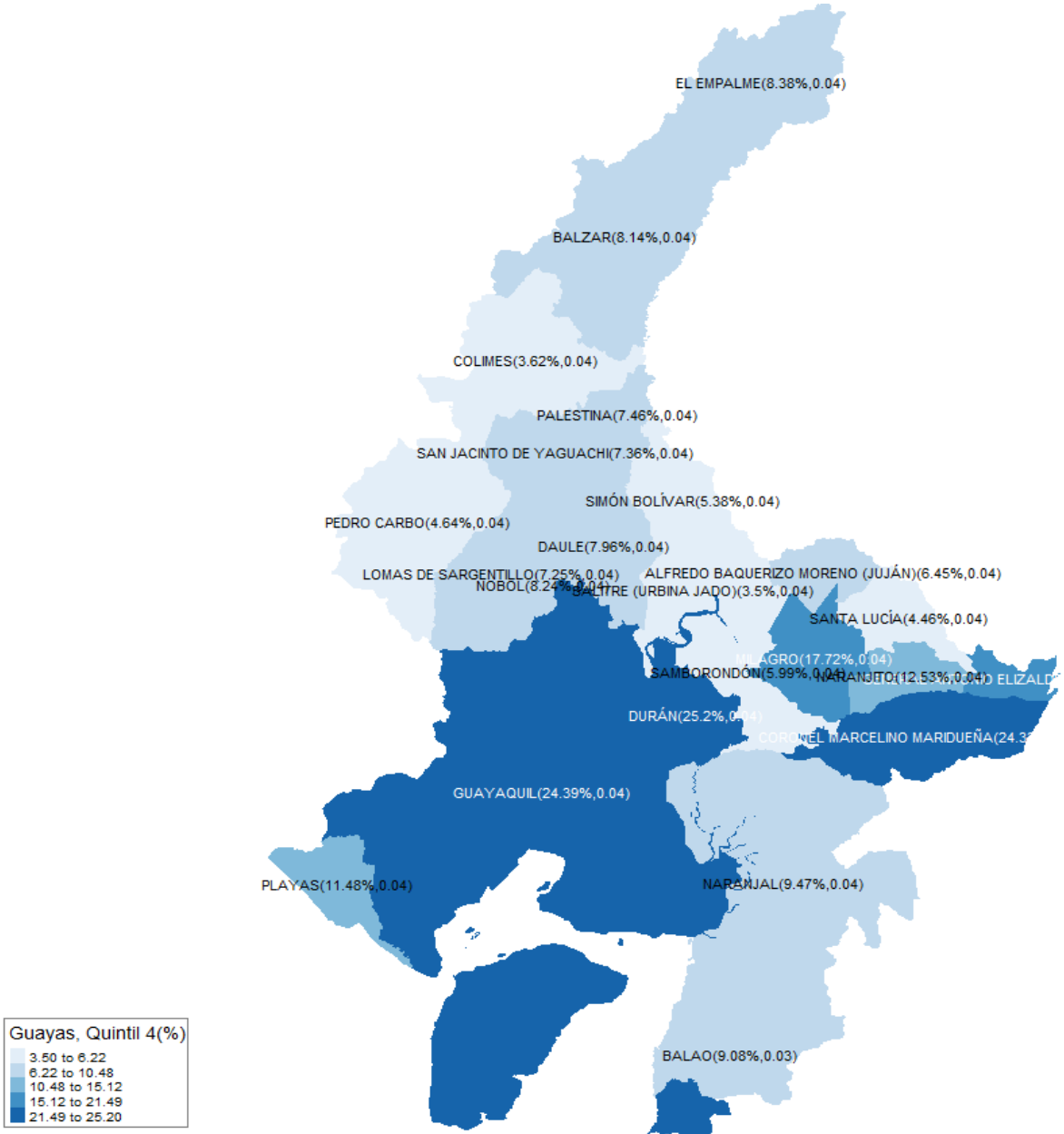
Fuente: Elaboración propia

Tercer Quintil



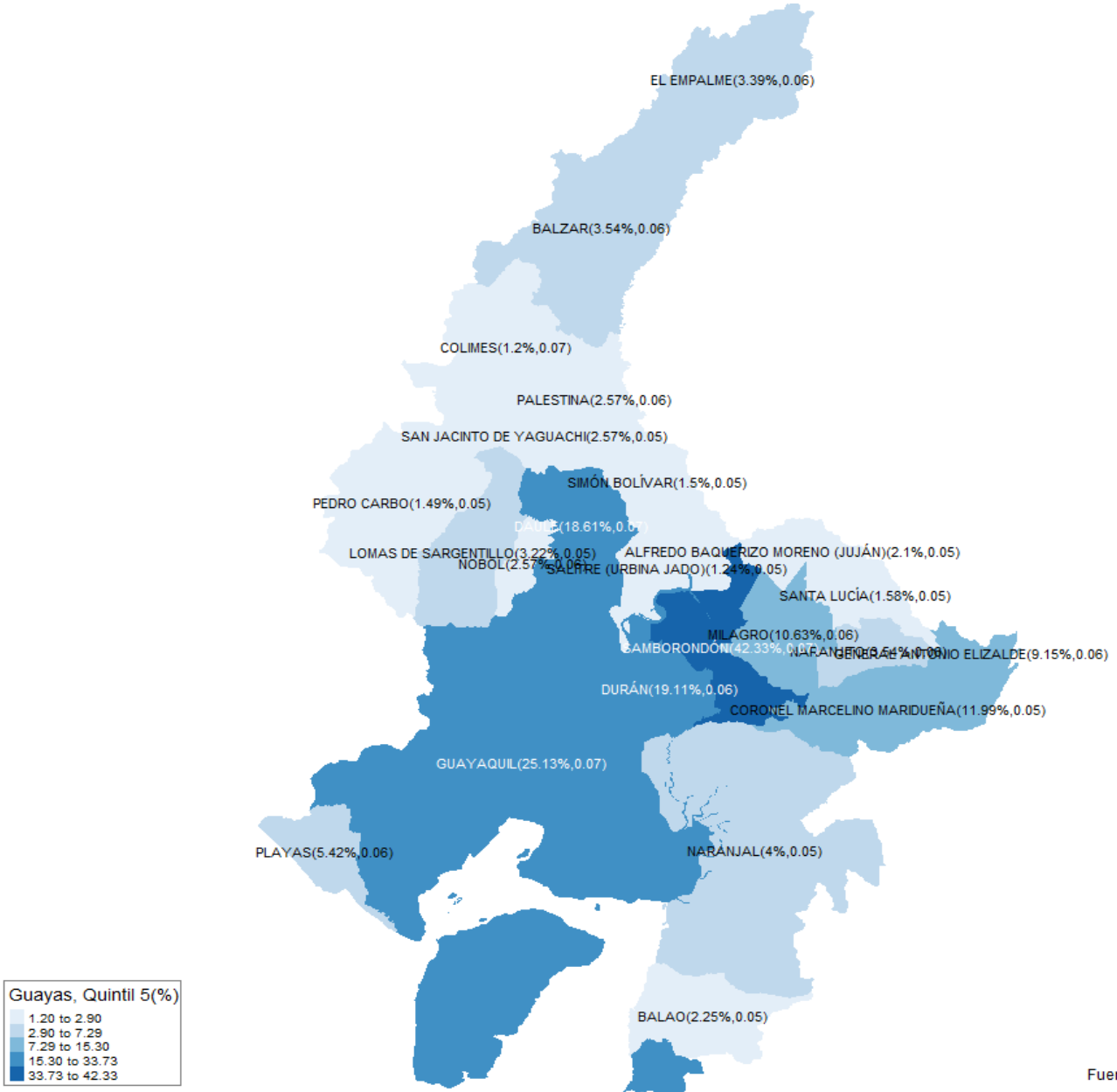
Fuente: Elaboración propia

Cuarto Quintil



Fuente: Elaboración propia

Quinto Quintil



Fuente: Elaboración propia

Sobre los autores:

Katherine Morales

Egresada de la carrera de Ingeniería Matemáticas con mención en Estadística e Investigación Operativa de la Escuela Politécnica Nacional (EPN). Pasante en procesamiento de imágenes en el Centro de Modelización Matemática MODEMAT, EPN. Pasante en manejo de bases de datos en ICA Consultores Cia Ldta. Pasante en el Departamento de Matemática de la Universidade Da Coruña, España en el área de Análisis Multivariante e Introducción a la Estadística Espacial. Capacitadora del lenguaje de programación R en la Asociación de Estudiantes de Matemática e Ingeniería Matemática AsoiMat de la EPN. Nivel intermedio en el manejo de datos espaciales y en sistemas de información geográfica (SIG). Nivel intermedio-avanzado en la automatización de modelos estadísticos, desarrollo de productos estadísticos, desarrollo de aplicaciones de consulta y visualización de información. Alto nivel en el lenguaje de programación R, RMarkdown y Shiny. Se desempeñó como Científica de Datos Jr en Lógika Inteligencia de Mercados.

Miguel Flores

Ph.D. (c) en Estadística e Investigación de Operaciones y Máster en Técnicas Estadísticas de la Universidad de La Coruña. Magister en Investigación Operativa con mención en Sistemas Logísticos y de Transporte de la EPN. Diplomado en Data Mining y Descubrimiento del Conocimiento de la Universidad Iberoamericana Ciudad de México e Ingeniero en Estadística Informática de la ESPOL. Docente e Investigador del Departamento de Matemática de la EPN. Fundador de la comunidad R Users Group - Ecuador® y capacitador de la Sociedad Ecuatoriana de Estadística en R. Consultor senior en Estadística e Investigación de Operaciones en DS Analytics. Experiencia profesional de más de 17 años en el campo consultoría para empresas privadas, ministerios, secretarías y ONG's y más de 20 años de experiencia docente impartiendo cursos de estadística y uso de R en Ecuador, España, Colombia y Perú. Especialista en el desarrollo de metodologías y paquetes estadísticos en R que permiten el tratamiento de nuevos objetos estadísticos en espacios abstractos (FDA) en la era del Big Data cuyo principal interés es resolver problemas en la industria 4.0 y metrología 4.0. Actualmente asesora a empresas españolas en el área de monitoreo de la eficiencia energética y estudios Interlaboratorio.