

ANÁLISIS DE LOS HURTOS EN COLOMBIA DURANTE EL AÑO 2017

Análisis de los hurtos en Colombia durante el año 2017: una comparación entre los modelos de regresión lineal múltiple y la regresión ponderada geográficamente

Marlon Augusto Aceros Bueno, Nelson Ricardo López Herrera

Trabajo de Grado para Optar el título de Especialista en Estadística

Director

Marianela Luzardo Briceño

PhD en Estadística

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialista en Estadística

Bucaramanga

2018

Tabla de Contenido

	Pág.
1. Justificación	14
2. Antecedentes	15
3. Objetivos	18
3.1 Objetivo general	18
3.2 Objetivos específicos	18
4. Marco teórico	18
4.1 Regresión lineal múltiple y datos espaciales	18
4.2 Regresión geográficamente ponderada	21
4.2.1 Descripción del modelo	21
4.2.2 Ajuste computacional del modelo	26
4.2.2.1 Calibración del modelo	26
4.2.2.2 Medidas de bondad de ajuste	27
4.2.2.3 Test de variabilidad geográfica	28
4.2.2.4 Análisis de residuales	28
5. Metodología	30
5.1 Datos	30
5.2 Procedimientos	33
6. Resultados	35
6.1 Análisis exploratorio de los datos	35
6.2 Modelos lineales múltiples evaluados	41
6.3 Análisis del modelo de regresión lineal múltiple final	43

6.3.1 Diagnóstico del modelo de regresión lineal múltiple	46
6.3.1.1 Normalidad en los residuales	46
6.3.1.2 Homocedasticidad	47
6.3.1.3 Evaluación de la multicolinealidad	48
6.3.1.4 Prueba de autocorrelación en los residuales.....	48
6.3.1.5 Prueba de autocorrelación espacial	48
6.4 Resultados del modelo GWR.....	51
6.4.1 Intercepto del modelo GWR	51
6.4.2 Coeficiente variable logsgpppercapita.....	53
6.4.3 Coeficientes variable logmatri1000p	55
6.5 Diagnóstico del modelo GWR	57
6.6 Comparación entre los modelos de regresión lineal múltiple y el modelo GWR.....	62
7. Conclusiones	62
8. Recomendaciones	65
Referencias.....	66

Lista de Figuras

Figura 1. Diagrama esquemático del tipo de ancho de banda adaptable. Fuente: Pavlyuk (2009)	25
Figura 2. Ejemplo ilustrativo del resultado de un modelo GWR. Fuente: Regresión ponderada geográficamente (GWR)-Ayuda ArcGIS for Desktop (http://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-statistics-toolbox/geographically-weighted-regression.htm).	26
Figura 3 Ejemplo de la salida gráfica del proceso de cálculo del Índice de Morán en ArcMap. Fuente: http://pro.arcgis.com/es/pro-app/tool-reference/spatial-statistics/h-sa-tool-popup.htm ..	30
Figura 4 Matriz de correlación de las variables	36
Figura 5 Mapa de distribución de la variable loghurtos.	37
Figura 6 Mapa de distribución de la variable sgppercapita	38
Figura 7 Mapa de distribución de la variable matriculados en Instituciones de Educación Superior	39
Figura 8 Mapa de distribución de la variable categoría.....	40
Figura 9 Resumen del modelo final	44
Figura 10 Diagramas de dispersión loghurtos vs lossgppercapita y loghurtos vs logmatri1000p	45
Figura 11 Diagramas de caja de la variable loghurto vs categoria	46
Figura 12 qqplot de los residuales del modelo lineal múltiple	47
Figura 13 Plot de los residuales vs los valores predichos.....	47
Figura 14 Resultado del Factor de Inflación de Varianza.....	48
Figura 15 Resultado del índice de Morán	49
Figura 16 Mapa de distribución de los residuales del modelo lineal múltiple.....	50

Figura 17 Mapa de distribución del intercepto del modelo GWR	52
Figura 18 Mapa de distribución del coeficiente de la variable logsgppercapita	54
Figura 19 Mapa de distribución del coeficiente de la variable logmatri1000p	56
Figura 20 resumen y medidas de diagnóstico del modelo GWR	57
Figura 21 Resultado del test de variabilidad espacial. Resultados obtenidos a partir del programa GWR4.	58
Figura 22 Mapa de distribución de los residuales del modelo GWR	59
Figura 23 Resultado del Índice de Morán aplicado a los residuales estandarizados del modelo GWR	60
Figura 24 Coeficiente de determinación R2 local según el modelo GWR	61

Lista de Tablas

Tabla 1 Interpretación del índice de Morán según el p-valor y el <i>Z-score</i>	29
Tabla 2 Variables evaluadas	31
Tabla 3 Criterios para categorizar a los municipios según el artículo séptimo de la Ley 1551 de 2012.....	33
Tabla 4 Modelos lineales evaluados	41
Tabla 5 Mínimo, máximo, rango y valor promedio de los estimadores de las variables predictoras del modelo GWR	58
Tabla 6 Principales medidas de bondad de ajuste para la comparación entre modelos.....	62

RESUMEN

TÍTULO: ANÁLISIS DE LOS HURTOS EN COLOMBIA DURANTE EL AÑO 2017: UNA COMPARACIÓN ENTRE LOS MODELOS DE REGRESIÓN LINEAL MÚLTIPLE Y LA REGRESIÓN PONDERADA GEOGRÁFICAMENTE*

AUTOR: MARLON AUGUSTO ACEROS BUENO, NELSON RICARDO LÓPEZ HERRERA**

PALABRAS CLAVE: FACTORES DE LA CRIMINALIDAD, HURTO, ESTADÍSTICAS CRIMINALES, MEDICIÓN DE LA CRIMINALIDAD, ESTADÍSTICA, GWR, OLS

DESCRIPCIÓN:

Según información procedente del observatorio del delito de la policía nacional de Colombia, los hurtos a personas y celulares han presentado una tendencia al alza desde el año 2003 (Norza, Peñalosa y Rodríguez, 2017). Situación que motivó la realización del presente estudio que consistió en identificar variables socio-económicas segregadas a nivel municipal que puedan relacionarse (o explicar) el hurto a personas y celulares durante el año 2017 mediante un modelo de regresión lineal múltiple y una regresión ponderada geográficamente (GWR).

Como resultado, la regresión lineal múltiple explica el 69.5% de la variabilidad del logaritmo del hurto a personas y celulares en 532 municipios, por medio de las variables: matriculados en instituciones de educación superior por cada mil personas, presupuesto per cápita asignado por el Sistema General de Participaciones y la categoría del municipio. Por otro lado, los resultados obtenidos mediante el modelo de regresión geográficamente ponderada explican el 50.16% del logaritmo del hurto con el uso de las mismas variables predictoras, excluyendo la categoría.

De acuerdo con los resultados, se concluyó que el modelo de regresión múltiple logró un mayor coeficiente de determinación en comparación con el modelo espacial. Dicha mejoría es explicada por la inclusión de una variable categórica que no es aceptada en el modelo geográfico; de igual forma, este último modelo demuestra que existe heterogeneidad espacial resultado de las diferencias sociales y económicas presentes en el territorio.

* Trabajo de Grado

** Facultad de Ciencias. Escuela de Matemáticas. Especialización en Estadística. Director: Marianela Luzardo Briceño, Doctora en Estadística

ABSTRACT

TITLE: ANALYSIS OF THE THEFTS IN COLOMBIA DURING THE YEAR 2017: A COMPARISON BETWEEN THE MULTIPLE LINEAR REGRESSION MODELS AND THE GEOGRAPHICALLY WEIGHTED REGRESSION *

AUTHOR: MARLON AUGUSTO ACEROS BUENO, NELSON RICARDO LÓPEZ HERRERA**

KEYWORDS: Crime factors, Theft, Criminal statistics, Crime measurement, Statistics, GWR, OLS

DESCRIPTION:

According to information from the observatory of the crime of the Colombian national police, robberies to people and cell phones have shown an upward trend since 2003 (Norza, Peñalosa and Rodríguez, 2017). Situation that motivated the realization of the present study that consisted in the comparison of a multiple linear regression model and a geographically weighted regression model, calculated from segregated socio-economic variables at the municipal level, used as predictors of theft during the year 2017.

As a result, multiple linear regression explains 69.5% of the variability of the logarithm of theft to people and cell phones in 532 municipalities, by means of the variables: enrolled in higher education institutions per thousand people, budget per capita allocated by the System General of Participations and the category of the municipality. On the other hand, the results obtained by means of the geographically weighted regression model explain 50.16% of the logarithm of the theft with the use of the same predictor variables, excluding the category.

According to the results, it was concluded that the multiple regression model achieved a higher coefficient of determination in comparison with the spatial model. This improvement is explained by the inclusion of a categorical variable that is not accepted in the geographic model; similarly, the latter shows that there is spatial heterogeneity resulting from the social and economic differences present in the territory.

* Specialization Thesis

** Facultad de Ciencias. Escuela de Matemáticas. Especialización en Estadística. Director: Marianela Luzardo Briceño, Doctora en Estadística

Introducción

El observatorio del delito de la policía nacional de Colombia es el ente encargado de recolectar, tratar, analizar y evaluar la información relacionada con el crimen y su propósito es conocer los factores relacionados con el surgimiento, transformación y variación del delito para contribuir a la toma de decisiones institucionales y la formulación de políticas públicas en materia de prevención y control del crimen (Policía Nacional de Colombia). A partir de esta labor se generan bases de datos de carácter público que fueron usadas en este estudio.

Al ser el hurto un delito de tipo monetario es razonable pensar que puede estar relacionado con indicadores económicos que reflejen las condiciones de inversión social y calidad de vida en los municipios. Este fenómeno ha sido estudiado por Sandoval y Barón (2008) quienes argumentan que el crimen depende de las oportunidades a las que los individuos tengan acceso en la sociedad y que incentiven el desarrollo personal y económico de los individuos que les permita satisfacer sus necesidades básicas.

A partir de los datos y estudios realizados por el mencionado observatorio, desde el 2003 se ha evidenciado un incremento en los casos de hurtos a personas y celulares en el país (Norza et al., 2017). Debido a esta tendencia, se considera importante la realización de estudios que identifiquen los principales factores asociados con estos indicadores, así mismo, explorar las diferencias que pueden existir entre territorios dada la heterogeneidad social, económica y cultural presente en el territorio nacional.

Es por esto que en este trabajo se realiza una comparación entre un modelo de regresión lineal múltiple y un modelo de regresión ponderada geográficamente (GWR por sus siglas en inglés), en el que se pueden identificar variaciones espaciales entre el hurto y las variables explicativas a partir

de la variación de los coeficientes del modelo. Dichos modelos pueden contribuir a la planificación de políticas públicas adecuadas al contexto de cada municipio y que pueden ser promovidas por las diferentes entidades gubernamentales que tendrán impacto en la criminalidad de los municipios del país.

1. Justificación

Desde el año 2003, en Colombia se ha presentado una tendencia al alza en delitos entre los que cabe destacar el hurto a residencias, motocicletas, entidades comerciales y a personas, mientras que el homicidio y el hurto a automotores han disminuido ligeramente (Norza et al., 2017). Dicho comportamiento impulsa la realización de investigaciones que ayuden a entender el contexto en el que ocurren los delitos y las variables relacionadas con el crimen, para así, poder entender las causas directas e indirectas que propician y/o facilitan la actividad delincuencial. En este trabajo se busca encontrar las posibles relaciones existentes entre algunas variables socioeconómicas-demográficas y el hurto en los municipios de Colombia.

Dada la extensión nacional del estudio y el nivel municipal de los datos, es posible que se encuentren diferencias significativas entre los distintos municipios dadas las grandes discrepancias socioculturales y económicas entre ellos.

En este orden de ideas, debido a la heterogeneidad del territorio nacional, entendiendo ésta como diferencias de tipo cultural, social y económica se espera que existan variaciones significativas en las estadísticas criminales entre municipios que no puedan ser explicadas adecuadamente por un modelo de regresión lineal múltiple. Por esta razón, en este trabajo se utilizará la técnica de regresión geográficamente ponderada (Brunsdon, Fotheringham y Charlton, 1996) para obtener modelos de regresión lineales que varían espacialmente para identificar las posibles diferencias de un lugar a otro.

La GWR permite mejorar los ajustes, neutralizar la dependencia espacial en los residuos y conocer la distribución en el espacio de las elasticidades en las variables explicativas y la significación local del modelo. Todo ello, sin renunciar al carácter directo y de respuesta rápida. Además, la combinación de precisión, bajo coste y sencillez les convierten en una alternativa

razonable para municipios de tamaño pequeño o de recursos limitados (Gutiérrez-Puebla, García-Palomares y Daniel-Cardozo, 2012), que necesitan evaluar de forma rápida y precisa el número de hurtos.

2. Antecedentes

Brunsdon et al. (1996) desarrollaron la técnica GWR diseñada con el propósito de tratar con datos no estacionarios espacialmente, en los cuales un solo modelo no puede explicar adecuadamente las relaciones entre las variables. En términos sencillos esta técnica consiste en calibrar múltiples modelos de regresión para capturar las diferencias existentes de un lugar a otro, permitiendo estimar los parámetros en cada lugar y representarlos gráficamente en un mapa.

Desde la formulación del modelo GWR, este ha sido ampliamente utilizado para la investigación en temas con variables georreferenciadas, donde se sospecha que puede haber cambios considerables en el comportamiento de las variables analizadas (*i.e.* sus coeficientes) en la región de estudio. Algunos ejemplos de estudios que emplean la técnica GWR han sido sobre la salubridad pública (Fotheringham, Charlton y Brunsdon, 1998; Wheeler y Tiefelsdorf, 2005), mercado inmobiliario (Brunsdon, Fotheringham y Charlton, 1999), en exploración geoquímica (Zhao, Wang y Cheng, 2014), en estudios de movilidad (Gutiérrez-Puebla et al., 2012), en estudios del rendimiento académico (Fotheringham, Charlton y Brunsdon, 2001), estudios biológicos-ecológicos (Ortiz-Yusty, Páez y Zapata, 2013) entre muchos otros campos.

Con respecto al estudio del comportamiento criminal Sandoval y Barón (2008) analizaron el estado del arte de los diferentes temas relacionados con la criminalidad, concluyendo que los fenómenos criminales

...tienen origen en aspectos de los individuos, los cuales dependiendo de las posibilidades que tengan en la sociedad pueden llegar a ser productivos para ésta ó pueden tener incentivos a ser criminales y delincuentes dadas las limitaciones que encuentren a progresar personal y laboralmente; así mismo a partir de las características propias de las zonas geográficas la delincuencia tiende a ser una forma de vida para las personas que aprovechan los recursos que brindan las regiones, lo cual puede llegar a propagarse hacia las zonas aledañas y convertirse en aros de criminalidad que a su vez pueden generar violencia producto del aprovechamiento que desean tener las redes delincuenciales de los dividendos generados por las actividades ilícitas. (Sandoval y Barón, 2008, p. 105).

Los aspectos relacionados con la criminalidad también han sido analizados mediante el uso de la GWR, tal es el caso del estudio del efecto del terrorismo en el crecimiento económico de Turquía (Öcal y Yildirim, 2010) y el estudio de los crímenes violentos en la ciudad de Portland, Estados Unidos realizado por Cahill y Mulligan (2007) que demuestran la utilidad de la técnica para modelar aspectos sociales como el crimen y reconocen un mejor ajuste en comparación con los modelos de regresión lineales.

En Colombia, la GWR ha servido para analizar los factores asociados con los cultivos de coca y comprobar que dichas relaciones varían espacialmente (Rincón-Ruiz, 2013; Rincón-Ruiz, Pascual y Flantua, 2013). Fowler (2013) analizó algunas variables socioeconómicas y su relación con las tasas de homicidios a nivel municipal durante el año 2005.

Por otro lado Norza et al. (2017) analizaron en un estudio descriptivo el comportamiento de los principales delitos de alto impacto en la sociedad Colombiana durante el 2015 y 2016.

A partir de lo anterior y basados en las consultas bibliográficas preliminares se evidenció la ausencia de estudios que evalúen la relación entre factores socioeconómicos y los hurtos a personas y celulares durante el año 2017 en Colombia y sus variaciones espaciales.

3. Objetivos

3.1 Objetivo general

Analizar la relación entre los factores socioeconómicos y el hurto en los diferentes municipios de Colombia mediante modelos de regresión lineal múltiple y regresión geográficamente ponderada.

3.2 Objetivos específicos

- Ajustar un modelo de regresión lineal múltiple para explicar el hurto a personas y celulares en Colombia a partir de variables socio-económicas.
- Ajustar un modelo de regresión ponderada geográficamente que incorpore las variaciones geográficas de la relación entre el número de hurtos a personas y celulares durante el año 2017 en Colombia y las variables disponibles.
- Comparar los dos modelos ajustados para así concluir sobre el ajuste y capacidad predictiva de cada uno.

4. Marco teórico

4.1 Regresión lineal múltiple y datos espaciales

En Charlton y Fotheringham (2009) la regresión es descrita como un conjunto amplio de métodos para modelar las relaciones entre una variable dependiente (Y) y un conjunto de variables independientes (X). En la forma más sencilla, un modelo de regresión lineal múltiple puede expresarse de la siguiente forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_m x_{mi} + \varepsilon_i \quad \text{para } i = 1 \dots n \quad (1)$$

En la ecuación anterior y_i y x_i son las variables respuesta e independientes (predictoras) respectivamente, medidas en la localidad i , ε_i representa el error, β_0 y β_{mi} son los coeficientes

estimados para las m variables de tal forma que se minimice el valor de $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ en las n observaciones. $\hat{y}_i$ es el valor predicho por el modelo para la i -ésima observación dado el i -ésimo valor de X .

Las predicciones son obtenidas mediante una combinación lineal de las variables independientes. El término $(y_i - \hat{y}_i)$ es denominado residual para la i -ésima observación. Los residuales deben ser independientes, presentar una distribución normal con media cero y varianza constante. Dichos modelos son comúnmente ajustados usando el procedimiento de mínimos cuadrados ordinarios (OLS por sus iniciales en inglés). Según Brunsdon et al. (1996) el estimador OLS se expresa como:

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (2)$$

Siendo $\hat{\beta}$ el vector de los coeficientes estimados, X es la matriz de diseño, la cual contiene los valores de las variables independientes y una columna con valores 1, y es el vector de los valores observados y $(X^T X)^{-1}$ es la inversa de la matriz de varianza-covarianza (Charlton, Fotheringham y Brunsdon, 2006).

Para medir la bondad de ajuste del modelo a los datos observados se utiliza el coeficiente de determinación R^2 el cual toma valores desde 0 hasta 1 para medir la proporción de la variación en los datos observados que es explicada por el modelo.

Para determinar si cambios en las variables predictoras tienen efecto en las predicciones, los parámetros del modelo $\hat{\beta}$ deben ser significativamente diferentes de 0, para esto se divide el parámetro estimado para cada variable por su respectivo error estándar. El resultado sigue una distribución *t-student* y puede ser comparado con los valores críticos dado el número de grados de libertad del modelo (Charlton et al., 2006).

Una de las suposiciones del modelo es que las observaciones deberían ser independientes, esto generalmente no ocurre con los datos espaciales ya que según la ley de la geografía “*Todo se relaciona con todo, pero las cosas cercanas se relacionan más que las lejanas*” (Tobler, 1970). La correlación espacial puede estar presente en las variables del modelo (*i.e.* lugares cercanos tienen valores similares) y en los residuales. Estas características de los datos espaciales influyen en las estimaciones de los parámetros en un modelo tradicional ya que, si los residuales del modelo tienen una estructura espacial, la estimación de los parámetros será ineficiente, por lo cual los errores estándar de los parámetros serán muy grandes, produciendo que parámetros potencialmente significativos parezcan no serlo. (Gutiérrez-Puebla et al., 2012)

Cuando la regresión lineal es utilizada con datos espaciales la localización en la que fueron tomados los datos no se tiene en cuenta en el proceso de modelado (Brunsdon, Fotheringham y Charlton, 1998) ya que se asume que el proceso examinado es constante en el espacio y por lo tanto se utiliza un solo modelo (Charlton et al., 2006).

En los modelos de regresión lineal múltiple se asume que las relaciones entre las variables son las mismas en la zona de estudio, es decir, hay homogeneidad espacial. Sin embargo, dicha suposición puede no cumplirse ya que los procesos que generan los datos pueden variar con el espacio, lo cual es denominado como heterogeneidad espacial (Charlton y Fotheringham, 2009), la cual puede ser el resultado de diferencias intrínsecas en el espacio debido a diferencias culturales o a contextos políticos distintos de un lugar a otro (Charlton et al., 2006).

Una forma de tener en cuenta la localización de los datos en los modelos de regresión lineal es a través del uso de variables *dummy* que expresen las clases territoriales en que fueron medidos los datos (barrios, municipios, departamentos, etc.), y el uso de modelos multinivel. Estos enfoques

pueden ser útiles para reconocer diferencias entre las unidades territoriales pero no permiten que los coeficientes del modelo varíen espacialmente (Brunsdon et al., 1998).

4.2 Regresión geográficamente ponderada

La regresión geográficamente ponderada fue desarrollada por Brunsdon et al. (1996) como un método para explorar la no estacionariedad espacial, es decir, la condición en la cual un modelo “global” no puede explicar adecuadamente las relaciones entre los conjuntos de variables. Esta situación puede darse en modelos con información georreferenciada en los cuales los coeficientes de la regresión varían en el espacio.

4.2.1 Descripción del modelo La GWR extiende el método de regresión tradicional permitiendo estimar parámetros locales en vez de globales (Charlton et al., 2006), de este modo el modelo puede ser expresado con la siguiente ecuación :

$$y_i = a_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) X_{ik} + \varepsilon_i \quad k = 1, 2, \dots, n \quad (3)$$

Donde (u_i, v_i) denota las coordenadas del i -ésimo punto en el espacio y $\beta_k(u_i, v_i)$ son los coeficientes en el punto i (Brunsdon et al., 1996). Por lo tanto, se permite que exista una superficie continua de valores de parámetros y las mediciones de esta superficie se toman en ciertos puntos para indicar la variabilidad espacial de la superficie (Charlton et al., 2006). De este modo, el modelo de regresión lineal múltiple es un caso particular de la GWR cuando se asume que los parámetros son constantes.

En la GWR los valores próximos al punto i son ponderados de acuerdo a su proximidad, es decir, los más próximos influyen más que los lejanos. Algebraicamente el estimador se puede expresar como:

$$a(u_i, v_i) = (X^t W(u_i, v_i) X)^{-1} X^t W(u_i, v_i) y \quad (4)$$

Donde $W(u_i, v_i)$ es una matriz $n \times n$ compuesta por ceros a excepción de los elementos en la diagonal, los cuales representan las ponderaciones geográficas de los datos observados en el punto i (Charlton et al., 2006).

$$W(u_i, v_i) = \begin{pmatrix} w_{i1} & 0 & 0 & \cdots & 0 \\ 0 & w_{i2} & 0 & \cdots & 0 \\ 0 & 0 & w_{i3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & w_{in} \end{pmatrix}$$

Donde w_{in} corresponde al peso del dato en el punto n en la calibración del modelo alrededor del punto i . Las ponderaciones cambian con i , lo cual marca la diferencia con respecto a la técnica de mínimos cuadrados ponderados donde la matriz de ponderación es constante (Charlton et al., 2006).

En el caso de un modelo de regresión tradicional OLS (*i.e.* global) la matriz de ponderaciones está implícitamente definida como:

$$W(u_i, v_i) = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Lo cual quiere decir que cada dato tiene un peso equivalente a 1, lo cual se traduce en que no hay variaciones espaciales en los parámetros estimados.

Para definir las ponderaciones w_{in} se pueden adoptar varios enfoques. Brunson et al. (1996) mencionan las siguientes opciones:

- Excluir de la calibración del modelo a los datos que se encuentren a una distancia (d) mayor que un valor fijo: Este enfoque consiste en calibrar modelos de regresión solo con los datos que se

encuentran dentro de una distancia fija estipulada por el investigador. Después de determinar la distancia de exclusión (d), se pueden tomar 2 opciones:

- Función de ponderación discreta: Consiste en asignar un peso igual a 1 a los datos que se encuentran dentro de la distancia de exclusión y asignar un valor de 0 a los puntos excluidos (a distancias mayores que d). En este caso se tiene:

$$\begin{aligned} W_{ij} &= 1 \quad \text{si } d_{ij} \leq d \\ W_{ij} &= 0 \quad \text{si es lo contrario} \end{aligned} \quad (5)$$

Donde d_{ij} es la distancia entre el lugar donde se va a calibrar el modelo y los datos. De este modo, todos los datos dentro de la ventana de análisis tendrán el mismo peso y los datos por fuera no serán tenidos en cuenta.

- Función de ponderación continua con exclusión de datos: consiste en formular una función de ponderación que asigne pesos a los datos dependiendo de su cercanía al punto donde se está calibrando el modelo y, que dicho peso disminuya con la distancia hasta alcanzar el valor de exclusión a partir del cual ya no se tendrán en cuenta los datos. Un ejemplo de esto sería:

$$\begin{aligned} W_{ij} &= \left[1 - \frac{d_{ij}^2}{d^2} \right]^2 \quad \text{si } d_{ij} \leq d \\ W_{ij} &= 0 \quad \text{si es lo contrario} \end{aligned} \quad (6)$$

- No excluir datos para la calibración de los modelos: consiste en tener en cuenta todos los datos para estimar el valor de los parámetros en toda la zona de estudio, pero asignando mediante una función continua pesos que disminuyen al aumentar la distancia al punto i donde se va a calibrar el modelo. Un ejemplo de este es el siguiente:

$$W_{ij} = \exp\left(-\frac{d_{ij}^2}{h^2}\right) \quad (7)$$

Donde h es denominado como el ancho de banda. Si i coincide con j (*i.e.* el modelo se va a calibrar en un el punto i donde se midió el dato j), el peso del dato en ese punto será igual a 1. El peso de los otros datos disminuirá según una curva Gaussiana al aumentar la distancia entre i y j . Para datos muy lejos del punto i el peso tenderá hacia 0.

La asignación de funciones discretas para la ponderación tiene el inconveniente de que pueden presentar discontinuidad, es decir, los coeficientes estimados podrían cambiar bruscamente de un sitio a otro. Esto puede reflejar una condición real en el fenómeno estudiado o ser un artefacto producto de la disposición de los puntos de muestreo (Charlton et al., 2006).

Los métodos anteriormente descritos asignan la misma función de ponderación espacial para todos los puntos, esto tiene el inconveniente que en regiones con datos muy dispersos puede resultar en la estimación de los parámetros con muy pocos datos. Para evitar este problema se puede utilizar funciones de ponderación espacialmente adaptables (Figura 1) con anchos de banda (h) pequeños en regiones con datos densamente distribuidos y anchos de banda mayores en zonas con datos dispersos (Charlton et al., 2006). Un ejemplo de esto es:

$$W_{ij} = \left[1 - \left(\frac{d_{ij}}{h_i} \right)^2 \right]^2 \quad \text{si } d_{ij} < h_i \quad (8)$$

$$W_{ij} = 0 \quad \text{si es lo contrario}$$

Donde h_i es la distancia del n -ésimo vecino más cercano al punto i .

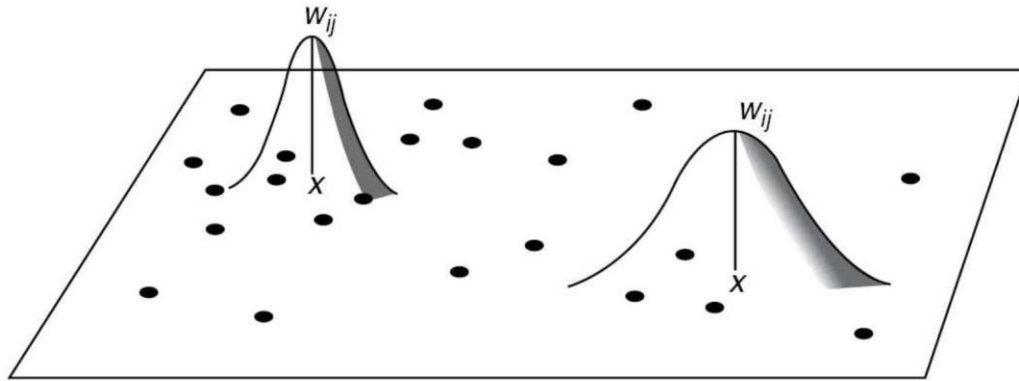


Figura 1. Diagrama esquemático del tipo de ancho de banda adaptable. Fuente: Pavlyuk (2009)

Los resultados obtenidos de la GWR dependerán de la función de ponderación utilizada. Por ejemplo, cuando h (el ancho de banda) tiende a infinito, no habrá disminución en los pesos al aumentar la distancia y todas las ponderaciones tenderán a 1, lo cual resulta en que los parámetros estimados se volverán uniformes y por lo tanto la GWR será equivalente a una regresión tradicional (Charlton et al., 2006). Por lo tanto, se debe buscar un ancho de banda óptimo.

Con la técnica GWR no sólo se obtienen coeficientes locales, también se obtienen versiones locales de los diagnósticos utilizados en la regresión tradicional como la bondad de ajuste (R^2), además permite la estimación de los parámetros en cualquier lugar del espacio, incluso en lugares diferentes a donde fueron medidos los datos (Charlton et al., 2006), esto se ilustra en la Figura 2 donde se observa un ejemplo del resultado de un modelo GWR visualizado en un Sistema de Información Geográfica (SIG). Nótese que la variable respuesta en el ejemplo es la tasa de crimen y las variables predictoras son la población y los ingresos. En este ejemplo las variables son ingresadas al software en modo de polígonos que representan unidades territoriales en donde fueron tomados los datos. Como resultado del modelo se obtiene 3 superficies ráster para cada parámetro (β) en las cuales los valores están codificados según la intensidad del color, para así visualizar de forma más sencilla la heterogeneidad espacial (diferencias en el valor de los parámetros de un punto a otro).

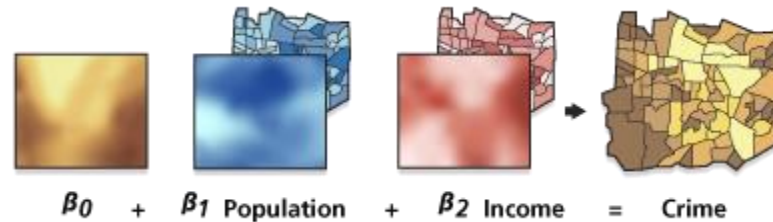


Figura 2. Ejemplo ilustrativo del resultado de un modelo GWR. Fuente: Regresión ponderada geográficamente (GWR)-Ayuda | ArcGIS for Desktop (<http://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-statistics-toolbox/geographically-weighted-regression.htm>)

4.2.2 Ajuste computacional del modelo El programa GWR4 es un programa estadístico de uso libre que permite la manipulación y procesamiento de datos georreferenciados diseñado para la estimación de modelos GWR. la cual incluye varias herramientas para calibrar y evaluar los modelos.

4.2.2.1 Calibración del modelo De acuerdo con Nakaya (2016) el programa GWR configurar las siguientes opciones para realizar la calibración de los modelos:

- *Kernel type*: Esta opción permite establecer el tipo de función que se utilizará para definir las vecindades al punto i donde se va a calibrar el modelo. Las opciones disponibles son:
- Distancia Fija Gaussiana: utiliza una ponderación decreciente a partir del punto que se quiere estimar sin llegar al valor de cero según a ecuación (7).
- Distancia Fija *bi-square*: utiliza la ecuación (8) para establecer las ponderaciones.
- Adaptativo *bi-square*: utiliza la ecuación (6) para establecer las ponderaciones.
- Adaptativo Gaussiano: usa la ecuación (7) modificada para considerar un ancho de banda dependiente de la posición (h_i).
- Método de selección para el ancho de banda:

- AICc: la extensión del *kernel* es determinada utilizando el Criterio de Información de Akaike corregido. El cual es una medida del rendimiento del modelo y sirve para comparar entre modelos distintos (con diferentes variables explicativas, pero con la misma variable respuesta). También sirve para comparar un modelo GWR con un modelo de regresión tradicional OLS (Environmental Systems Research Institute [ESRI], 2016). Cuanto menor sea el AICc, mejor es el modelo.
- CV: la extensión del *kernel* es determinada utilizando el método de Validación Cruzada.
- BIC/MDL: usa el Criterio de Información Bayesiano (BIC) para elegir el ancho de banda óptimo.

4.2.2.2 Medidas de bondad de ajuste Como resultado del proceso de modelación en el programa GWR4¹ se obtiene los siguientes indicadores que permiten evaluar el modelo:

- Resumen diagnóstico de un modelo de regresión múltiple estimado por mínimos cuadrados ordinarios (OLS).
- *Bandwidth* o *Neighbors*: indica el ancho de banda o el número de vecinos utilizado por el programa.
- Suma del cuadrado de los residuales: este valor corresponde a la suma de los residuales al cuadrado del modelo.
- R^2 : coeficiente de determinación de Pearson. Es una medida de la bondad de ajuste del modelo. Toma valores entre 0 y 1 indicando la proporción de la variabilidad de los datos que es explicada por el modelo.
- R^2 *Adjusted*: Este valor corresponde al coeficiente de determinación de Pearson ajustado, el cual hace una corrección al valor de R^2 para compensar su aumento espurio debido al simple hecho de tener múltiples variables explicativas. En la GWR el número de grados de libertad (utilizados para calcular el R^2 ajustado) es función del ancho de banda, por este motivo el ajuste puede ser

¹ Programa estadístico de uso libre disponible en <https://gwrtools.github.io/gwr4-downloads.html>

bastante diferente a un modelo global OLS. Debido a esto se recomienda el AICc como una mejor medida para comparar modelos (ESRI, 2016).

- ANOVA: Utilizado para comparar el rendimiento estadístico del modelo global (OLS) con el modelo GWR.

4.2.2.3 Test de variabilidad geográfica En la aplicación GWR4 la variabilidad espacial de los coeficientes se evalúa a través de la comparación entre modelos con parámetro fijo y parámetro variable en el espacio; el método usado por esta prueba es el índice de Información de Akaike corregido (Nakaya, 2016).

El programa ArcMap es un Sistema de Información Geográfica que permite la manipulación, visualización y procesamiento de datos geográficos. Dentro de sus características incluye herramientas para la generación de mapas y la función I de Morán utilizada para detectar patrones de agrupamiento o dispersión.

4.2.2.4 Análisis de residuales El estadístico I de Moran es un estadístico desarrollado por Moran (1950) para evaluar la presencia de autocorrelación espacial en datos medidos en áreas teniendo en cuenta la localización y el valor de los datos. El índice está definido como (ESRI, 2016):

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij} Z_i Z_j}{\sum_{i=1}^n \sum_{j=1}^n W_{ij} \sum_{i=1}^n Z_i^2}$$

Donde Z_i y Z_j son las desviaciones del valor en la posición i y en la posición j respectivamente con respecto al valor de la media (*i.e.* $Z_i = x_i - \bar{X}$). W_{ij} es la matriz de vecindad la cual define si las posiciones i, j son vecinas ($W_{ij} = 1$) o no ($W_{ij} = 0$), n es el número total de datos.

Cuando un valor y el de sus vecinos son ambos mayores o menores que el valor de la media, el producto entre las dos desviaciones tendrá signo positivo, de lo contrario será negativo. Cuanto

mayor sean las desviaciones de la media, mayor será el producto entre las desviaciones. Si los valores tienden a un clúster espacial (*i.e.* valores altos cerca a valores altos o, valores bajos cerca a valores bajos), el índice de Morán será positivo, por el contrario, si los valores altos tienden a estar cerca a valores bajos, el índice será negativo. Además, si los productos positivos entre las desviaciones igualan a los negativos, el índice tenderá a cero. Por lo tanto, el índice de Morán tendrá valores comprendidos entre -1.0 y +1.0. Un valor positivo del índice I de Moran indica una tendencia hacia el *clustering* mientras que un valor negativo indica una tendencia hacia la dispersión (ESRI, 2016).

El índice de Morán también viene acompañado de un Z-Score y un p-valor. La hipótesis nula es H_0 : *Los valores se distribuyen aleatoriamente*

Dependiendo del valor del Z-Score y el p-valor, se puede concluir lo expresado en la Tabla 1 y en la Figura 3.

Tabla 1 Interpretación del índice de Morán según el p-valor y el Z-score

P-Valor $> \alpha$	No se puede rechazar la hipótesis nula, es decir, es probable que los valores se distribuyan debido a un proceso aleatorio.
P-valor $< \alpha$	Se puede rechazar la hipótesis nula, cuando el Z-score positivo, la distribución espacial de los datos está más agrupada que lo esperado en una distribución aleatoria. En caso contrario indica que los valores están más dispersos que lo esperado en un proceso aleatorio

La presencia de correlación espacial en los residuales del modelo GWR puede ser señal de un modelo mal definido por la ausencia de una variable clave en el modelo (ESRI, 2016). Debido a

esto, tras haber realizado un modelo GWR se debe comprobar que los residuales se distribuyen espacialmente de forma aleatoria con el uso del Índice de Moran.

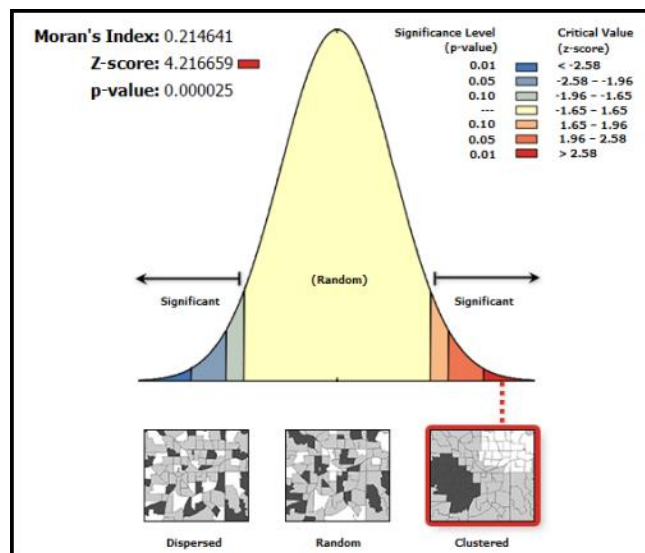


Figura 3 Ejemplo de la salida gráfica del proceso de cálculo del Índice de Morán en ArcMap. Fuente: <http://pro.arcgis.com/es/pro-app/tool-reference/spatial-statistics/h-sa-tool-popup.htm>

5. Metodología

Este estudio es de tipo cuantitativo-correlacional (Dankhe, 1986) en el que se hace uso de información secundaria emitida por diferentes instituciones de orden nacional segregada a nivel municipal.

5.1 Datos

En la Tabla 2 se muestra la descripción de las variables socio-económicas utilizadas para la estimación de los modelos de regresión. Esta información proviene de fuentes secundarias generadas por entidades gubernamentales entre las que se encuentran el Departamento Administrativo Nacional de Estadística (DANE), el Departamento Nacional de Planeación (DPN),

el Ministerio de Educación Nacional y de Migración Colombia y el grupo de información de la Policía Nacional de Colombia. La información fue agrupada a nivel municipal y georreferenciada según el Marco Geoestadístico Nacional del año 2017 suministrado por el DANE.

Tabla 2 Variables evaluadas

Nombre	Variables	Tipo	Transformación	Descripción	Fuente
hurtos	loghurtos	Cuantitativa tipo razón	Logarítmica	Corresponde al logaritmo de la cantidad de hurtos a personas y celulares reportados durante el año 2017	Base de datos del grupo Observatorio del Delito de la Policía Nacional ²
sgppercapita	logsgppercapita	Cuantitativa tipo razón	Logarítmica	Corresponde al presupuesto en millones asignado por el Sistema General de Participaciones	Departamento Nacional de Planeación ³
matri1000p	logmatri1000p	Cuantitativa tipo razón	Logarítmica	Corresponde al logaritmo de la cantidad de personas matriculadas en instituciones de educación superior	Datos del Ministerio de Educación Nacional – Sistema Nacional de Información de la Educación Superior ⁴
Categoría*	categoria	Cualitativa ordinal		Categoría del municipio según la Contaduría General de la Nación.	Contaduría General de la Nación ⁵
Extranjeros	Totalextran	Cuantitativa tipo conteo		Corresponde a la cantidad de extranjeros hospedados en los municipios	Registro de ingreso de extranjeros según ciudad de hospedaje, Ministerio de relaciones exteriores de Colombia ⁶

² <https://www.policia.gov.co/grupo-informaci%C3%B3n-criminalidad/estadistica-delictiva>

³ https://sicodis.dnp.gov.co/ReportesSGP/SGP_ReportePorEntidadyFuente.aspx

⁴ <https://www.mineducacion.gov.co/sistemasinfo/Informacion-a-la-mano/212400:Estadisticas>

⁵

http://www.contaduria.gov.co/wps/wcm/connect/759e1fc3-bd8b-4d63-a95b-f7c37d509140/CT01+-+Categorizacion+2018+-+DNP.xlsx?MOD=AJPERES&CONVERT_TO=url&CACHEID=759e1fc3-bd8b-4d63-a95b-f7c37d509140

⁶ http://migracioncolombia.gov.co/index.php/es/?option=com_content&view=article&id=718

Nombre	Variables	Tipo	Transformación	Descripción	Fuente
Cobertura neta total	cober	Cuantitativa tipo razón	Logarítmica	Porcentaje de estudiantes matriculados en el sistema educativo; sin contar los que están en extraedad (por encima de la edad correspondiente para cada grado)	Ministerio de Educación ⁷
Cobertura neta secundaria	COBERTURA_NETA	Cuantitativa tipo razón		Porcentaje de estudiantes matriculados en la educación secundaria; sin contar los que están en extraedad (por encima de la edad correspondiente para cada grado). Corresponde al porcentaje de cobertura neta en educación preescolar, básica, media y secundaria	Ministerio de Educación ⁸
Cobertura neta educación básica	CobNeta	Cuantitativa tipo razón			Ministerio de Educación ⁹
Población total	PobTotal	Cuantitativa tipo razón		Corresponde a la población total del municipio	DANE. Proyecciones de la población total por municipio para el año 2017 ¹⁰
Población rural	PobRural	Cuantitativa tipo razón		Corresponde a la cantidad de residentes en el sector rural del municipio	DANE. Proyecciones de la población total por municipio para el año 2017
Departamento	Depto	Cualitativa nominal		Corresponde al departamento en el cual se encuentra el municipio	Marco Geoestadístico Nacional del 2017 ¹¹
Región	REGION	Cualitativa nominal		Corresponde a la región geográfica	Marco Geoestadístico

⁷ <https://www.dane.gov.co/Educaci-n/ESTADISTICAS-EN-EDUCACION-BASICA-POR-MUNICIPIO/nudc-7mev>

⁸ <https://www.dane.gov.co/Educaci-n/ESTADISTICAS-EN-EDUCACION-BASICA-POR-MUNICIPIO/nudc-7mev>

¹⁰ https://www.dane.gov.co/files/investigaciones/poblacion/proyepobla06_20/Municipal_area_1985-2020.xls

¹¹ <https://geoportal.dane.gov.co/v2/?page=elementoDescargaMGN>

Nombre	Variables	Tipo	Transformación	Descripción	Fuente
				en la cual se encuentra el municipio	Nacional del 2017 ¹²

*La variable categoría corresponde a la clasificación otorgada por la Contaduría General de La República teniendo en cuenta la normatividad vigente definida por el artículo primero de la Ley 617 de 2000 y el artículo séptimo de la Ley 1551 de 2012. La categorización de los municipios se realiza bajo los siguientes parámetros: los límites definidos de población en ingresos corrientes de libre destinación (en adelante ICLD) y el valor máximo definido de los gastos de funcionamiento como proporción a los ICLD para las diferentes entidades territoriales, importancia económica y situación geográfica según lo estipulado en la Tabla 3.

Tabla 3 Criterios para categorizar a los municipios según el artículo séptimo de la Ley 1551 de 2012

Categoría	Población	ICLD en (SMLMV)*	Importancia Económica
Especial (0)	≥ 500.001	> 400.000	Grado 1
Primera (1)	100.001 - 500.000	100.000 - 400.000	Grado 2
Segunda (2)	50.001 - 100.000	50.000 - 100.000	Grado 3
Tercera (3)	30.001 - 50.000	30.000 - 50.000	Grado 4
Cuarta (4)	20.001 - 30.000	25.000 - 30.000	Grado 5
Quinta (5)	10.001 - 20.000	15.000 - 25.000	Grado 6
Sexta (6)	≤ 10.000	hasta 15.000	Grado 7

*Salario Mínimo Legal Mensual Vigente, Fuente: Contaduría General de la República

5.2 Procedimientos

A continuación, se presentan los pasos que se siguieron para construir y elegir los modelos finales.

En primer lugar, se realizó un análisis exploratorio de los datos, a partir de la evaluación de la correlación entre las variables, con el objetivo de identificar la posible presencia de relaciones lineales entre la variable predicha y las predictoras, al igual que la posible multicolinealidad entre predictoras.

Una vez realizado esto, se modelaron diferentes regresiones lineales múltiples combinando las distintas variables. Este proceso se realizó mediante la función *step* de la librería *stats* del programa

¹² <https://geoportal.dane.gov.co/v2/?page=elementoDescargaMGN>

R (R Core Team, 2018). Adicionalmente se utilizó un proceso empírico de prueba y error, con el cual se definió el modelo de regresión lineal múltiple que satisface los supuestos de normalidad, homocedasticidad y no multicolinealidad entre variables predictoras. Para complementar, se identificaron los valores atípicos en los residuales del modelo final y se comprobó que corresponden a información real característica de los municipios, por lo cual, no fueron eliminados. Es importante mencionar que los modelos fueron elaborados con los datos de 532 municipios los cuales contaban con la información completa en todas las variables.

Para satisfacer la condición de normalidad y homocedasticidad, las variables: cantidad de matriculados en IES y número de extranjeros hospedados por municipio, fueron recalculadas tomando en cuenta la población total de cada municipio (i) mediante la ecuación (9), con el propósito de obtener una tasa que facilite la comparación entre municipios.

$$x'_i = \frac{x_i}{Población_i} \times 1000 \quad (9)$$

Adicionalmente, se estandarizaron las variables usadas en los modelos mediante la ecuación (10).

$$x'_i = \frac{x_i - \bar{X}}{\sigma} \quad (10)$$

Finalmente, se aplicó una transformación logarítmica a las variables hurta personas y celulares, población total, población rural, NBI, matriculados en IES por cada 1000 personas y SGP percapita; con la intención de mitigar el sesgo a la derecha presente en las variables, con lo cual se logró cumplir con los supuestos requeridos en los modelos lineales. Una vez obtenido el modelo final de regresión lineal múltiple, se visualizaron geográficamente los residuales en ArcMap y se aplicó el índice de Morán (Moran, 1950) que evalúa la autocorrelación espacial e indica la conveniencia de realizar la regresión geográficamente ponderada.

En el programa ArcMap se georreferenció la base de datos con la información geográfica municipal obtenida del Marco Geoestadístico Nacional, se calcularon los centroides de los municipios y se convirtió la base de datos a formato *.dbase*; información que fue procesada en el programa estadístico GWR4 (Nakaya, 2015) en el que se obtuvo el modelo espacial; los resultados obtenidos en este proceso fueron llevados a ArcMap para su visualización geográfica.

El análisis de normalidad en los residuos se realizó mediante el test de Shapiro-Wilk (Shapiro y Wilk, 1965) y la homocedasticidad a partir del test Breusch-Pagan (Breusch y Pagan, 1979).

6. Resultados

6.1 Análisis exploratorio de los datos

Tal como se dijo anteriormente, en primer lugar, se realizó un análisis de correlación entre las diferentes variables estudiadas. En este sentido, en la Figura 4 se observa que la única variable altamente correlacionada con la variable dependiente (hurto) es la población total (PobTotal), las demás variables presentan correlaciones por debajo de 0.5. Tampoco existen correlaciones importantes entre las variables predictoras. La relación entre el presupuesto per cápita asignado en el Sistema General de Participaciones (sgp_percapita) y el hurto, aunque es muy baja, presenta el comportamiento esperado: a mayor presupuesto por persona, menor el número de hurtos.

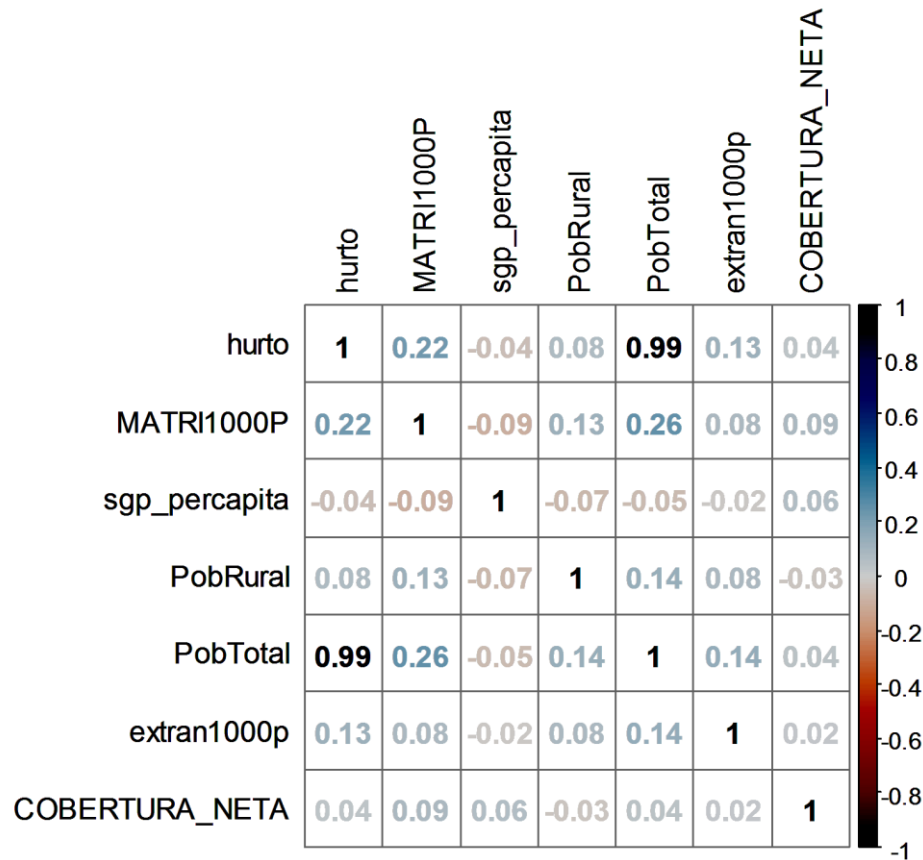


Figura 4 Matriz de correlación de las variables

En la Figura 5 se observa que existe variabilidad espacial en la cantidad de hurtos a personas y celulares (loghurtos). La mayor cantidad de delitos registrados se localizan principalmente en las ciudades capitales y municipios de categoría especial y categoría uno (Figura 11). Es importante resaltar que existen municipios con baja población que presentan altos números de hurtos registrados como el caso de algunos municipios localizados en la región de los llanos orientales.

Al realizar el análisis descriptivo de la variable hurto se observa que en promedio se registraron en el 2017 en los 532 municipios un promedio de 2.1 hurtos por cada mil habitantes, el valor mínimo fue de 0.03 hurtos por cada mil habitantes presentados en el municipio de Moñitos del departamento de Córdoba y el máximo fue 15.6 registrado en Pasto en el departamento de Nariño. La desviación estándar de los hurtos por cada mil personas es de 2.52.

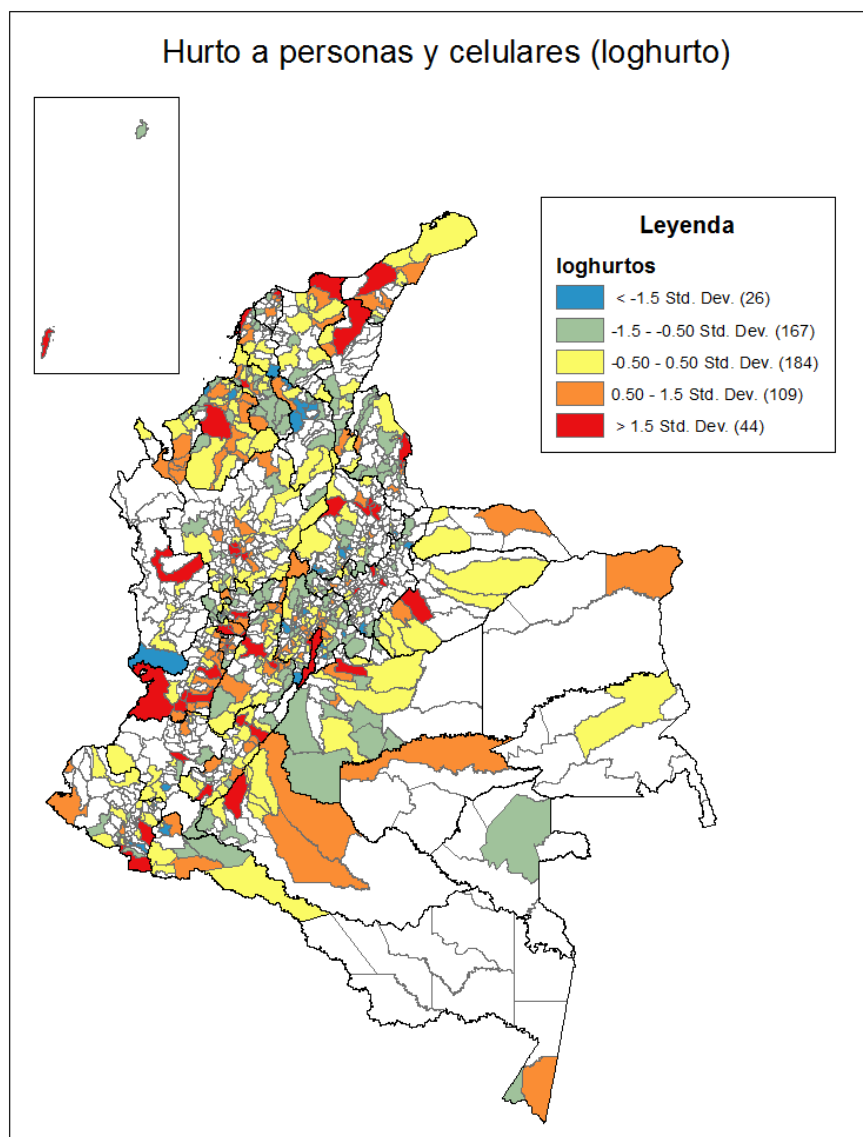


Figura 5 Mapa de distribución de la variable loghurto.

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

A partir de la Figura 6 se puede evidenciar que existe una importante dispersión en los recursos económicos per cápita asignados por el Sistema General de Participaciones. En promedio el presupuesto per cápita asignado por el SGP es de 530.090 pesos, el mínimo valor es de 136.461 asignado al municipio de Sopó en el departamento de Cundinamarca, el valor máximo es de 2'059.951 pesos asignados al municipio de San Miguel en el departamento de Santander. La desviación estándar de la variable sgppercapita es de 232.439.

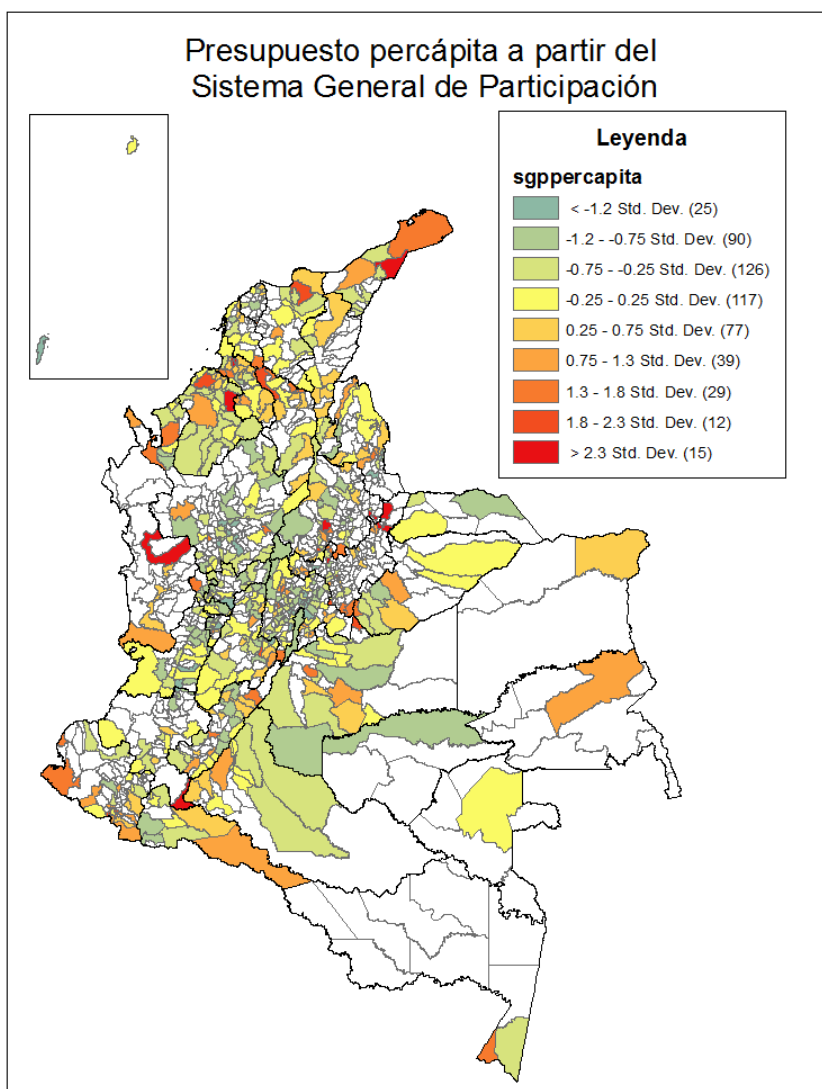


Figura 6 Mapa de distribución de la variable sgppercapita

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

En la Figura 7 se observa que las principales ciudades del País (Bogotá, Cali, Barranquilla, Medellín, Bucaramanga y Cartagena) se concentra el mayor número de estudiantes de educación superior. El valor promedio de matriculados en IES es de 28.1 por cada mil habitantes, el valor mínimo se presenta en el municipio de Marinilla en el departamento de Antioquia con 0.018, el máximo es de 535 en el municipio de Nilo en Cundinamarca.

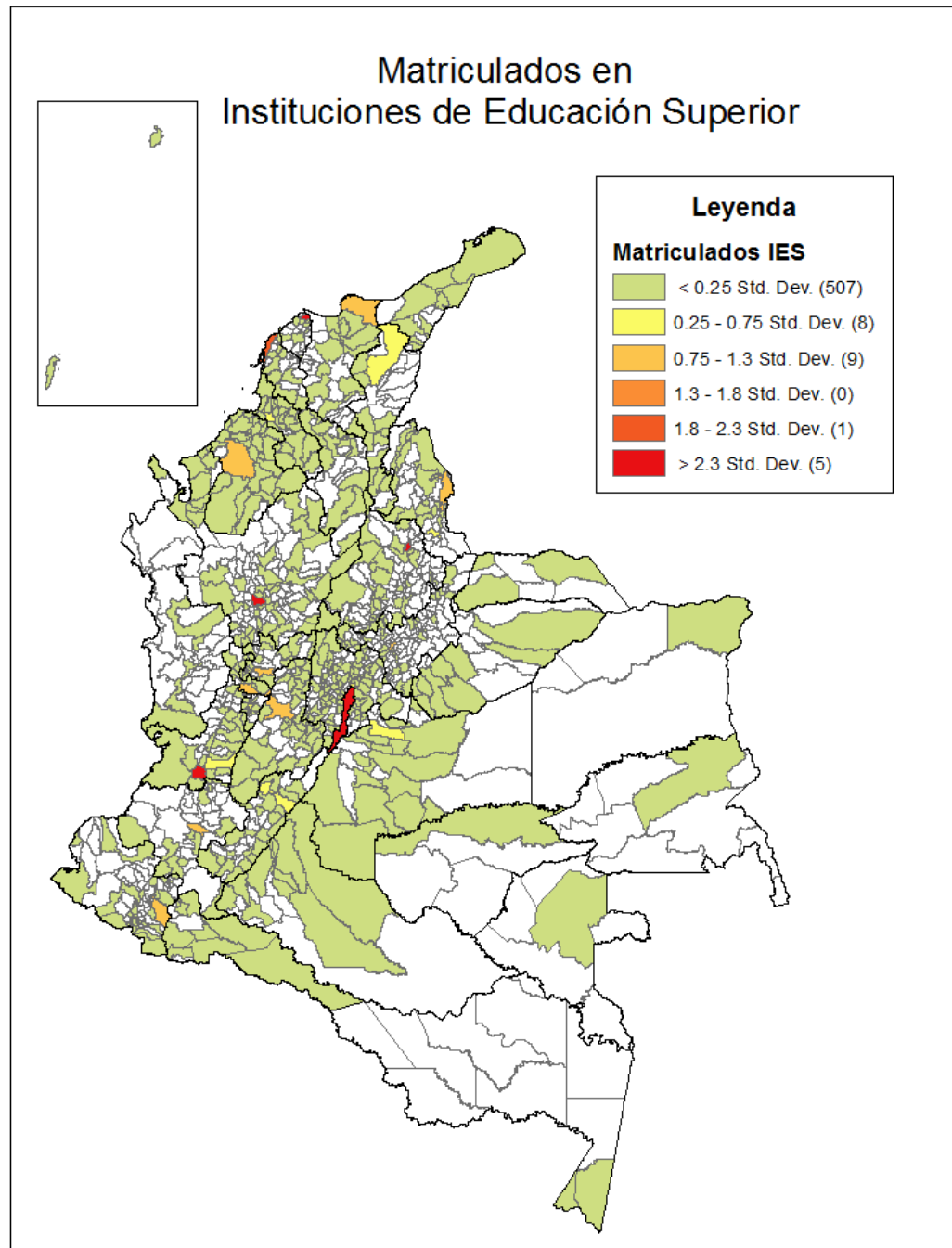


Figura 7 Mapa de distribución de la variable matriculados en Instituciones de Educación Superior
 Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

El número de municipios en cada categoría y su localización se presenta en la Figura 8.

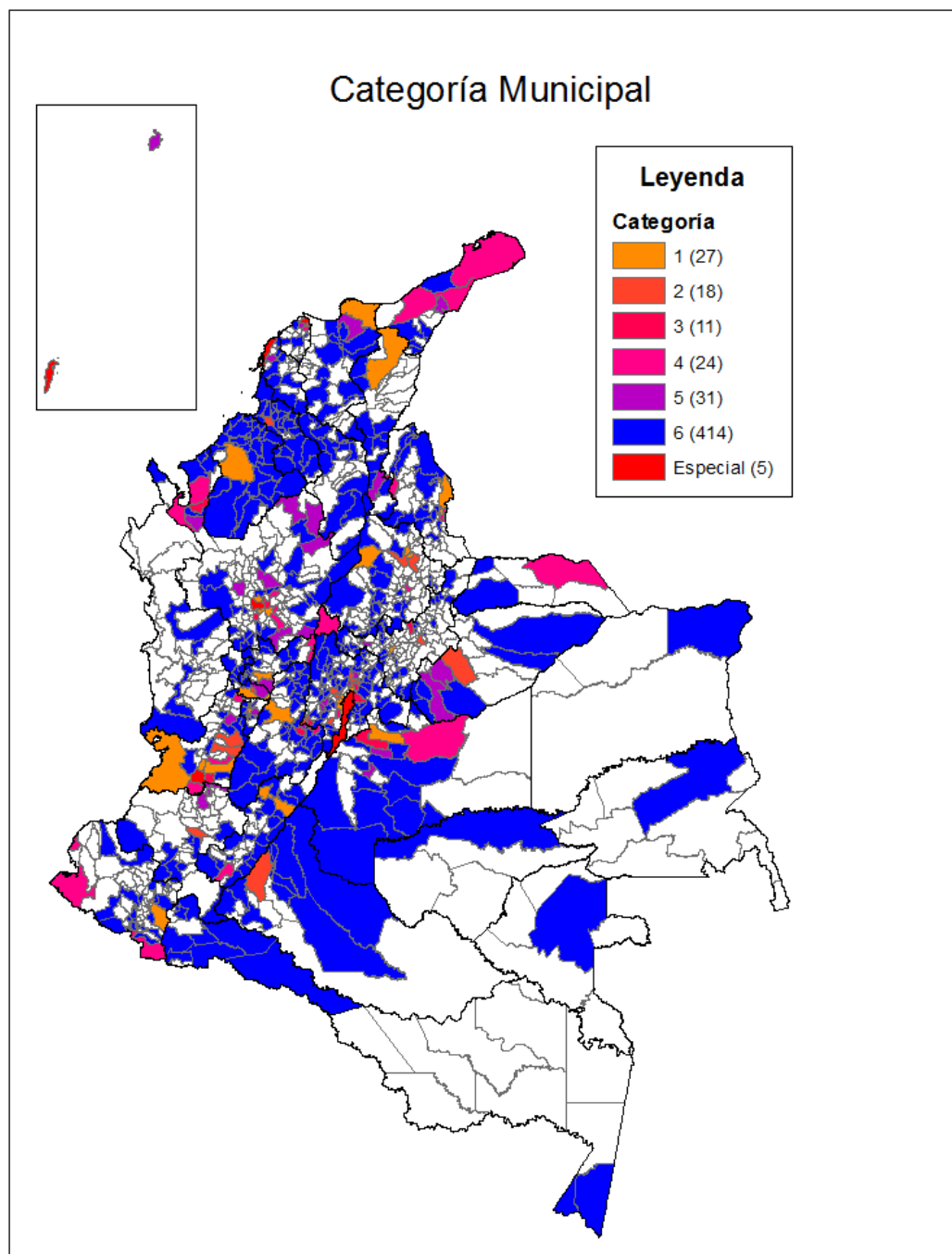


Figura 8 Mapa de distribución de la variable categoría

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

6.2 Modelos lineales múltiples evaluados

En la Tabla 4 se presentan los modelos lineales generados con el resultado las pruebas de normalidad Shapiro-Wilk (Shapiro y Wilk, 1965), homocedasticidad Breusch-Pagan (Breusch y Pagan, 1979) y los valores AIC y el R^2 ajustado en porcentaje; en rojo se señalan las variables que resultaron ser no significativas. Nótese que entre los modelos evaluados solo “modelosteplog8” y “modelosteplog9” cumplen con los supuestos de normalidad y homocedasticidad.

Al compararlos, por principio de parsimonia, se escogió el “modelosteplog9” como final, debido a que la inclusión de la variable PobRural puede mostrar una relación espuria, ya que muestra una relación contraria a los esperado y, adicionalmente su exclusión no afecta considerablemente el AIC ni el valor del R^2 ajustado.

Tabla 4 Modelos lineales evaluados

Modelo	Fórmula	AIC	R ² (%)	Normalidad residual	Varianza Constante
modelopoblación	loghurto~PobTotal	2202.5	16	No	No
modelosaturado	loghurto ~ sgppercapitamillones +Matri1000P +extran1000p+ COBERTURA_NETA + NBI_Prop +Categoria +PobTotal+ PobRural+ Depto+ REGION	1581	75.85	No	No
modelostep	loghurto~ sgppercapitamillones + Matri1000P + extran1000p + COBERTURA_NETA + NBI_Prop + Categoria + PobTotal + PobRural + Depto	1581.8	75.85	No	No
modelostep2	loghurto~ sgppercapitamillones + Matri1000P + extran1000p + COBERTURA_NETA + NBI_Prop + Categoria + PobTotal + PobRural	1684.8	68.9	Si	No
modelostep3	loghurto~ sgppercapitamillones + Matri1000P + COBERTURA_NETA + NBI_Prop + Categoria + PobTotal + PobRural	1683.7 1	69	Si	No
modelopob2	loghurto~log(PobTotal)	1558.1	75	No	Si

Modelo	Fórmula	AIC	R ² (%)	Normalidad residual	Varianza Constante
modelosaturado 2	loghurto ~ logsgppercapita + logmatri1000p +extran1000p+ logcober + log(NBI_Prop) +Categoria +log(PobTotal)+ log(PobRural)+ Depto+ REGION	1275	86.3	No	No
modelosteplog	loghurto ~ logmatri1000p + extran1000p + logcober + log(NBI_Prop) + log(PobTotal) + log(PobRural) + Depto	1272	86.3	No	No
modelosteplog2	loghurto ~ logmatri1000p + extran1000p + log(NBI_Prop) + log(PobTotal) + log(PobRural) + Depto	1273.4	86.29	No	No
modelosteplog3	loghurto ~ logmatri1000p + extran1000p + log(NBI_Prop) + log(PobTotal) + log(PobRural)	1330	83.83	No	No
modelosteplog4	loghurto ~ logmatri1000p + extran1000p + log(NBI_Prop) + log(PobRural)	1826	58.86	No	No
modelosteplog5	loghurto ~ logmatri1000p + extran1000p + log(NBI_Prop) + log(PobRural)+categoria	1688	68.6	Si	No
modelosteplog6	loghurto ~ logmatri1000p + log(NBI_Prop) + log(PobRural)+Categoria	1687	68.63	Si	No
modelosteplog7	loghurto ~ logmatri1000p + log(NBI_Prop) + log(PobRural)+Categoria+logsgppercapita	1648.7 7	70.86	Si	Si
modelosteplog8	loghurto~logmatri1000p+log(PobRural)+Categoria+logs gppercapita	1647.8	70.86	Si	Si
modelosteplog9	loghurto~logsgppercapita+logmatri1000p +Categoria	1671	69.5	Si	Si
modelosteplog1 0	loghurto~ logsgppercapita + logcober + logmatri + REGION	1750.9	64.06	Si	No
modeloestan	hurtoestan ~ coberestan + nbiestan + sgpperestan + matriestan + extranestan	242.49	90.9	No	No
modeloestan1	hurtoestan ~ nbiestan + matriestan + extranestan	238.7	90.95	No	No
modeloestan2	hurtoestan ~ nbiestan + matriestan + extranestan+Categoria	201.15	91.66	No	No
modeloestan3	hurtoestan ~ nbiestan + coberestan + matriestan + Categoria	1262.7 8	38.6	No	No
modeloestan4	hurtoestan ~ extranestan+Categoria	199.58	91.65	No	No
modeloestan5	hurtoestan~ nbiestan + coberestan +matriestan+ sgppercapit amillones	1491	4	No	No

Modelo	Fórmula	AIC	R ² (%)	Normalidad residual	Varianza Constante
modeloestan5	hurtoestan~nbiestan+coberestan+matriestan+sgppercapit amillones+Categoría	1264.6	38	No	No
modeloestan5	hurtoestan ~ nbiestan + coberestan + matriestan	1489	5	No	No

El modelo final (loghurto~logmatri1000p + logsgppercapita + Categoría) (denominado modelostep9 en la Tabla 4), logra explicar un 65.9% de la variabilidad en el hurto a partir de la tasa de matriculados en educación superior por cada mil personas, el presupuesto per cápita asignado en el Sistema General de Participaciones y la categoría del municipio. A continuación, se presenta el análisis y diagnóstico del modelo.

6.3 Análisis del modelo de regresión lineal múltiple final

En la Figura 9, según el valor del *F-Statistics*, con un nivel de significancia del 5% se concluye que el modelo final está bien especificado, rechazando la hipótesis nula ($H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$). Al observar el valor *t* y su correspondiente *p*-valor se confirma que todas variables incluidas en el modelo son significativas a un nivel de significancia del 5%.

El modelo final se puede expresar como:

$$\begin{aligned} \loghurto = & 24.94971 - (1.30122)\logsgppercapita + (0.21779)\logmatri1000p \\ & - (1.30239)Categoría1 - (2.30343)Categoría2 - (3.63049)Categoría3 \\ & - (3.19047)Categoría4 - (4.26926)Categoría5 - (5.25923)Categoría6 \end{aligned}$$

* Teniendo como grupo de referencia los municipios con categoría especial

```

> modelsteplog9=lm(formula = loghurto ~ logmatril000p +logsgppercapita+Categoria, data = datos1)
> summary(modelsteplog9)

Call:
lm(formula = loghurto ~ logmatril000p + logsgppercapita + Categoria,
    data = datos1)

Residuals:
    Min       1Q   Median       3Q      Max
-3.3238 -0.7662  0.0124  0.7889  4.0965

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.94971    1.73823   14.354 < 2e-16 ***
logmatril000p  0.21779    0.02217    9.822 < 2e-16 ***
logsgppercapita -1.30122    0.12931  -10.063 < 2e-16 ***
Categorial    -1.30239    0.52063   -2.502  0.0127 *
Categoria2     -2.30343    0.54364   -4.237 2.68e-05 ***
Categoria3     -3.63049    0.59066   -6.146 1.57e-09 ***
Categoria4     -3.19047    0.52899   -6.031 3.08e-09 ***
Categoria5     -4.26926    0.51773   -8.246 1.33e-15 ***
Categoria6     -5.25923    0.48621  -10.817 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.152 on 523 degrees of freedom
Multiple R-squared:  0.6996,    Adjusted R-squared:  0.695
F-statistic: 152.2 on 8 and 523 DF,  p-value: < 2.2e-16

```

Figura 9 Resumen del modelo final

La interpretación del modelo anterior indica que, permaneciendo las demás variables constantes, un aumento del 1% en la tasa de matriculados por cada mil personas, en promedio se espera un aumento del 0.21% en la cantidad de hurtos. Esta tendencia se puede observar en la Figura 10.

Al permanecer las demás variables constantes, un aumento del 1% en la inversión del Sistema General de Participaciones por habitante, se espera una reducción del 1.3% en promedio en la cantidad de hurtos en el país; relación observada en la Figura 10.

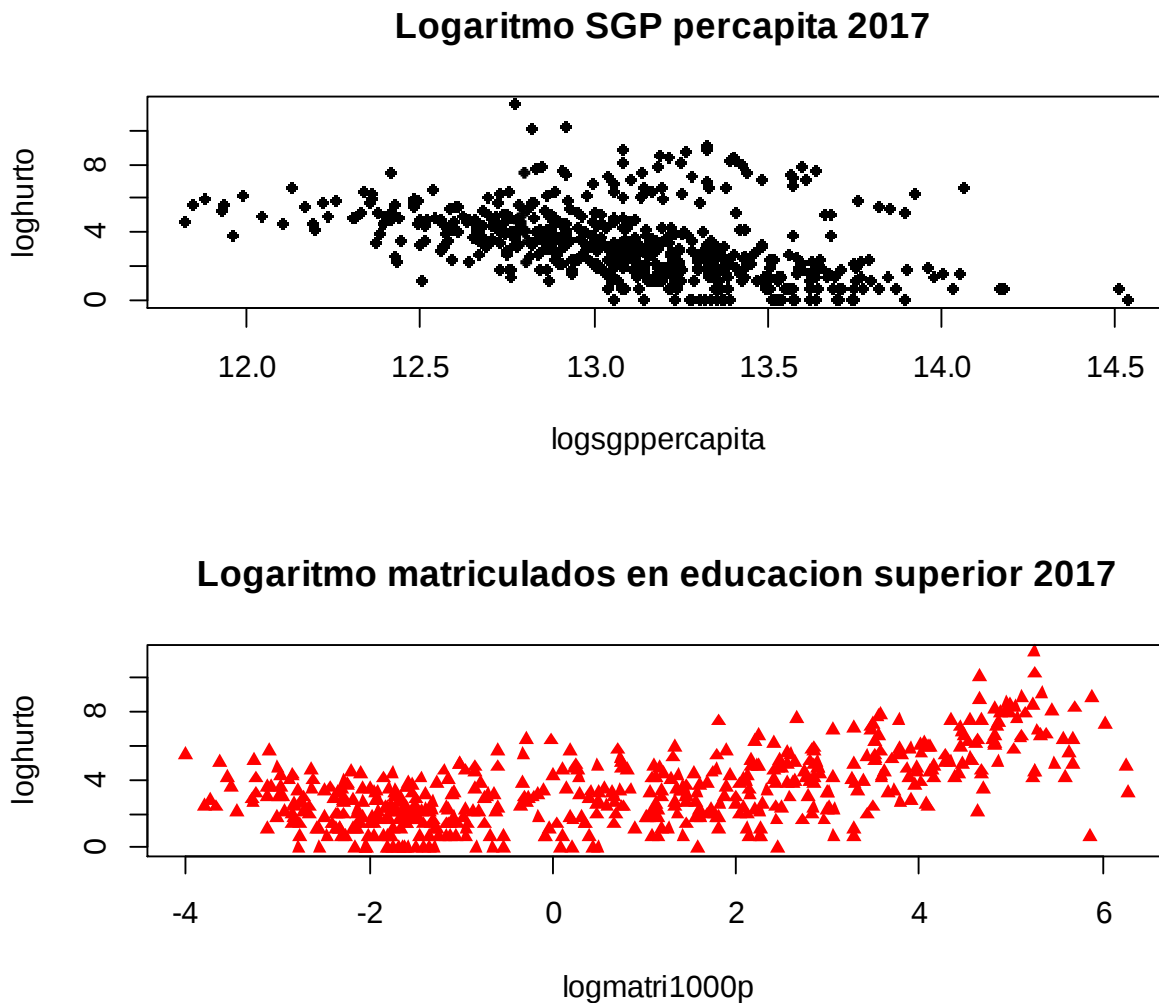


Figura 10 Diagramas de dispersión loghurtos vs lossgppercapita y loghurtos vs logmatri1000p

Así mismo, el modelo obtenido indica que los municipios clasificados en la categoría especial presentan en promedio un valor mayor en la variable loghurto en comparación con los municipios de las otras categorías; de esta manera, los municipios de la categoría 1 tienen en promedio una disminución promedio del 1.3% en la variable loghurto, los municipios con categoría 2 un 2.3% menos hurtos, los municipios de la categoría 4 presentan en promedio 3.19% menos, la categoría 5 un 4.26% menos y los municipios de sexta categoría ostentan en promedio 5.2% menos hurtos; diferencias observadas en la Figura 11.

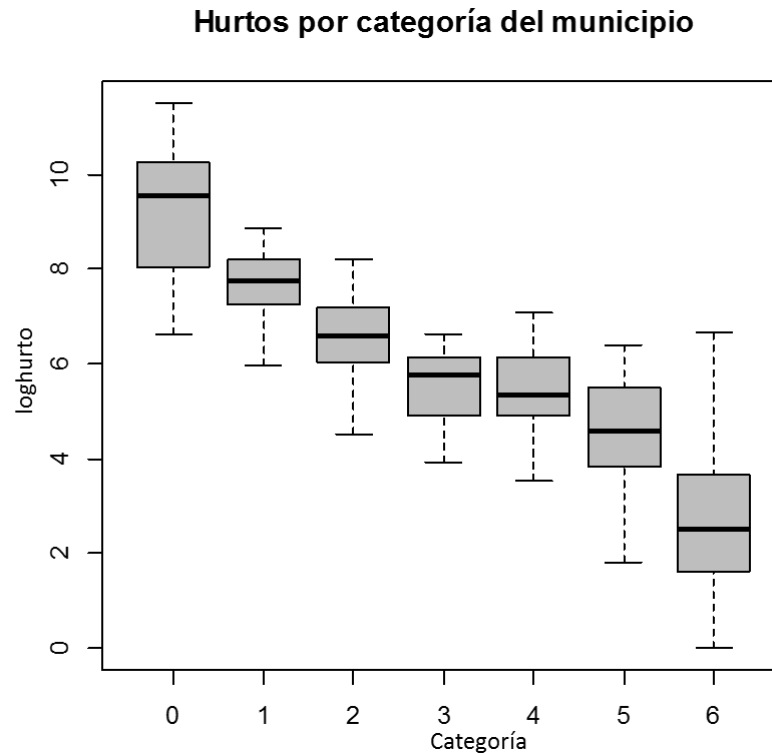


Figura 11 Diagramas de caja de la variable loghurto vs categoria

6.3.1 Diagnóstico del modelo de regresión lineal múltiple Para la validar el modelo lineal obtenido se realizaron múltiples pruebas estadísticas para garantizar el cumplimiento de los supuestos de normalidad, homocedasticidad, no autocorrelación en los residuales, ausencia de multicolinealidad perfecta entre las variables explicativas. Dado el objetivo del presente documento, se evaluó la presencia de autocorrelación espacial mediante el índice de Morán (Moran, 1950).

6.3.1.1 Normalidad en los residuales A partir del análisis grafico por medio del *qqplot* (Figura 12) y el *test* de Shapiro-Wilk (Shapiro y Wilk, 1965) con un P-valor de 0.6874, no se rechaza la hipótesis nula H_0 : *los residuales tienen una distribución normal*, a un nivel de significancia del 5%.

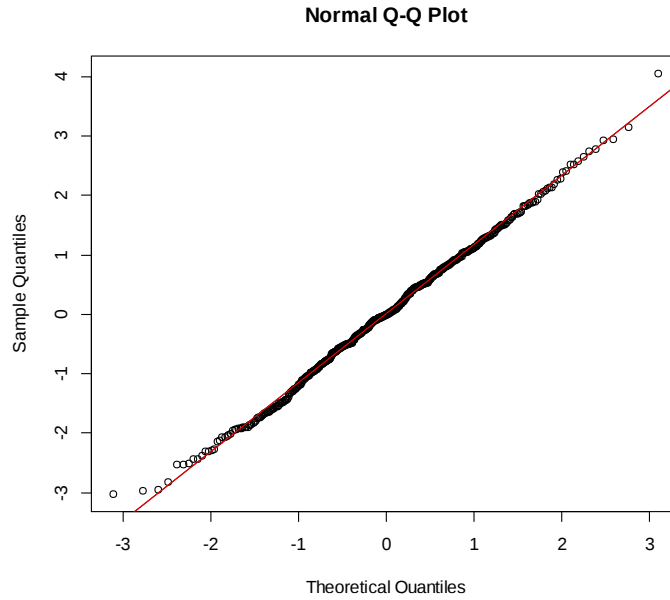


Figura 12 qqplot de los residuales del modelo lineal múltiple

6.3.1.2 Homocedasticidad Para evaluar la homocedasticidad se utilizó el plot de los residuales vs cada una de las variables numéricas predictoras y el *test* de Breusch-Pagan (Breusch y Pagan, 1979) cuya hipótesis nula es H_0 : *los datos son homocedasticos*. Con un P-valor de 0.2982 no se rechaza la hipótesis nula de homocedasticidad de los datos a un nivel de significancia del 5%.

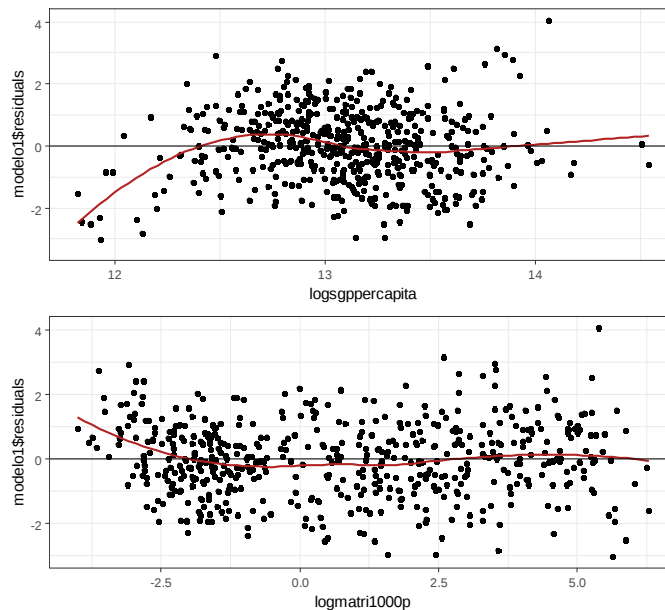


Figura 13 Plot de los residuales vs los valores predichos

6.3.1.3 Evaluación de la multicolinealidad Al observar la matriz de correlación entre parejas de regresores se observan correlaciones bajas entre predictoras. Además, valores del Factor de Inflación de Varianza (VIF) menores a 10 (Figura 14) indican que no existe multicolinealidad en el modelo.

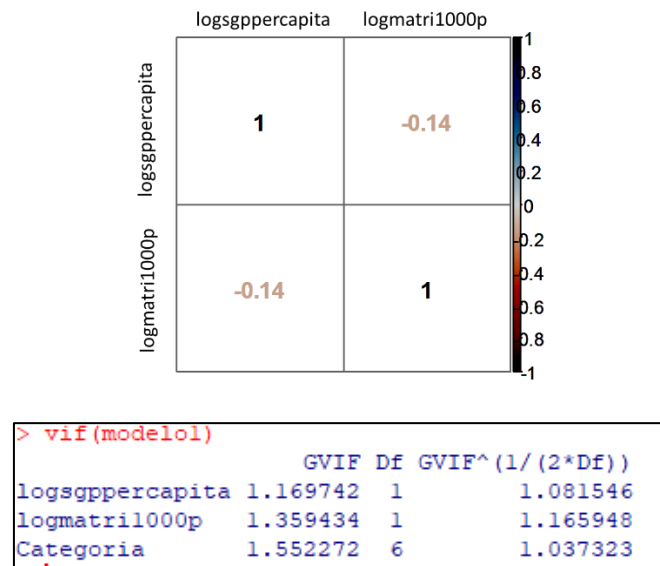


Figura 14 Resultado del Factor de Inflación de Varianza

6.3.1.4 Prueba de autocorrelación en los residuales Para evaluar la autocorrelación de los residuales se utilizó el test de Durbin-Watson (1950, 1951, 1971) que tiene como hipótesis nula de no autocorrelación. Con un P-valor de 0.001605 se concluye que los residuales presentan autocorrelación.

6.3.1.5 Prueba de autocorrelación espacial Teniendo en cuenta el resultado del test de autocorrelación de Durbin-Watson aplicado a los residuales del modelo, se decidió evaluar la autocorrelación espacial mediante el índice de Morán (1950). El resultado obtenido (Figura 15) indica que existe autocorrelación espacial en los residuales del modelo ya que hay menos del 1% de probabilidad que el agrupamiento espacial de los residuales sea aleatorio, rechazando la hipótesis nula H_0 : los datos se distribuyen aleatoriamente

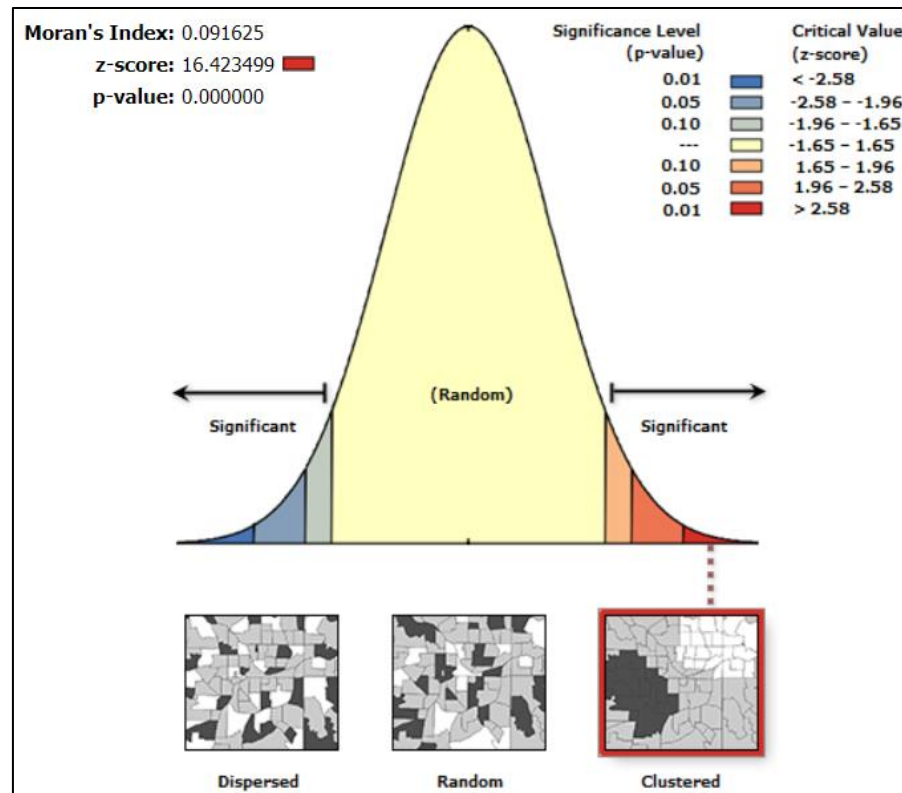


Figura 15 Resultado del índice de Morán

En la Figura 16 se observa la distribución espacial de los residuales, es de resaltar los *cluster* de altos residuales presentes en la región caribe y en el departamento de Caquetá; resaltados con circunferencias de color negro.

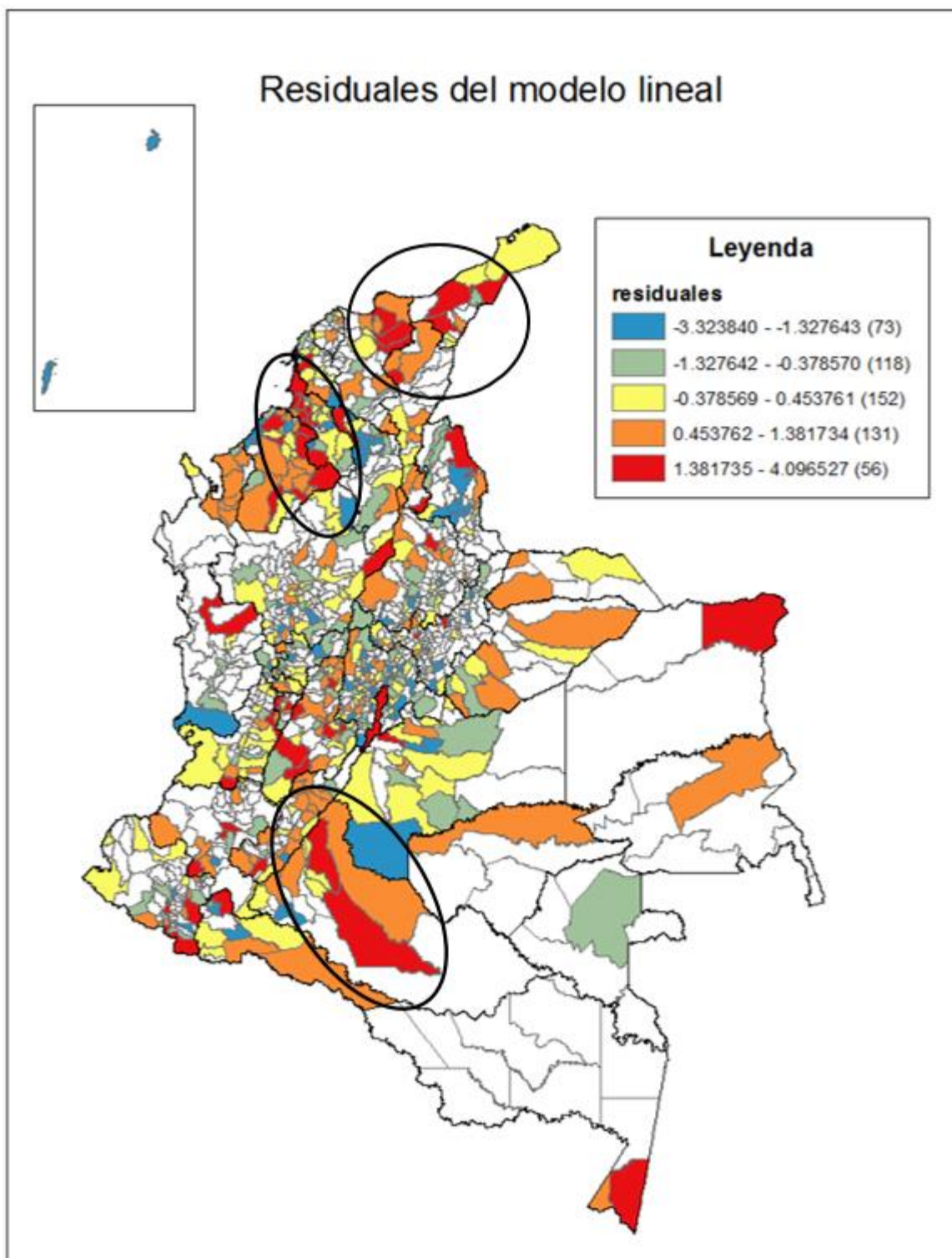


Figura 16 Mapa de distribución de los residuales del modelo lineal múltiple.
Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

6.4 Resultados del modelo GWR

A continuación, se presenta los mapas de distribución de los coeficientes estimados por el modelo GWR.

6.4.1 Intercepto del modelo GWR En la Figura 17 se observa la heterogeneidad espacial en el intercepto, el cual presenta una desviación estándar de 2.680319, considerablemente mayor que los coeficientes de las variables logaritmo del presupuesto per cápita asignado por el Sistema General de Participaciones y el Logaritmo de la tasa de matriculados en educación superior por cada 1000 personas.

Los valores más bajos se encuentran agrupados hacia el sur y suroeste del país, siendo los municipios de Puerto Nariño y Leticia en el Amazonas y, Puerto Leguízamo en el Putumayo, los de menor valor.

Por otro lado, los valores más altos se concentran en el nororiente del país, siendo los municipios de Sardinata y Ocaña en el departamento de Norte de Santander y, San Alberto en el Cesar, los que presentan el mayor valor.

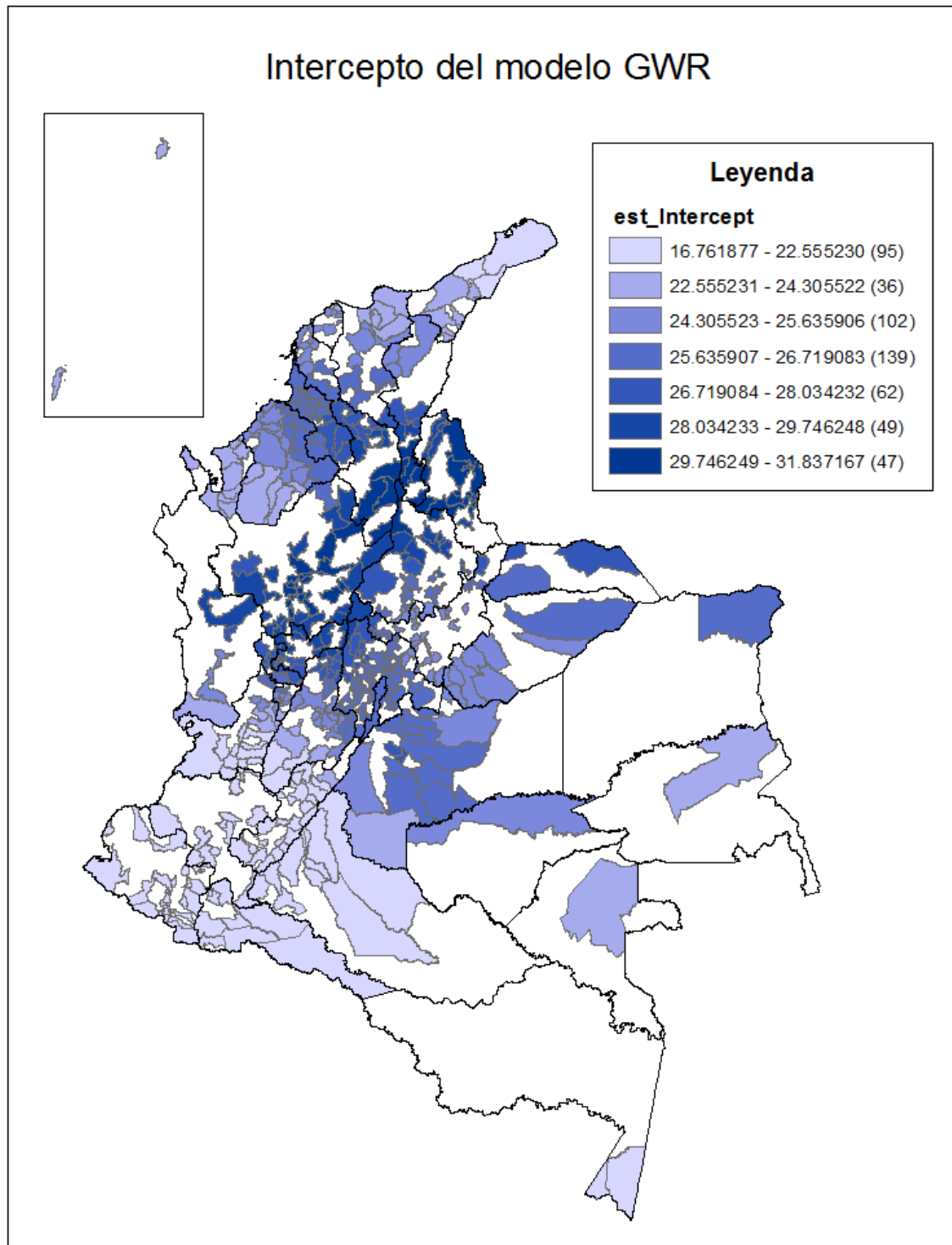


Figura 17 Mapa de distribución del intercepto del modelo GWR

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

6.4.2 Coeficiente variable logsgpppercapita En la Figura 18 se observa que el coeficiente de la variable presupuesto per cápita asignado por el Sistema General de Participaciones, transformada logarítmicamente presenta una correlación negativa con el hurto en todo el país, por lo cual tiene un comportamiento atenuante en relación al hurto, sin embargo, esta relación es variable espacialmente. Los municipios con mayor valor se concentran hacia el sur y suroeste del país, siendo los municipios de Puerto Nariño y Leticia en el Amazonas y, Puerto Leguízamo en el Putumayo, los de mayor valor.

Los valores más altos se concentran en el nororiente del país, siendo los municipios de Sardinata y Ocaña en el departamento de Norte de Santander y, San Alberto en el Cesar, los que presentan el mayor valor. Esto indica que en estas regiones esta variable influye en menor medida sobre los hurtos.

Por otro lado, los valores más bajos se concentran hacia la zona del Magdalena medio y Norte de Santander, siendo los municipios de Sardinata y Gramalote en Norte de Santander y, san Alberto en el Cesar los municipios con los valores más bajos.

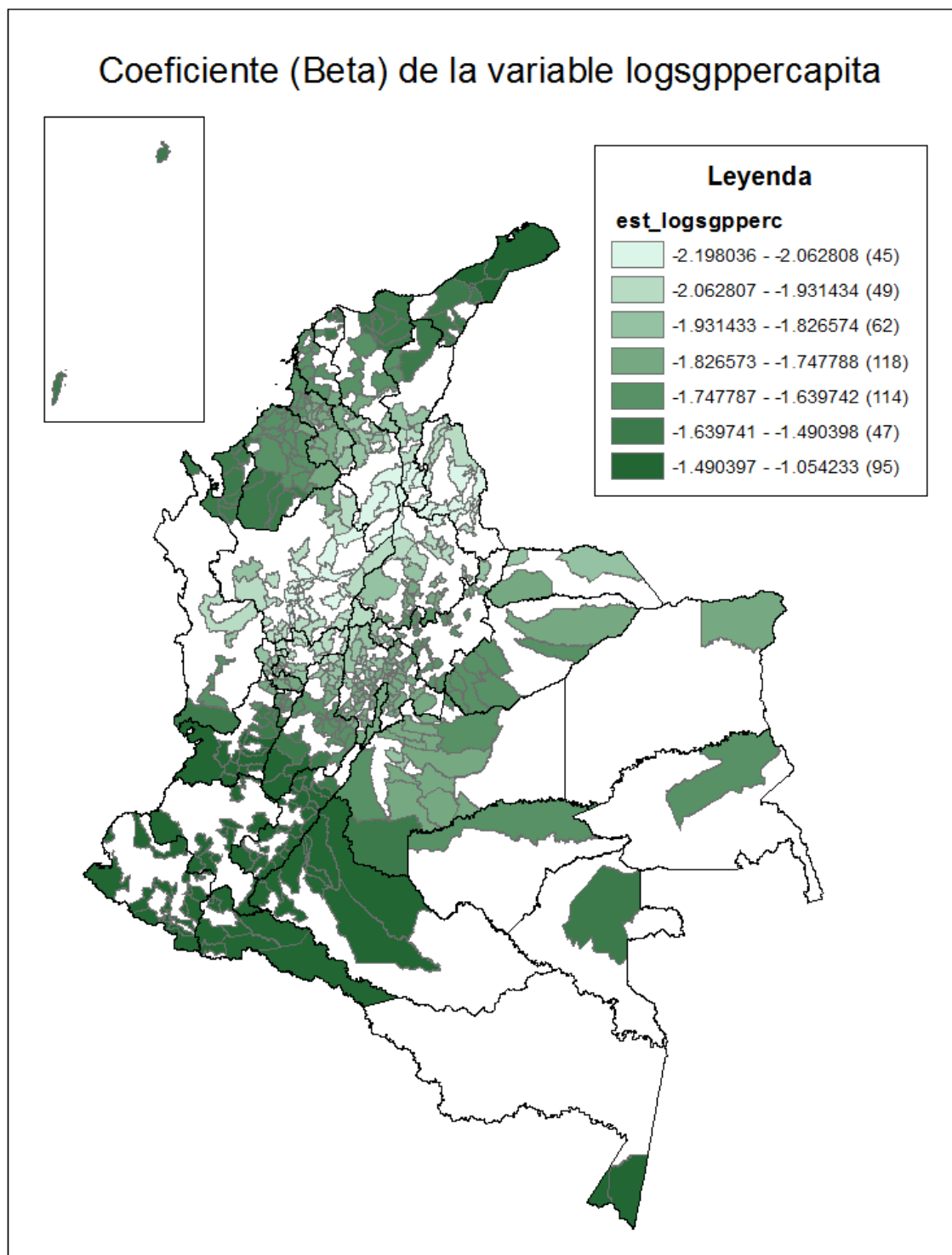


Figura 18 Mapa de distribución del coeficiente de la variable logsgppercapita

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

6.4.3 Coeficientes variable logmatri1000p En la Figura 19 se observa que la tasa de matriculados en instituciones de educación superior por cada mil personas presenta una relación directa con el hurto; sin embargo, esta relación varía espacialmente en menor medida en comparación con la variable logsgppercapita y el intercepto del modelo.

Los valores más altos se concentran en el norte y suroeste del país, siendo los municipios de Tumaco y El Charco en el departamento de Nariño y, Guapi en el Cauca los que presentan los mayores coeficientes.

Por otra parte, los menores valores se concentran principalmente en los departamentos de Antioquia y los Santanderes; siendo los municipios de Amalfi, Yolombó y Yondó en Antioquia los que presentan los menores coeficientes.

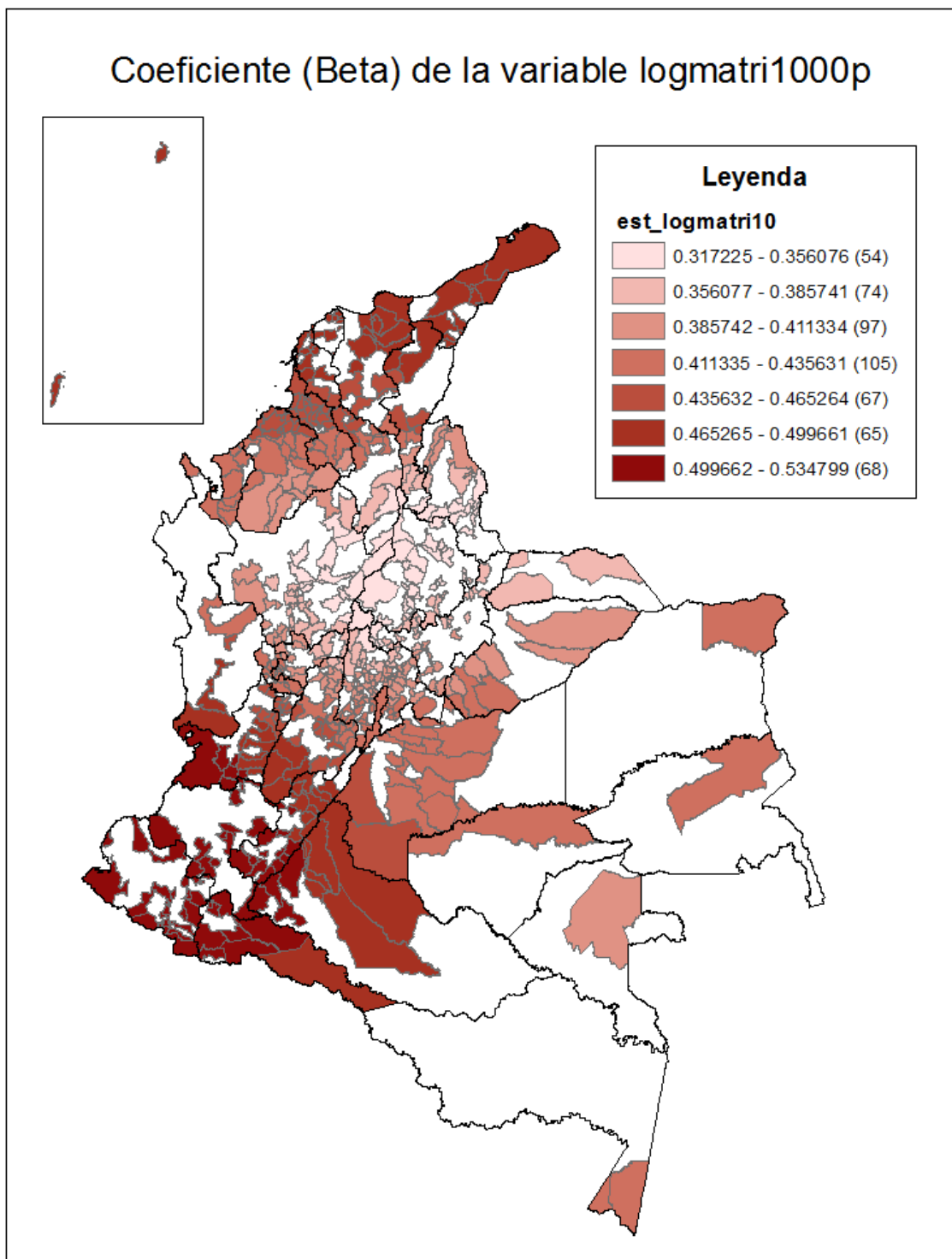


Figura 19 Mapa de distribución del coeficiente de la variable logmatri1000p

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

6.5 Diagnóstico del modelo GWR

En los procesos espaciales no se cumplen los supuestos de normalidad y homocedasticidad

...debido a: lo irregular de las unidades geográficas (los municipios) por diferencias significativas de áreas, y fundamentalmente por el papel del espacio, que no es sólo contenedor de información sino factor explicativo de los procesos que en él se presentan. Por lo tanto, los procesos espaciales tienden a ser no estacionarios, es decir que presentan variación regional o local (Estrada y Durán, 2015, p. 10)

El modelo resultante (Figura 20) obtuvo un valor de R^2 igual a 0.50 y un R^2 ajustado de 0.48, lo cual indica que aproximadamente el 50% de la variabilidad del hurto a personas y celulares es explicado por las variables predictoras.

GWR (Geographically weighted regression) result			

Bandwidth and geographic ranges			
Bandwidth size:	248.374222		
Coordinate	Min	Max	Range

X-coord	167428.000000	1658641.000000	1491213.000000
Y-coord	89436.000000	1980257.000000	1890821.000000
Diagnostic information			
Residual sum of squares:	1151.156315		
Effective number of parameters (model: trace(S)):	16.758168		
Effective number of parameters (variance: trace(S'S)):	10.734758		
Degree of freedom (model: n - trace(S)):	515.241832		
Degree of freedom (residual: n - 2trace(S) + trace(S'S)):	509.218422		
ML based sigma estimate:	1.470995		
Unbiased sigma estimate:	1.503540		
-2 log-likelihood:	1920.390077		
Classic AIC:	1955.906413		
AICc:	1957.204478		
BIC/MDL:	2031.851766		
CV:	2.328926		
R square:	0.501662		
Adjusted R square:	0.479323		

Figura 20 resumen y medidas de diagnóstico del modelo GWR

A partir del modelo GWR se obtienen valores de los coeficientes para las variables en cada municipio. Los resultados mostrados en la Tabla 5 indican que la variable *logsgppercapita* es la que presenta mayor variación geográfica, con una desviación estándar de 0.2120; por otro lado, la tasa de matriculados por cada mil personas presenta una menor dispersión representada en una desviación estándar de 0.0535.

Tabla 5 Mínimo, máximo, rango y valor promedio de los estimadores de las variables predictoras del modelo GWR

Variable	Mínimo	Máximo	Rango	Media
Intercepto	16.761877	31.837167	15.075289	25.750007
logmatri1000p	0.317225	0.534799	0.217574	0.426423
logsgppercapita	-2.198036	-1.054233	1.143803	-1.737462

En el cálculo del modelo GWR se incluyó la prueba de variabilidad geográfica que tiene como objetivo evaluar si las variables predictoras presentan variabilidad significativa en su coeficiente Beta a nivel espacial. El resultado mostrado en la Figura 21 indica que existe variación significativa en el intercepto y en la variable *logsgppercapita* (*logsgpperc*), por el contrario, la variable *logmatri1000p* (*logmatri10*) no varía significativamente. Esto se deduce a partir del valor *DIFF of Criterion* (AICc) menor a 2, que es el resultado de la diferencia entre los AICc de los modelos con el parámetro global o local. Lo anterior concuerda con la poca dispersión observada en el coeficiente de la variable *logmatri1000p* presentado en Tabla 5.

```
*****
Geographical variability tests of local coefficients
*****
```

Variable	F	DOF for F test		DIFF of Criterion
Intercept	994.233346	5.440	515.242	-1287.633180
logmatri10	2.035058	4.589	515.242	0.199089
logsgpperc	119.371867	4.763	515.242	-385.477611

Figura 21 Resultado del test de variabilidad espacial. Resultados obtenidos a partir del programa GWR4.

En la Figura 22 se observa el mapa de distribución de los residuales estandarizados del modelo GWR, los cuales, a partir del resultado del índice de Morán mostrado en la Figura 23 no se encuentra distribuidos aleatoriamente ya que existe agrupamiento espacial. Este comportamiento puede ser debido a la ausencia de variables clave en el modelo.

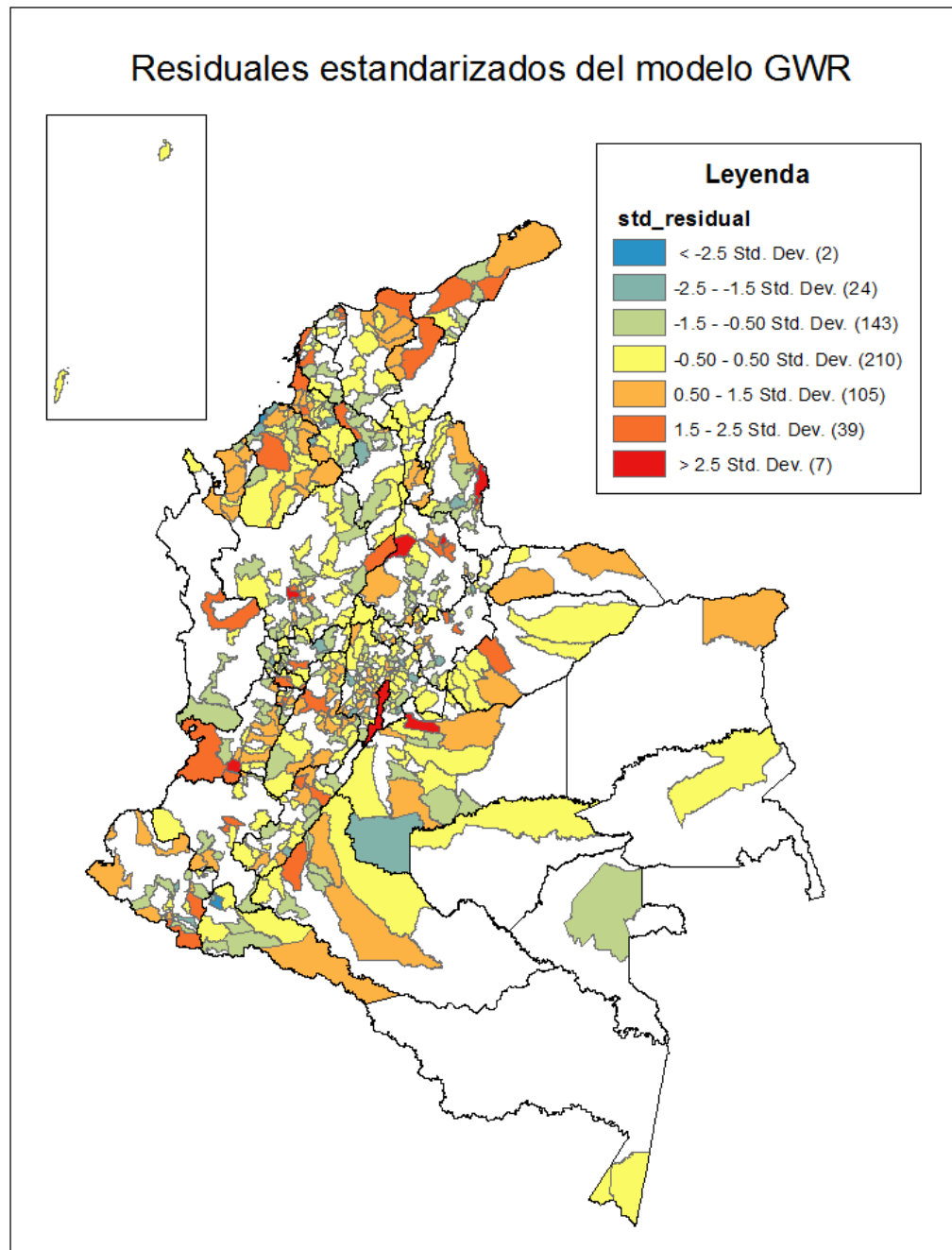


Figura 22 Mapa de distribución de los residuales del modelo GWR

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

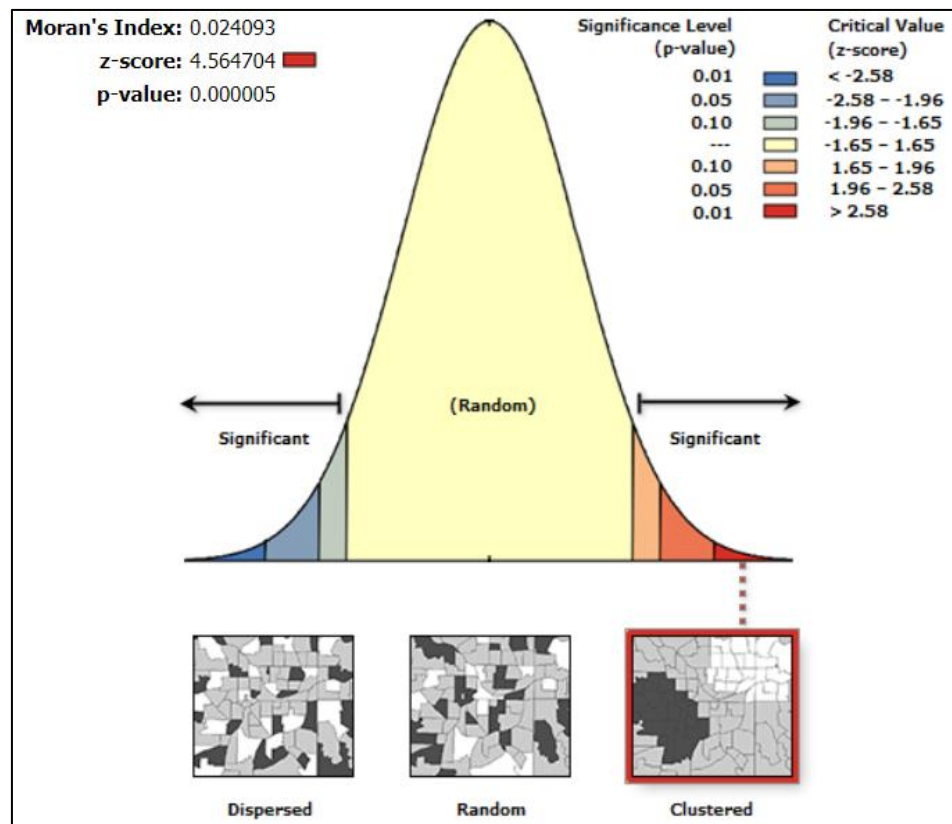


Figura 23 Resultado del Índice de Morán aplicado a los residuales estandarizados del modelo GWR

Finalmente, el coeficiente de determinación R^2 del modelo GWR mostrado en la Figura 24 indica que existen diferencias en el ajuste del modelo a nivel espacial. El menor ajuste se observa en el departamento de Córdoba y la región del Urabá, donde las variables predictoras logran explicar tan solo entre 39 y el 43% aproximadamente de la variabilidad del hurto. Por otro lado, el mejor ajuste del modelo GWR se presenta hacia el centro del país en los departamentos de Cundinamarca, Boyacá y el sur de Santander principalmente; donde se alcanzan cifras entre el 53 y 57% de variabilidad explicada.

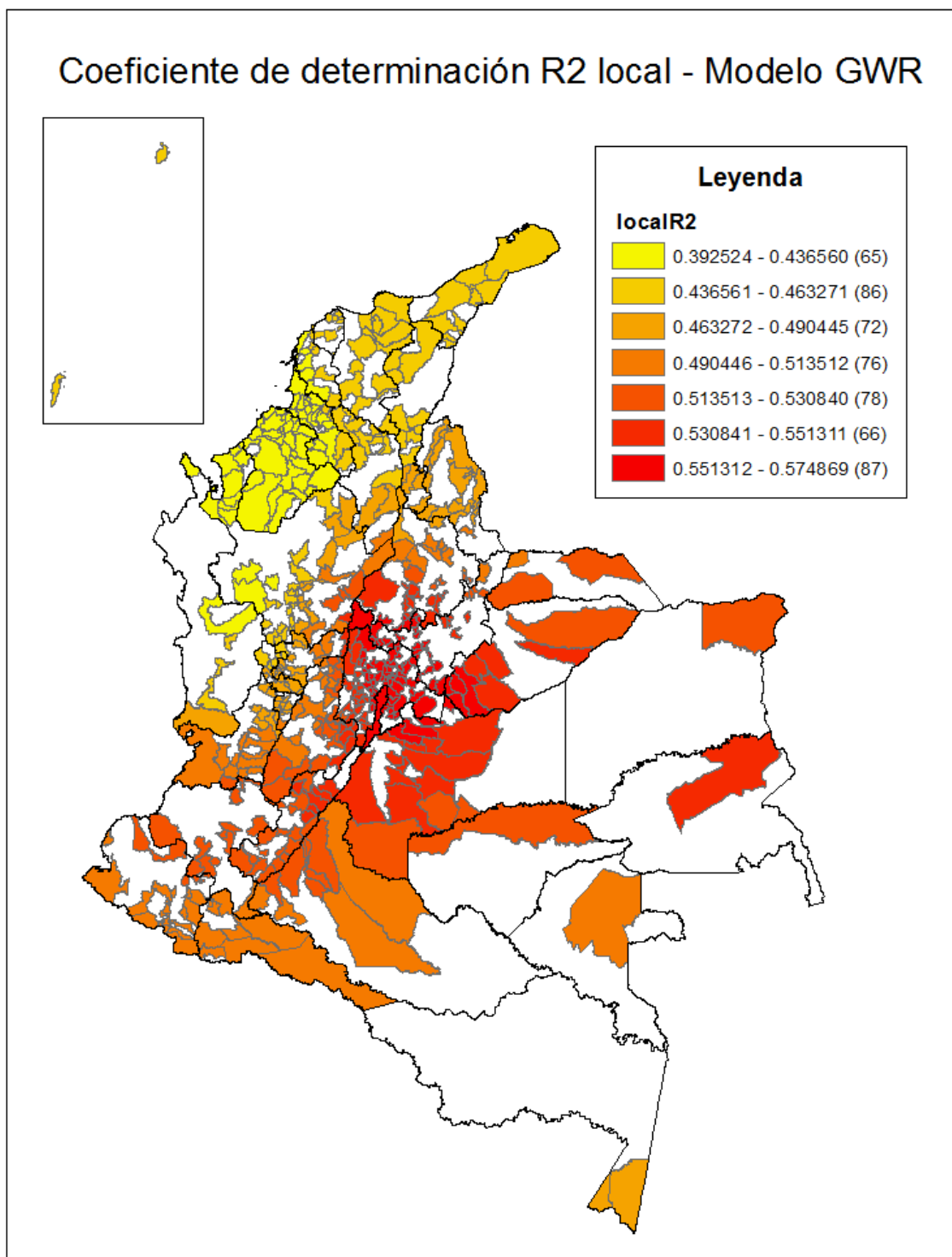


Figura 24 Coeficiente de determinación R^2 local según el modelo GWR

Nota: Los datos en paréntesis corresponden al número de municipios. Los municipios en blanco no fueron incluidos en el análisis por falta de información.

6.6 Comparación entre los modelos de regresión lineal múltiple y el modelo GWR

Una primera comparación entre los modelos se realizó utilizando las mismas variables: el logaritmo del hurto a personas y celulares como variable predicha, el logaritmo de la tasa de matriculados en educación superior por cada mil personas y el logaritmo del presupuesto per cápita asignado por el Sistema General de Participaciones como variables predictoras. Es importante mencionar que el modelo OLS usado en esta comparación no cumple con los supuestos de normalidad y homocedasticidad en los residuales. Además, la proporción de variabilidad explicada fue menor y en comparación con el modelo GWR, que obtuvo un menor AICc. Ver resultados en la Tabla 6.

El modelo de regresión lineal múltiple final incluye la variable Categoría, con la cual el modelo aumenta considerablemente el R^2 , disminuye el AIC (Tabla 6) y hace que el modelo cumpla con los supuestos de normalidad y homocedasticidad en los residuales como se evidencia en el apartado 4.2 del capítulo de resultados.

Tabla 6 Principales medidas de bondad de ajuste para la comparación entre modelos

Modelo	Predictoras	R^2 (%)	R^2 ajustado	AIC
OLS global	logmari1000p, logsgppercapita	46.13	0.4582	1969.858134
GWR (local)	logmari1000p, logsgppercapita	50.16	0.4793	1955.906413
OLS final	logsgppercapita logmari1000p Categoría	69.96	0.695	1672.6

7. Conclusiones

El modelo de regresión lineal múltiple estimado por mínimos cuadrados ordinarios que tiene como variables predictoras el presupuesto per cápita asignado por el Sistema General de Participaciones y la tasa de matriculados en instituciones de educación superior por cada mil habitantes, requirió

de la transformación logarítmica y la inclusión de la variable categórica categoría del municipio (definida por la Contaduría General de la Nación), cumple con los supuestos de normalidad y homocedasticidad de los residuales y, no multicolinealidad en las variables predictoras. Este modelo logra explicar el 69.5% de la variabilidad del logaritmo de los hurtos a personas y celulares registrados por la Policía Nacional durante el año 2017 en los 532 municipios incluidos en el análisis. Sin embargo, este tipo de modelos realiza una valoración global sin tener en cuenta la localización de los datos, por lo tanto, no tiene en cuenta las posibles heterogeneidades propias de los territorios, que es reflejada en la presencia de autocorrelación espacial de los residuales.

El modelo de regresión ponderada geográficamente estimado mediante el uso de un *kernel* de tipo adaptativo *bi-square* según el criterio de selección AICc, con las variables predictoras: presupuesto per cápita asignado por el Sistema General de Participaciones y la tasa de matriculados en instituciones de educación superior por cada mil habitantes, ambas transformadas logarítmicamente, sin incluir la variable categoría, explica el 50.16% de la variación del logaritmo del hurto con un AIC de 1955.9. De esta forma, se obtiene un mejor ajuste en comparación con el mismo modelo de regresión lineal múltiple estimado mediante mínimos cuadrados ordinarios, el cual presenta un coeficiente de determinación de 46.1 y un AIC de 1969.8. Con lo cual se puede concluir que al aplicar la técnica GWR a datos que presentan autocorrelación espacial, se obtienen mejores resultados.

A pesar del mejor ajuste logrado mediante la técnica GWR en comparación con el mismo modelo estimado por mínimos cuadrados ordinarios, el método de regresión espacial está limitado al uso de variables numéricas, por tanto, al no permitir el uso de variables categóricas, se restringe

el uso de posible información clave para el modelo. Esto se comprobó en los resultados obtenidos por el modelo de regresión lineal múltiple estimado por mínimos cuadrados ordinarios, en el que se incluyó la variable categoría del municipio, logrando de esta forma cumplir con los supuestos de normalidad y homocedasticidad en los residuales y, ausencia de multicolinealidad en las variables predictoras. Así mismo, su inclusión aumentó el coeficiente de determinación a 69.96%, lo cual representa una mejora de 19.44% con respecto al modelo GWR y del 23.83% respecto al modelo de regresión múltiple sin la variable categórica.

Del análisis realizado y los modelos evaluados, se puede deducir que para los 532 municipios incluidos en el modelo, las variables: cantidad de extranjeros por cada mil habitantes, población rural y cobertura neta en educación preescolar, básica, media y secundaria no presentan relación significativa con el hurto a personas y celulares durante el año 2017.

En los tres modelos finales se observa una relación inversa entre el hurto a personas y celulares y el presupuesto per cápita asignado a través del Sistema General de Participaciones, con lo que se concluye que, a mayor inversión en los municipios, sin importar su localización, se tiende a reducir las cifras de hurtos.

El logaritmo de los matriculados en instituciones de educación superior por cada mil habitantes, presenta una relación directa con el logaritmo del hurto a personas y celulares; relación que posiblemente sea el reflejo de una tendencia en este tipo de hurtos a ser cometidos en zonas de alto flujo de personas como lo son los centros universitarios. Sin embargo, este resultado amerita

estudios posteriores sobre la conducta criminal y ocurrencia de delitos que permitan aclarar el porqué de esta relación.

Los resultados obtenidos indican que cuanto mejor es la categoría de los municipios, mayor es la ocurrencia de hurtos a personas y celulares.

Los modelos obtenidos pueden ser utilizados como insumo en posteriores estudios de política pública cuyo objetivo sea mitigar el hurto a nivel municipal teniendo en cuenta la heterogeneidad del territorio.

8. Recomendaciones

Los estudios sobre criminalidad, especialmente los que atentan contra el patrimonio económico, requieren de la inclusión de variables e información relacionada con la calidad de vida y pobreza; ya que es razonable pensar que este tipo de delitos está motivado por el insuficiente acceso a condiciones de vida dignas asociadas a bajos ingresos. Se recomienda para futuros estudios la inclusión de algunas de estas variables: distintos indicadores de pobreza y tasas de desempleo, que no fueron incluidas en este estudio debido a la ausencia de datos actualizados y segregados a nivel municipal.

Así mismo, para futuros estudios sobre criminalidad se recomienda tener en cuenta variables disuasivas que dificultan la ejecución de los delitos como el pie de fuerza policial, militar y judicial presente de los municipios. Esta información no fue tomada en cuenta en este estudio debido a la falta de información de fácil acceso público.

Referencias

- Breusch, T. S. y Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 47, 1287. <https://doi.org/10.2307/1911963>
- Brunsdon, C., Fotheringham, A. S. y Charlton, M. (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis*, 28, 281–298. <https://doi.org/10.1111/j.1538-4632.1996.tb00936.x>
- Brunsdon, C., Fotheringham, A. S. y Charlton, M. (1998). Geographically Weighted Regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47, 431–443. <https://doi.org/10.1111/1467-9884.00145>
- Brunsdon, C., Fotheringham, A. S. y Charlton, M. (1999). Some Notes on Parametric Significance Tests for Geographically Weighted Regression. *Journal of Regional Science*, 39, 497–524. <https://doi.org/10.1111/0022-4146.00146>
- Cahill, M. y Mulligan, G. (2007). Using Geographically Weighted Regression to Explore Local Crime Patterns. *Social Science Computer Review*, 25, 174–193. <https://doi.org/10.1177/0894439307298925>
- Charlton, M. y Fotheringham, A. S. (2009). *Geographically weighted regression (White Paper)*. Recuperado de https://www.geos.ed.ac.uk/~gisteac/fspat/gwr/gwr_arcgis/GWR_WhitePaper.pdf
- Charlton, M., Fotheringham, A. S. y Brunsdon, C. (2006). *Geographically weighted regression: NCRM Methods Review Papers/NCRM/006*. Recuperado de <http://eprints.ncrm.ac.uk/90/>
- Dankhe, G. L. (1986). Investigación y comunicación. En C. Fernández-Collado y G. L. Dankhe (Eds.), *La comunicación humana: ciencia social* (pp. 385–454). México: McGraw-Hill.

- Durbin, J. y Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, 37, 409. <https://doi.org/10.2307/2332391>
- Durbin, J. y Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, 38, 159. <https://doi.org/10.2307/2332325>
- Durbin, J. y Watson, G. S. (1971). Testing for Serial Correlation in Least Squares Regression. III. *Biometrika*, 58, 1. <https://doi.org/10.2307/2334313>
- Environmental Systems Research Institute. (2016). *Regresión ponderada geográficamente GWR)-Ayuda*. Recuperado de <http://desktop.arcgis.com/es/arcmap/10.3/tools/spatial-statistics-toolbox/geographically-weighted-regression.htm>
- Estrada, L. y Durán, C. (2015). *Estudio sobre las relaciones espaciales locales entre la pobreza multidimensional, la ruralidad y la capacidad institucional*. Recuperado de https://www.researchgate.net/publication/317721631_Estudio_sobre_las_relaciones_espaciales_locales_entre_la_pobreza_multidimensional_la_ruralidad_y_la_capacidad_institucional
- Fotheringham, A. S., Charlton, M. y Brunsdon, C. (1998). Geographically Weighted Regression: A Natural Evolution of the Expansion Method for Spatial Data Analysis. *Environment and Planning A*, 30, 1905–1927. <https://doi.org/10.1068/a301905>
- Fotheringham, A. S., Charlton, M. y Brunsdon, C. (2001). Spatial Variations in School Performance: A Local Analysis Using Geographically Weighted Regression. *Geographical and Environmental Modelling*, 5, 43–66. <https://doi.org/10.1080/13615930120032617>
- Fowler, B. (2013). *Understanding Colombian Violence Through Geographic Information Systems and Statistical Approaches (Master Thesis)*. Western Kentucky University, Bowling Green, Kentucky.

- Gutiérrez-Puebla, J., García-Palomares, J.C. y Daniel-Cardozo, O. (2012). Regresión Geográficamente Ponderada (GWR) y estimación de la demanda de las estaciones del Metro de Madrid. En *XV Congreso Nacional de Tecnologías de la Información Geográfica*. Madrid.
- Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. *Biometrika*, 37, 17–23.
<https://doi.org/10.2307/2332142>
- Nakaya, T. (2015). GWR4: eapartment of Geography, Ritsumeikan University.
- Nakaya, T. (2016). *GWR4.09 User Manual: GWR4 Windows Application for Geographically Weighted Regression Modelling*. Recuperado de
https://raw.githubusercontent.com/gwrtools/gwr4/master/GWR4manual_409.pdf
- Norza, E., Peñalosa, M. J. y Rodríguez, J. D. (2017). Exégesis de los registros de criminalidad y actividad operativa de la Policía Nacional en Colombia, año 2016 *Revista Criminalidad*, 59, 9–40.
- Öcal, N. y Yildirim, J. (2010). Regional effects of terrorism on economic growth in Turkey: A geographically weighted regression approach. *Journal of Peace Research*, 47, 477–489.
<https://doi.org/10.1177/0022343310364576>
- Ortiz-Yusty, C. E., Páez, V. y Zapata, F. A. (2013). Temperature and precipitation as predictors of species richness in northern andean amphibians from Colombia. *Caldasia*, 35, 65–80.
- Pavlyuk, D. (2009). Statistical Analysis of the Relationship between Public Transport Accessibility and Flat Prices in Riga. *MPRA*,
- Policía Nacional de Colombia. *Observatorio del Delito - Policía Nacional*. Recuperado de
<https://www.policia.gov.co/observatoriodeldelito>
- R Core Team. (2018). Vienna, Austria: R Foundation for Statistical Computing. Recuperado de
<https://www.R-project.org/>

- Rincón-Ruiz, A. (2013). Coca crops, forests, people and anti-drug policy in Colombia: towards the understanding of a complex relationship (Ph.D. Thesis). Universitat Autònoma de Barcelona. Recuperado de https://ddd.uab.cat/pub/tesis/2013/hdl_10803_117303/arr1de1.pdf
- Rincón-Ruiz, A., Pascual, U. y Flantua, S. (2013). Examining spatially varying relationships between coca crops and associated factors in Colombia, using geographically weight regression. *Applied Geography*, 37, 23–33. <https://doi.org/10.1016/j.apgeog.2012.10.009>
- Sandoval, L. E. y Barón, D. M. (2008). Una revisión al estudio de la delincuencia y criminalidad. *Revista Facultad de Ciencias Económicas: Investigación y Reflexión*, 16, 105–117. Recuperado de <https://dialnet.unirioja.es/descarga/articulo/4237385.pdf>
- Shapiro, S. S. y Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 52, 591. <https://doi.org/10.2307/2333709>
- Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46, 234–240.
- Wheeler, D. y Tiefelsdorf, M. (2005). Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7, 161–187. <https://doi.org/10.1007/s10109-005-0155-6>
- Zhao, J., Wang, W. y Cheng, Q. (2014). Application of geographically weighted regression to identify spatially non-stationary relationships between Fe mineralization and its controlling factors in eastern Tianshan, China. *Ore Geology Reviews*, 57, 628–638. <https://doi.org/10.1016/j.oregeorev.2013.08.005>