10.15446/rcdg.v33n1.96975

Cuadernos de Geografía:
Revista Colombiana de Geografía

Correlación de variables geográficas y variación lingüística a partir de un modelo espacial del Atlas Lingüístico-Etnográfico de Colombia*

Javier Orlando Fernández Campos ¤
Johnatan E. Bonilla §
Luz Angela Rocha Salamanca ¥

Resumen

La dialectología es una rama de la lingüística que analiza la variación geográfica y sociolingüística de las lenguas en el espacio. Ahora bien, tradicionalmente el componente espacial en la dialectología suele limitarse a la ubicación de la variación lingüística sin tener en cuenta otros aspectos geográficos que se pueden medir. Teniendo esto en cuenta, este artículo presenta los resultados de una investigación realizada para determinar la existencia de una relación cuantitativa relevante entre variables lingüísticas y variables geográficas mediante el diseño de un modelo espacial; en él se incorporan datos léxicos del Atlas Lingüístico-Etnográfico de Colombia (ALEC), metadatos recolectados durante la expedición del atlas e información de distintos fenómenos geográficos de los territorios explorados. Entre los resultados, la evaluación explicada por el Índice de Autocorrelación Espacial sugiere que las variables aptitud agroclimática, precipitación, distancia geográfica y vías de acceso muestran la mayor dependencia espacial bivariante en relación con la distancia lingüística. El tratamiento de esta interacción dentro del modelo geográfico mixto autorregresivo de regresión espacial ratifica dicha dependencia corroborando así la relación entre la variación lingüística y los fenómenos geográficos.

Palabras clave: datos geográficos, dialectología, econometría espacial, español de Colombia, geografía, geolingüística.

Ideas destacadas: artículo de investigación que plantea la existencia de una relación cuantitativa entre variables de tipo geográfico y variables lingüísticas a través del diseño, desarrollo e implementación de un modelo espacial que incorpora datos del tomo III del Atlas Lingüístico Etnográfico de Colombia, metadatos e información de distintos fenómenos geográficos.

RECIBIDO: 28 DE JUNIO DE 2021. | EVALUADO: 24 DE ENERO DE 2022. | ACEPTADO: 1 DE JUNIO DE 2022.

CÓMO CITAR ESTE ARTÍCULO

Fernández Campos, Javier Orlando; Bonilla, Johnatan E.; Rocha Salamanca, Luz Angela. 2024. "Correlación de variables geográficas y variación lingüística a partir de un modelo espacial del Atlas Lingüístico-Etnográfico de Colombia" *Cuadernos de Geografía: Revista Colombiana de Geografía* 33 (1): 179-197. https://doi.org/10.15446/rcdg.v33n1.96975

* Proyecto resultado de la investigación de la Maestría en Ciencias de la Información y las Comunicaciones con énfasis en Geomática de la Universidad Distrital Francisco José de Caldas, Bogotá – Colombia.

¤ Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. ✉ jofernandezc@correo.udistrital.edu.co – ORCID:0000-0001-9239-6783.

§ Grupo de Investigación Lingüística – Language and Translation Technology Team, Instituto Caro y Cuervo – Universidad de Gante, Bogotá-Colombia Gante-Bélgica. ✉ johnatan.bonilla@caroycuervo.gov.co – ORCID: 0000-0002-8166-3548.

¥ Facultad de Ingeniería, Universidad Distrital Francisco José de Caldas, Bogotá-Colombia. ✉ lrocha@udistrital.edu.co – ORCID: 0000-0001-5274-4819.

✉ Correspondencia: Javier Orlando Fernández Campos, Calle 3 #5-23 Villapinzón - Cundinamarca, Colombia.

Correlation of Geographic Variables and Linguistic Variation from a Spatial Model of the Atlas Lingüístico-Etnográfico de Colombia

Abstract

Dialectology is a discipline of linguistics that examines the geographical and sociolinguistic variation of languages in space. Dialectology usually only considers the location of linguistic variation without considering other geographical aspects that can be measured. This article presents the results of an investigation into the existence of a relevant quantitative relationship between linguistic and geographic variables through the design of a spatial model that incorporates lexical data from the Atlas Lingüístico-Etnográfico de Colombia (Linguistic-Ethnographic Atlas of Colombia) (ALEC), metadata collected during the expedition of the atlas and public information on different geographic phenomena regarding the explored territories. Among the results, the evaluation explained by the spatial autocorrelation index suggested that the variables: agro-climatic suitability, precipitation, geographic distance, and access roads show the most significant bivariate spatial dependence concerning linguistic distance. Furthermore, the treatment of this interaction within the geographic autoregressive mixed model of spatial regression ratified this dependence, thus confirming the relationship between linguistic variation and geographic phenomena.

Keywords: Colombian Spanish, dialectology, geographical data, geography, geolinguistics, spatial econometrics.

Highlights: research article that addresses a quantitative relationship between geographic and linguistic variables through the design, development, and implementation of a spatial model that incorporates data from volume III of the Ethnographic Linguistic Atlas of Colombia, metadata, and information on different geographic phenomena.

Correlação de variáveis geográficas e da variação linguística a partir de um modelo espacial do Atlas Linguístico-Etnográfico da Colômbia

Resumo

A dialetologia é uma disciplina de linguística que analisa a variação geográfica e sociolinguística das línguas no espaço. Contudo, a componente espacial da dialetologia limita-se normalmente à localização da variação linguística sem considerar outros aspectos geográficos que possam ser medidos. Neste sentido, este artigo apresenta os resultados de uma investigação para determinar a existência de uma relação quantitativa relevante entre as variáveis linguísticas e geográficas. Através da concepção de um modelo espacial que incorpora dados lexicais do Atlas Lingüístico-Etnográfico da Colômbia (ALEC), metadados recolhidos durante a expedição do atlas e informação pública sobre diferentes fenómenos geográficos dos territórios explorados. Entre os resultados, a avaliação explicada pelo índice de autocorrelação espacial sugeriu que as variáveis: adequação agroclimática, precipitação, distância geográfica, e estradas de acesso mostram a dependência espacial bivariada mais significativa no que diz respeito à distância linguística. Além disso, o tratamento desta interação dentro do modelo geográfico de regressão espacial autorregressiva mista ratificou esta dependência, corroborando assim a relação entre a variação linguística e os fenómenos geográficos.

Palavras-chave: dados geográficos, dialetologia, econometria espacial, espanhol colombiano, geografia, geolingüística.

Ideias destacadas: artigo de investigação que aborda uma relação quantitativa entre variáveis geográficas e linguísticas através da concepção, desenvolvimento e implementação de um modelo espacial que incorpora dados do volume III do Atlas Etnográfico Linguístico da Colômbia, metadados, e informação sobre diferentes fenómenos geográficos.

Introducción

La variación lingüística es un fenómeno que se puede enfocar de manera multidimensional con herramientas de análisis y descripción basadas en modelos matemáticos y estadísticos (Nerbonne 2006). Además, en los últimos años la implementación de herramientas tecnológicas para el procesamiento y análisis de grandes volúmenes de datos ha permitido comprobar que la variación de una lengua representada en sus dialectos no se reduce a procesos de cambio interno, sino que se condiciona por aspectos extralingüísticos de tipo geográfico, temporal, económico y social (Dubert-García y Sousa 2016). Si bien los factores extralingüísticos no guardan una relación intrínseca con el desarrollo del sistema lingüístico, aspectos como la clase social, la edad, la raza o la religión ayudan a redimensionar las características de la variedad de una lengua y sus dinámicas de difusión y cambio (Hernández Campoy 1999).

En el campo espacial, la separación geográfica conduce de forma natural e inevitable a la separación lingüística (Labov 1994). La definición de límites o puntos por donde pasa un mismo fenómeno lingüístico (isoglosa) está dada por barreras físicas. En este sentido, dos puntos separados por grandes accidentes geográficos como cordilleras o cuerpos de agua muestran mayor diferencia lingüística en comparación con aquellos cuya barrera geográfica es una vía principal de comunicación de núcleos urbanos. Por ejemplo, para el caso de Colombia la oposición tierras bajas/tierras altas o costeño/andino ha sido definitiva para la distinción dialectal entre las regiones Pacífico y Caribe con las regiones de cordillera (Montes Giraldo 1982; Mora et ál. 2004; Bonilla y Bejarano Bejarano 2022).

De acuerdo con Anselin (1988), la relación entre distintos fenómenos geográficos puede representarse como dependencia o autocorrelación en el sentido de consecuencia de una interacción funcional entre lo que ocurre en un punto determinado y su afectación en otro lugar. Frente a esto, en lingüística se ha postulado el principio dialectológico, según el cual las variedades cercanas geográficamente tienden a ser más similares que las distantes (Nerbonne y Kleiweg 2007). Para demostrarlo, el estudio de las variedades dialectales se realiza a través de la aplicación de técnicas cuantitativas y estadísticas como la dialectometría (Goebl 2006). Mediante esta técnica se ha demostrado que en Colombia la correlación entre distancia geográfica y distancia lingüística se ajusta al postulado dialectológico, puesto que las distancias lingüísticas empiezan a volverse constantes después de los 600 km aproximadamente (Bonilla 2023). Sin embargo, de acuerdo con Rodríguez-Díaz et ál. (2018) no hay suficiente evidencia estadística teniendo en cuenta que la mayoría de los análisis dialectométricos se realizan utilizando métodos débiles ante el ruido aleatorio en los datos. En este orden, para analizar la existencia de dicha dependencia es necesario incluirla dentro de modelos espaciales de regresión que permitan conocer efectos de difusión, procesos de dispersión, interacciones, externalidades y jerarquías, entre otros (Pérez Pineda 2006).

Con base en lo anterior el presente artículo expone el diseño e implementación de un modelo espacial cuantitativo que incorpora datos geolingüísticos del tomo III del Atlas Lingüístico-Etnográfico de Colombia (ALEC) del Instituto Caro y Cuervo (ICC 1983), e incluye información léxica y espacial, con el fin de corroborar de qué manera la variación dialectal está relacionada intrínsecamente con factores geográficos.

Marco teórico

Dialectología y dialectometría

La dialectología es la disciplina de la lingüística encargada del estudio de los dialectos en un contexto social y geográfico (Heeringa 2004). De la dialectología se desprende un método de análisis denominado dialectometría, el cual es un método cuantitativo desarrollado por Séguy (1973) cuyo enfoque es la medición de distancias dialectales (Nerbonne 2006, 2010) utilizando herramientas computacionales y estadísticas (Wieling y Nerbonne 2015).

En los últimos años, la dialectometría ha integrado datos adicionales en el análisis lingüístico (Wieling 2012), desde el uso de diseños de regresión en los que se incluye la geografía como una medida de distancia, hasta el acercamiento metodológico con la sociolingüística, que incorpora factores sociales (Wieling 2012) como la edad, sexo o estrato socioeconómico. Este enfoque de modelos de regresión mixtos permite evaluar la importancia de la información lingüística frente a enfoques individuales sociales y geográficos (Wieling y Nerbonne 2015).

La dialectometría usualmente recoge datos de atlas lingüísticos dado que el componente espacial está implícito. Cada mapa de un atlas lingüístico se compone de un polígono con puntos o convenciones para señalar la presencia en localidades específicas de ciertos rasgos o variantes; estos rasgos pueden ser fonéticos, es decir, relacionados con los sonidos de las lenguas, léxicos en

el caso de variación de formas de decir o denominar un concepto, o gramaticales, cuando atienden a un fenómeno de construcción de palabras u oraciones. Para este caso la dialectometría es el instrumento que servirá para el análisis de la información de tipo léxico contenida en el ALEC y proporcionará la metodología inicial para el establecimiento de distancias, diferencias y similitudes léxicas, el cálculo de las medidas de similitud (índice relativo de identidad y distancia media geográfica) (Goebl 2006) y los fundamentos para encontrar la relación con los datos geográficos.

Medidas de similitud lingüística: el Índice Relativo de Identidad (iri)

De acuerdo con Goebl (1987) existen dos posibilidades para medir la proximidad o semejanza entre dos vectores de una matriz de datos lingüísticos:

1. Una medida de similitud no ponderada, también denominada isocrática.
2. Una medida de similitud ponderada, también llamada anisócrata.

El Índice Relativo de Identidad (IRI_{jk}) es una medida de similitud no ponderada entre vectores (j,k), conocida en alemán como "*Relativer Identitätswe*", nombre adoptado por Goebl (1987) en investigaciones dialectométricas realizadas a principios de la década de los setenta; según Goebl (1987) el IRI_{jk} se basa en el concepto taxométrico del uso de coidentidades (COI) y codiferencias (COD_{jk}) entre un par de vectores (j,k) de referencia, y su fórmula es la siguiente:

$$IRI_{jk} = 100 \cdot \frac{\sum_{i=1}^{\tilde{p}} (COI_{jk})i}{\sum_{i=1}^{\tilde{p}} (COI_{jk})i \; + \; \sum_{i=1}^{\tilde{p}} (COD_{jk})i} \qquad (1)$$

Donde $\tilde{p}$ es el número de atributos en el vector j como en el vector k; $(COI_{jk})i$ es la coidentidad entre los puntos j y k en el atributo i; j es la codiferencia entre los puntos y en el atributo es el índice del vector de referencia, k es el índice del vector en comparación y, por último, i, es el índice del atributo.

Distancia geodésica

La distancia geodésica es la longitud de recta entre dos puntos sobre una superficie de revolución denominada elipsoide, con parámetros de referencia WGS84 (World Geodetic System 1984) para el presente desarrollo. Existen numerosas técnicas de medición; sin embargo, en esta investigación se utilizará la fórmula de Vincenty, fórmula que ha generado resultados precisos y con ventajas notables en el tiempo de cálculo (Esenbuga y Colak 2016).

El Atlas Lingüístico-Etnográfico de Colombia (ALEC)

El ALEC es un compendio de mapas resultado de la investigación geolingüística del ICC. El objetivo del atlas era caracterizar y describir el español hablado en Colombia ya que a la fecha no se contaba con datos sobre las variedades rurales del país. Un total de 1.500 preguntas relacionadas con distintos temas (campos semánticos) tales como cuerpo humano, festividades y distracciones, tiempo y espacio y, en general, sobre la ruralidad colombiana, permitieron recolectar hasta 1976 los fenómenos lingüísticos, folclóricos y etnográficos tradicionales de la época (Flórez 1983).

Los lugares (localidades) seleccionados para realizar las encuestas fueron 262; el criterio de selección principal fue su representación cartográfica y se intentó conservar dentro del mapa equidistancia entre los diversos lugares de encuesta. Sumado a lo anterior se tuvieron en cuenta otros aspectos, uno de tipo cronológico, referente a la fecha de fundación de la localidad, y otro de tipo geográfico, referido a la temperatura, altura sobre el nivel medio del mar, actividad económica predominante, población y vías de acceso (Flórez 1983). No obstante, las variantes cartografiadas correspondieron a 238 del total (264) y se entresacaron 24 debido a su cercanía espacial; estas localidades restantes, sin embargo, fueron tenidas en cuenta dentro de una sección en los mapas denominada 'Otras respuestas y Adiciones'.

Las personas que respondieron una o más encuestas lingüísticas (informantes) fueron en total 2.234, en su mayoría nativos del lugar encuestado (Flórez 1983). De ellos la mitad tenían entre treinta y sesenta años; por cada uno de los hombres se interrogó a una mujer y aproximadamente una quinta parte del total era analfabeta. Como resultado de la investigación se publicaron seis tomos "cada uno de 50 x 35 cm, que contienen 1.696 láminas con 1.523 mapas de información lingüística, etnográfica o mixta, adiciones de texto, material fotográfico e ilustraciones" (Bonilla y Bernal Chávez 2020).

La importancia de la obra para la lingüística colombiana es vital ya que del ALEC se han desprendido diversas investigaciones entre las que se destacan los trabajos sobre la división dialectal del país (Montes Giraldo 1982; Mora et ál. 2004), que ha sido revisada y discutida en los últimos años mediante técnicas dialectométricas y computacionales (Ávila et ál. 2015; Rodríguez-Díaz et ál. 2018;

Bonilla y Bejarano 2022; Bonilla 2023), así como discutida y referenciada en diversos manuales de lingüística hispánica a nivel mundial. En el mismo sentido, ha sido obra de referencia para estudios del español en contacto con lenguas indígenas (Rodríguez de Montes 1987; Lancheros 2018) y diversos análisis sobre la pronunciación, el léxico y la gramática del español colombiano (Ruiz Vásquez 2014; Bernal y Díaz-Romero 2017).

El desarrollo más reciente del ALEC se realizó entre 2015 y 2020 como parte del proyecto ALEC interactivo de la Línea de Investigación en Lingüística de Corpus y Computacional (LICC) del ICC, con el apoyo del grupo de investigación Núcleo de Investigación en Datos Espaciales (NIDE) de la Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas. Gracias a esta colaboración se realizó el tránsito de los datos impresos de los seis tomos del ALEC a una Base de Datos Espacial (BDE) administrada por un sistema de información geográfica[1] (Bonilla y Bernal Chávez 2020) que permite la carga, visualización y análisis de la información. Conectado a la BDE también se desarrolló un atlas en línea denominado ALEC Digital[2] (Rocha et ál. 2018; Bonilla et ál. 2020) que se encuentra abierto al público para consulta libre.

Pesos y correlación espacial

Los pesos espaciales son un elemento clave para el modelamiento de datos geográficos. Se definen como una estructura que representa las relaciones espaciales entre las observaciones en un conjunto de datos geográficos, en la cual se establece la vecindad entre estas observaciones en términos de su proximidad geográfica. Formalmente se expresan como una matriz W de *n x n* dimensiones en la que los elementos W_{ij} son los pesos espaciales (Anselin y Smirnov 2006).

Para la medición de autocorrelación espacial (Acevedo y Velásquez 2008) se implementa el Índice I de Moran, cuyos valores varían en el intervalo [1,-1]. Si el valor I = 1, existe una relación positiva y por tanto los valores se concentran en un espacio geográfico y hay correspondencia. Entre tanto, si el valor I = -1 existe una autocorrelación negativa y los valores están perfectamente dispersos. Ahora bien, si el valor I = 0, los valores son espacialmente aleatorios (Vilalta y Perdomo 2005). Para su cálculo se sigue la siguiente fórmula matemática:

$$I = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \cdot \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2} \quad (2)$$

Donde *n* es el número de localidades, o unidades geográficas; W_{ij} es la matriz que define las distancias, o la matriz de contigüidad de pesos espaciales (en este caso); *i* es el número de filas y *j* número de columnas, valores que explican la contigüidad entre dos localidades. La prueba de significancia estadística está dada por los valores de *Z* con el supuesto de una distribución normal (Goodchild 1987).

Econometría espacial: modelos espaciales

La econometría espacial se ocupa del tratamiento de la interacción espacial y la estructura de datos espaciales para su análisis y observación dentro de modelos geográficos, e incluye, dentro del análisis de los modelos de regresión, el análisis de los efectos de dependencia espacial y heterogeneidad y/o heteroscedasticidad (Anselin 2003). Existen distintos modelos que permiten incorporar la dependencia espacial de manera formal. A continuación se presenta una breve introducción de la estructura básica de los modelos a utilizar dentro de este desarrollo.

Modelo básico de regresión múltiple

Cuando existe la posibilidad de ausencia de dependencia espacial en una variable, esta se puede recoger en un grupo de regresores a partir de la siguiente estructura:

$$\begin{aligned} y &= x\beta + \mu \\ \mu &\approx N(0, \sigma^2 I) \end{aligned} \quad (3)$$

Donde *x* es una matriz de tamaño (K, N) de K variables y N observaciones y β el vector de parámetros de estas variables con tamaño (K,1). Como se puede observar, un modelo básico solo sería correcto en términos espaciales cuando el efecto espacial (en este caso la diferencia lingüística) esté explicado por valores de una o más variables en dicho lugar (i). De acuerdo con Chasco (2003), la aplicación de este modelo mediante el método de mínimos cuadrados produce un efecto de dependencia espacial significativo que no logra explicar la estructura de la variable respuesta. En este tipo de casos hay problemas de correlación espacial; por tanto, la forma más efectiva de abordarlo es mediante el uso de modelos de dependencia

1 http://atlasweb.caroycuervo.gov.co

2 http://alec.caroycuervo.gov.co/alec/

espacial como el modelo residual (modelos de error espacial) o el modelo de dependencia espacial sustantiva (modelos de retardo espacial).

Modelo de regresión con dependencia espacial en la perturbación aleatoria o modelo de error espacial

El modelo de dependencia espacial en la perturbación aleatoria permite definir la existencia de factores o variables no considerados en el modelo y enunciarlos en términos del error (Chasco 2003). De esta manera, la relación de dependencia entre la variable respuesta (y) no se explica solo por las variables independientes sino también por aquellas que se encuentran ausentes (dependencia espacial residual). Su representación en forma general es la siguiente:

$$\begin{aligned} y &= x\beta + \mu \\ \mu &= \lambda W_\mu + \epsilon \\ \epsilon &\approx N(0, \sigma^2 I) \end{aligned} \quad (4)$$

Donde μ es la perturbación aleatoria distribuida; λ el parámetro autorregresivo, en este caso asociado al retardo espacial (w_μ), y ϵ un vector que representa las perturbaciones aleatorias.

Modelo mixto regresivo-autorregresivo espacial o modelo del retardo espacial

Este modelo incorpora la influencia de las variables a través de la variable dependiente espacialmente retardada; es decir, a través de los valores que para cada punto *i* adoptan las variables en las localizaciones vecinas. Es adecuado cuando se está poniendo de manifiesto un proceso de difusión espacial de la diferencia lingüística, de tal forma que los valores de la variable *y* en una localidad *i* estarían incrementando la probabilidad de ocurrencia de valores en lugares vecinos:

$$\begin{aligned} y &= \rho W_y + X\beta + \mu \rightarrow y = (1 - \rho W)^{-1} X\beta + \mu \\ \mu &\sim N(0, \sigma^2 I) \end{aligned} \quad (5)$$

Para esta fórmula *y* es un vector (N,1) de observaciones de la variable explicada; W la matriz de pesos espaciales de la misma variable; X una matriz de K variables exógenas: el retardo espacial; ρ el coeficiente autorregresivo espacial (un valor escalar), valor que toma la intensidad de las interdependencias entre las observaciones, y μ la perturbación aleatoria.

Contrastes de dependencia espacial

Los contrastes de dependencia espacial incluyen cinco alternativas para hallar el mejor modelo para la información de entrada. Los dos primeros son de retardo espacial (LM-Lag y Robust LM-Lag), los dos siguientes son modelos alternativos referentes al uso del error espacial (LM-Error y Robus LM-Error) y, finalmente, el modelo LM-SARMA que combina el retardo espacial con el error espacial (Figura 1).

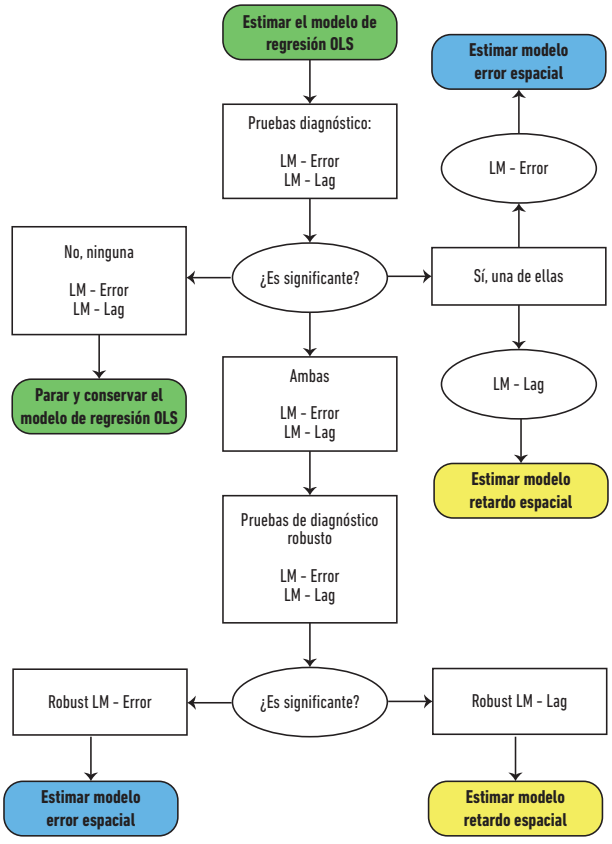

Figura 1. Diagrama de pruebas de diagnóstico de dependencia espacial.
Fuente: Anselin y Rey (2014).

Metodología

La metodología utilizada en esta investigación intenta responder la hipótesis que plantea la existencia de una relación cuantitativa (modelo espacial) entre variables geográficas y variables léxicas para explicar la variación dialectal de Colombia, con base en los datos léxicos del tomo III del ALEC. En este sentido, se desarrollaron diversas actividades que se ilustran en la Figura 2 y se exponen con detalle a lo largo de este capítulo.

Figura 2. Metodología propuesta.

Organización de datos y análisis estadístico

Para el desarrollo del modelo se capturaron los datos espaciales de 264 localidades cartografiadas originalmente en el ALEC. Los datos concernientes a fenómenos geográficos y lingüísticos se encuentran resumidos en siguiente tabla:

Tabla 1. Datos para utilizar dentro del desarrollo del modelo

Datos	Tipo	Entidad
Actividades económicas	Categórico	Metadatos ALEC - Instituto Caro y Cuervo.
Altura sobre el nivel medio del mar	Numérico	Instituto Geográfico Agustín Codazzi (IGAC).
Vías de acceso	Categórico	Metadatos ALEC - Instituto Caro y Cuervo.
Temperatura	Categórico	IDEAM
Aptitud agroclimática	Categórico	IDEAM
Índice de Disponibilidad Hídrica	Categórico	IDEAM
Riesgo por amenaza sísmica	Categórico	Servicio Geológico Colombiano.
Precipitación	Categórico	IDEAM
Distancia Lingüística Media	Numérico	Resultado IRI - Instituto Caro y Cuervo.
Distancia Geográfica	Numérico	Resultado fórmula de Vincenty - Instituto Caro y Cuervo.

Entre los metadatos de cada localidad que hacen parte del Manual del ALEC (Flórez 1983) se tuvieron en cuenta aquellos relacionados con las actividades económicas y vías de acceso para la época de las encuestas (1956-1978). Aunque parte de la información tenía registros nulos para algunas localidades se recurrió a entidades especializadas en el manejo de información espacial (Tabla 1) para completar los registros.

Con respecto a la información geográfica que no aparecía en el manual del ALEC, se trabajó con datos abiertos relacionados con temperatura, precipitación, aptitud agroclimática (Salvatore et ál. 2009) e Índice de Disponibilidad Hídrica (IDH) provenientes del Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) y con información de riesgo por amenaza sísmica del Servicio Geológico Colombiano (Tabla 1). Los datos utilizados pertenecen al intervalo temporal 1981-2010 por su proximidad a la época en que se realizó el trabajo de campo del ALEC. En cuanto a los datos de precipitación se representó la distribución espacial de la media anual sobre el territorio colombiano.

Con relación a la diferencia lingüística, en este desarrollo se calculó el Índice Relativo de Identidad (IRI_{jk}) (Goebl 1987). Los datos lingüísticos para su cálculo fueron tomados de los mapas léxicos del tomo III del ALEC que se encuentran en la BDE. Después de la depuración de elementos no estrictamente léxicos como los etnográficos, fonéticos y suplementos, se seleccionaron 100 mapas, en total, en la determinación del valor medio de similitud. Cabe resaltar que la media se utiliza para la visualización y agrupamiento; sin embargo, en esta investigación se empleó para reducir el número de valores al número de localidades (Tabla 2).

Por último, se obtuvo la distancia geográfica mediante el algoritmo de Vincenty, por el que se obtuvo una matriz cuadrada que relaciona cada valor medio de distancia entre las distintas localidades.

Tabla 2. Características de los datos a utilizar dentro del desarrollo

Datos	Mayor aparición		Menor aparición	
Actividades económicas	Agricultura (198)		Industria (5), comercio (5)	
Vías de acceso	Carretera (208)		Férrea (1), aérea (1)	
Temperatura	Entre 26 y 28 grados (64)		Entre 20 y 22 grados (9)	
Aptitud agroclimática	Extremadamente seco (112)		Moderadamente húmedo (12)	
Índice de Disponibilidad Hídrica	Semiseco (66)		Húmedo (4)	
Riesgo por amenaza sísmica	Alto (143)		Extremadamente bajo (1)	
Precipitación	Entre 1.000 y 1.500 (73)		Entre 7.000 y 9.000 (3)	
Datos	**Media**	**Máximo**	**Mínimo**	**Desviación estándar**
Distancia lingüística media	0,465078	0,617272	0,348222	0,048371
Distancia geográfica (km)	385,78	1.278,7	224,79	132,83
Altura sobre el nivel medio del mar (m)	1.266,24	3.833,12	80	1.064,5

Datos: elaborada a partir de datos IDEAM (2020); ICC (2020).

Diseño del modelo espacial

La etapa de diseño del modelo espacial comprendió la determinación de los pesos espaciales para cada una de las localidades de manera que se pudiesen incorporar las distintas relaciones espaciales entre las variables del modelo a utilizar. En este desarrollo se utilizaron los pesos espaciales basados en la contigüidad de las localidades con vértices y arcos compartidos (Anselin y Rey 2014) teniendo presente la relación de la variación lingüística con la distancia geográfica (Figura 3).

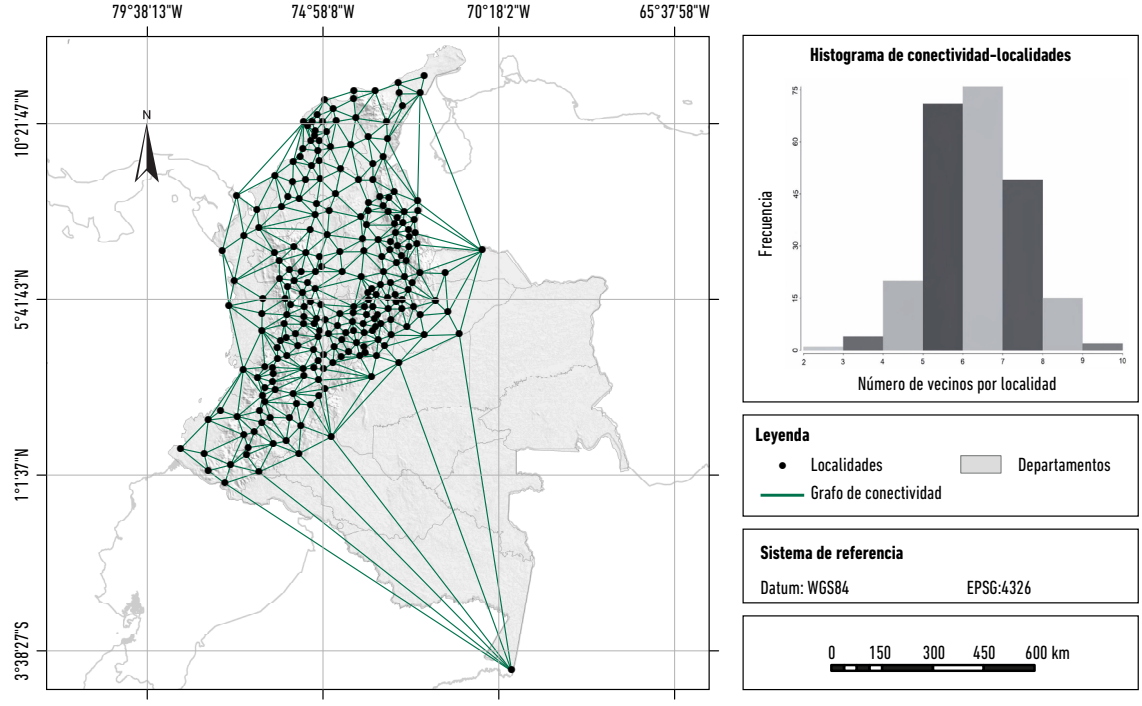

Figura 3 Mapa de conectividad e histograma de pesos espaciales.
Datos: elaborada a partir de datos ICC (2020); IGAC (2020).

El histograma de conectividad (véase Figura 3) muestra el número de observaciones para cada valor de cardinalidad; con patrón general simétrico y centro en 7, representa el 31,9 % del total. El valor mínimo de observaciones vecinas en un punto es dos y el máximo nueve. Definidos los pesos espaciales, mediante el Índice de Autocorrelación Espacial I de Moran se determinó la concentración o dispersión de los valores de cada variable dentro del territorio, así como su relación unívoca con la distancia lingüística media. Para su cálculo se hicieron en total 999 permutaciones esperando mayor confiabilidad técnica. Adicionalmente, con el Índice I de Moran Bivariante se analizó la relación entre cada una de las variables y la diferencia lingüística media, la que mostró como resultado una relación consistente entre variables geográficas y los metadatos sobre el territorio objeto de estudio.

Una vez analizada la dependencia espacial para las distintas variables geográficas, el siguiente paso consistió en identificar sus causas con el propósito de integrarla y analizarla dentro de un modelo espacial. Los resultados de las distintas pruebas para hallar el mejor modelo para la información de entrada (véase Figura 1) indicaron que los valores son significativos y robustos para todas las pruebas siendo el *p*-valor menor a α (α=0,05), lo que hizo rechazar la hipótesis nula de no dependencia espacial (Tabla 3). Se encontró que la prueba LM-LAG es más significativa que LM-ERR, así como la prueba robusta más significativa lo es para el rezago espacial. Los resultados permitieron considerar este modelo como el óptimo; no obstante, se realizó la construcción e implementación de modelos adicionales para verificar tal supuesto.

Tabla 3. Valores de pruebas de dependencia espacial

Prueba	Valor	Grados de Libertad	p-valor
LM - ERR	17,496	1	0,0288
LM - EL	10,98	1	0,0009209
LM - LAG	49,895	1	1,62e-09
LM - LE	43,379	1	4,51e-08
SARMA	60,876	2	6,04e-14

Construcción e implementación del modelo espacial

La construcción e implementación de los modelos espaciales se realizó utilizando el software estadístico R. Los modelos construidos se basaron en la generación de múltiples iteraciones de integración y descarte de variables independientes, teniendo presente tanto la matriz de pesos espaciales como el Índice de Correlación Espacial (I de Moran). Los modelos generados finalmente fueron los siguientes:

Tabla 4. Modelos construidos

Modelos generados
Modelo de regresión básico estimado por mínimos cuadrados ordinarios (MCO).
Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial.
Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial; variables con mayor significancia (VMS).
Estimación del modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR).
Modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) para variables significativas (VMS).

Selección y validación del modelo espacial

La selección del modelo óptimo se hizo con base en el coeficiente de determinación, el análisis de términos del error: heteroscedasticidad, normalidad y autocorrelación espacial, y, finalmente, el análisis de los términos estimados (Tabla 5).

Tabla 5. Resumen de estadísticos de prueba modelos generados

	Modelo MCO	Modelo de retardo espacial	Modelo de retardo espacial VMS	Modelo SARAR	Modelo SARAR VMS
Coeficiente de determinación					
R^2	0,525				
R^2 Ajustado	0,4227				
Pseudo R^2		0,6109	0,58062	0,62926	
ρ (dependencia espacial)		0,54529	0,61495	0,77318	
Test de normalidad					
Shapiro-WILK	0,986	0,990	0,993	0,989	0,990
p-valor	0,020	0,090	0,357	0,055	0,112
D'Agostino's K^2	8,049	5,919	3,582	6,362	4,679
p-valor	0,018	0,052	0,167	0,042	0,096
Heteroscedasticidad					
Breusch-Pagan	65,187	69,634	25,583	62,611	27,053
p-valor	0,0124	0,0046	0,0602	0,0212	0,0409

Correlación espacial					
I de Moran	43,583	-1,631	-12,771	-0,25413	-0,051
p-valor	1E-05	0,2448	0,2016	0,7994	0,9587

La Tabla 5 muestra cómo el R^2 clásico, el R^2 ajustado y el pseudo R^2 tienen valores que representan un éxito de predicción medio alto, con una varianza explicada que supera en su mayoría el 50 % (R^2 >0,5). Así mismo, el análisis de residuos para cada modelo al ejecutarse las pruebas de normalidad, de correlación y heteroscedasticidad, permite seleccionar como el modelo con mayor cohesión estadística al *Modelo de retardo espacial* VMS. Para este modelo la prueba de Shapiro-Wilk tiene un estadístico calculado de 0,993 con un *p*-valor de 0,35 (nivel de significancia α igual a 0,05). En este caso la hipótesis nula es que la distribución de los datos es normal; al ser el *p*-valor mayor a alfa ($p > \alpha$) no se rechaza la hipótesis nula de normalidad; por tanto, es probable que los residuos tengan esta distribución. El contraste de D'Agostino's K^2 reafirma dicha hipótesis, siendo el estadístico de 3,582 con un *p*-valor de 0,167. La hipótesis nula expresa que los datos provienen de una distribución normal y como el nivel de significancia es mayor a alfa (α) se puede afirmar que los residuos provienen de una distribución normal (Figuras 4 y 5).

Para la detección de homogeneidad o heteroscedasticidad espacial dentro de los residuos se utilizó la prueba de Breusch-Pagan, cuyo estadístico calculado para el modelo seleccionado fue 25,583 con un *p*-valor de 0,06018. La hipótesis nula plantea que existe homocedasticidad cuando el valor de α es mayor a 0,05; por tanto, existe homocedasticidad u homogeneidad en el modelo estimado seleccionado. Finalmente, para el propósito de contrastar la hipótesis de que los residuos se encuentran localizados de forma aleatoria en el espacio, existe el estadístico I de Moran que calcula la asociación de valores similares o disímiles entre regiones vecinas (Vayá y Moreno 2000). Dado que en los modelos espaciales generados el p-valor es mayor a alfa (α=0,05) no se puede rechazar la hipótesis nula; es probable que los residuos para estos modelos se distribuyan de forma aleatoria, es decir, que no estén correlacionados espacialmente.

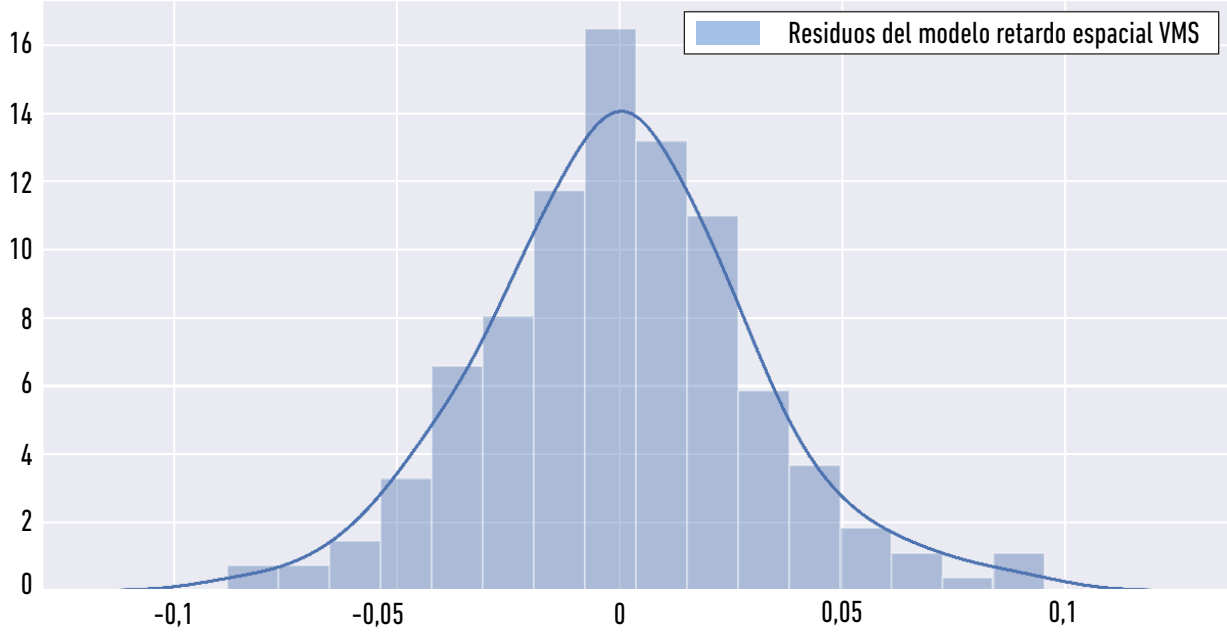

Figura 4. Histograma de residuos modelo retardo espacial VMS.
Datos: elaborada a partir de la distribución de los residuos del modelo de retardo espacial VMS aplicado a los datos del ICC (2020), IDEAM (2020), IGAC (2020).

La visualización de los residuos (Figura 5) se hizo con el uso de polígonos de Voronoi o de Thiessen; en ella se muestran los residuos teniendo presente el nivel de confianza de estos; para este modelo este nivel de confianza es igual a ±2σ, donde sigma es σ = 0,03013, es decir ±2σ=0,06026.

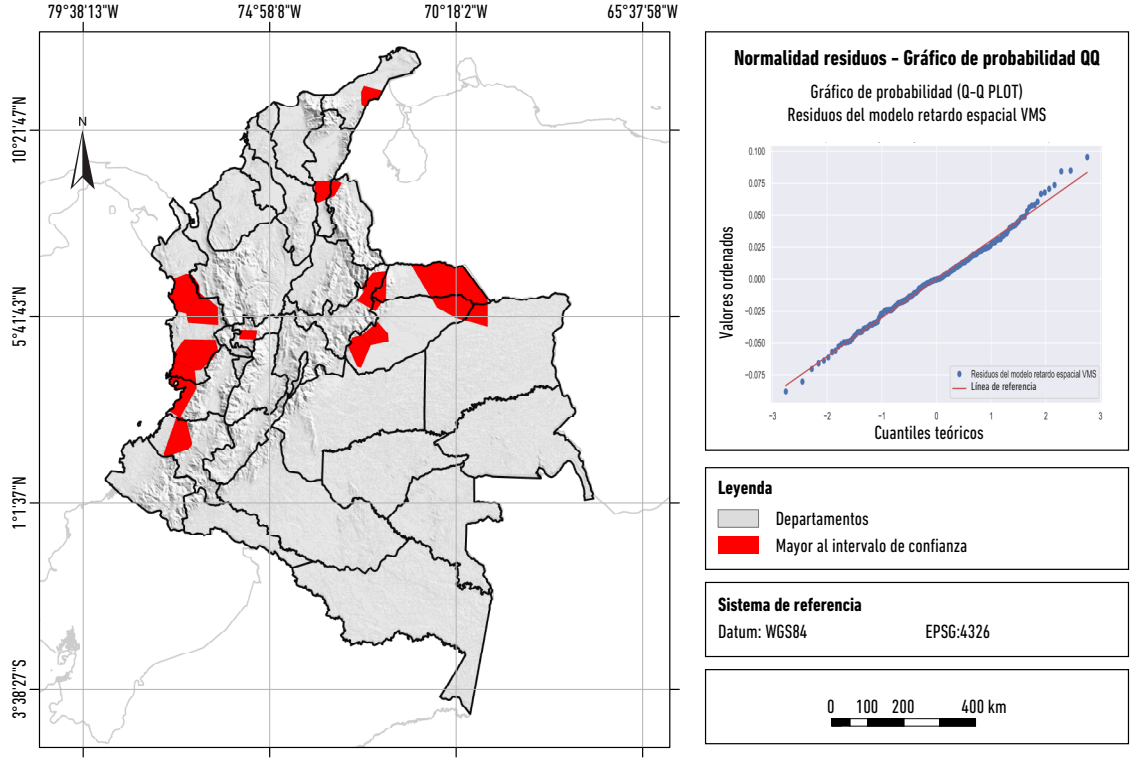

Figura 5. Residuos con el modelo de retardo espacial VMS y gráfico de probabilidad QQ.
Datos: elaborada a partir de datos ICC (2020); IGAC (2020).

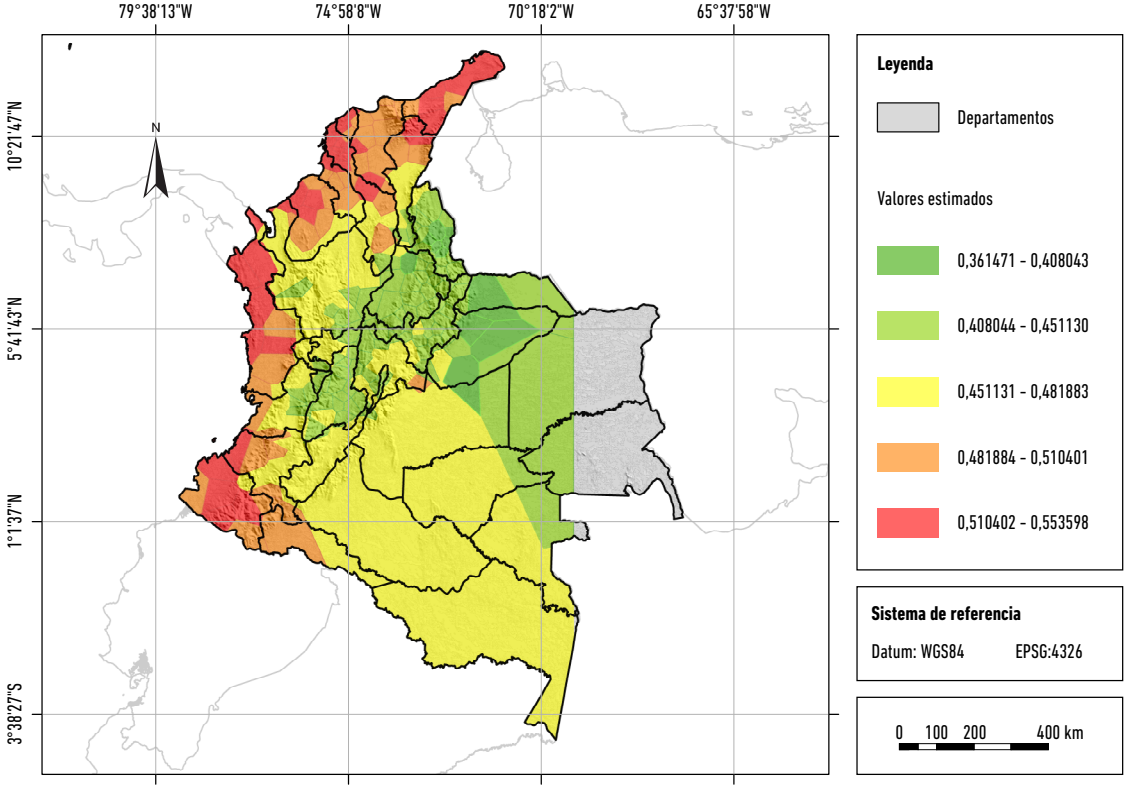

Figura 6. Valores estimados con el modelo de retardo espacial VMS.
Datos: elaborada a partir de datos ICC (2020); IGAC (2020).

La Figura 6 muestra los valores estimados con el modelo de retardo espacial VMS, exceptuando algunas áreas ubicadas en la zona suroccidental (departamento de Nariño). Se observa que la distancia lingüística media estimada se distribuye de una forma homogénea en la mayor parte del territorio; así mismo, existe una clara diferencia entre aquellas zonas en las que la distancia lingüística es mayor: regiones costeras (Costeña y Caribe), con respecto a aquellos lugares en que es relativamente baja, región central (Andina).

Resultados

I de Moran (correlación espacial)

La correlación espacial univariante y bivariante se presentó como un indicador de dependencia geográfica entre las variables que representan las observaciones dentro de cada localidad. Las categorías extremadamente seco y ligeramente húmedo de la variable Aptitud Agroclimática tienen un valor significativo de acuerdo con el índice I de Moran. De acuerdo con la prueba de correlación espacial bivariante existe una dependencia negativa cuando se trata de la relación entre la categoría ligeramente húmedo y la distancia lingüística media (Tabla 6). De esta categoría hacen parte 33 localidades en total, ubicadas en 10 departamentos, la mayor parte de ellas en la zona central (región Andina). La distancia lingüística media para estas localidades se encuentra en el intervalo [0,359525, 0,54849] y, a pesar de ser un intervalo en esencia amplio, la mayor parte de localidades concentra distancias lingüísticas medias relativamente bajas, lo que supone que el estadístico calculado es correcto. Con respecto a la variable IDH, se presenta la mayor correlación espacial en la categoría semiseco (Figura 8) con un p-valor = 0,001, lo cual supone una relación significativa; dentro del modelo seleccionado en forma equivalente se encuentra que una localidad con estas condiciones puede presentar menor distancia lingüística con respecto al total.

Tabla 6. I de Moran - objetos espaciales

Variable	I de Moran Global	*p*-valor	I de Moran Bivariante vs Distancia Lingüística media	*p*-valor
Actividades económicas				
Minería	0,057	0,054	0,003	0,446
Aptitud agroclimática				
Ligeramente húmedo	0,105	0,006	-0,113	0,001
Extremadamente seco	0,659	0,001	0,063	0,015
Moderadamente húmedo	0,260	0,001	0,018	0,245
Precipitación				
Entre 2.000 y 2.500	0,118	-0,013	-0,091	0,002
Entre 2.500 y 3.000	0,165	-0,068	-0,095	0,001
Entre 500 y 1.000	0,492	-0,128	-0,038	0,078
Temperatura				
Entre 20 y 22 grados	0,090	0,018	-0,058	0,025
Entre 26 y 28 grados	0,510	0,001	0,242	0,001
Índice de Disponibilidad Hídrica (IDH)				
Semihúmedo	0,262	0,001	0,0033	0,454
Semiseco	0,444	0,001	-0,261	0,001
Índice de Disponibilidad Hídrica (IDH)				
Vía camino de herradura	0,065	0,090	0,074	0,006
Vía carretera y férrea	0,084	0,024	0,023	0,182

Caminos carreteables	0,420	0,001	-0,206	0,001
Vía carretera	0,147	0,003	-0,041	0,070
Distancia geográfica				
Distancia geográfica	0,798	0,001	0,269	0,001

En cuanto a la variable actividades económicas la categoría minería ejerce gran influencia cuando la diferencia lingüística media es elevada; esto se ilustra en la Figura 7, en la que se presentan las localidades que ejecutaban dicha actividad en el momento de realizar la encuesta: Remedios y Zaragoza en el departamento de Antioquia, Cértegui y Nóvita en el departamento del Chocó e Iscuandé y Barbacoas en el departamento de Nariño. Es de destacar que las localidades se ubican en el noroeste, centro-oeste y suroeste del territorio, en lugares del Pacífico colombiano de composición poblacional afrodescendiente en los que existe una distancia lingüística media alta. Para este caso se puede plantear la hipótesis de la relación entre la distancia lingüística alta y zonas con alta influencia afrodescendiente, que por contacto llevan a una mayor distinción lingüística (Sharp 1970; Granda 1977; Romero 1991; Barbary y Urrea 2003). En el mismo orden, se puede afirmar que el modelo responde y verifica en sus resultados tal hipótesis, aunque la muestra no es lo suficientemente extensa para tomar este resultado como decisivo.

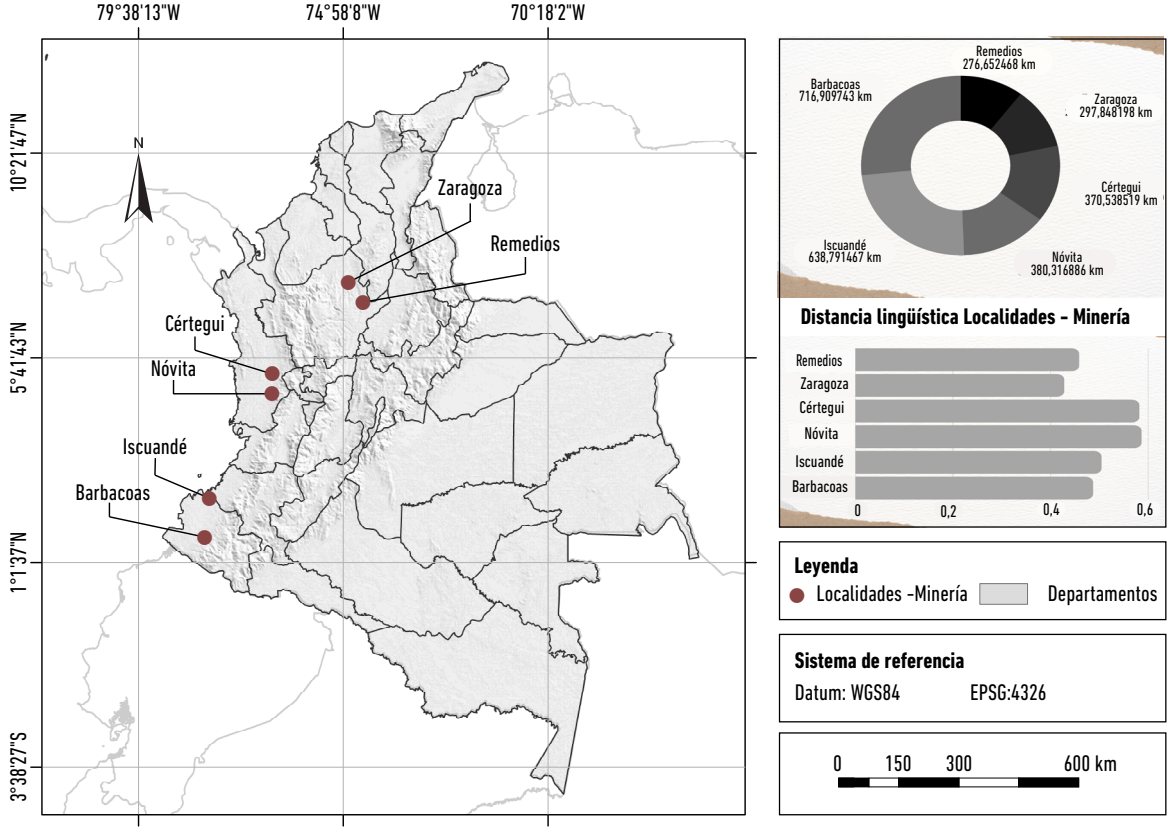

Figura 7. Distancia media geográfica y lingüística en localidades con actividad minera.
Datos: elaborada a partir de datos del ICC (2020); IGAC (2020).

Modelo espacial seleccionado

En los modelos generados a lo largo de esta investigación se siguió la metodología propuesta por Anselin y Rey (2014) (véase Figura 1), de tal manera que inicialmente se generó el modelo lineal por mínimos cuadrados ordinarios (MCO), se aplicaron las respectivas pruebas de dependencia espacial y por último se utilizaron pruebas de coherencia estadística para la selección y validación del mejor modelo (Tabla 7).

Tabla 7. Modelo del retardo espacial para las variables con mayor significancia

Variable	Estimado	Std.Error	z-valor	Pr(>|z|)
Intercepto	-0,0041	0,049	-0,0836	0,9334
Distancia geográfica				
Log (Distancia Geográfica (km)	0,0236	0,0082	2,8789	0,0039
Actividades económicas				
Minería	0,0254	0,012	1,9709	0,048
Aptitud agroclimática				
Extremadamente seco	0,0091	0,0052	1,7447	0,081
Ligeramente húmedo	-0,0068	0,0063	-1,084	0,2783
Moderadamente húmedo	0,0259	0,0125	2,0755	0,0379
Precipitación				
Entre 2.000 y 2.500	-0,011	0,0062	-1,77	0,0767
Entre 500 y 1.000	0,0064	0,0077	0,8244	0,4097
Amenaza sísmica				
Moderadamente alto	0,0068	0,0061	1,1166	0,2641
Vías de acceso				
Camino de herradura	0,0711	0,0202	3,515	0,0004
Caminos carreteables	0,0413	0,013	3,1595	0,0015
Carretera	0,0443	0,0165	2,6722	0,0075
Carretera y Férrea	0,0683	0,0162	4,1977	0,00002
Temperatura				
Entre 20 y 22 grados	-0,0058	0,0098	-0,5972	0,5503
Entre 26 y 28 grados	0,0021	0,005	0,4324	0,6654
IDH				
Semihúmedo	-0,0203	0,0108	-1,8735	0,0609
Semiseco	-0,0086	0,005	-1,7133	0,0866

Nota: Rho: 0,61495, LR test value: 74,334, p-value: 2,22e-16 - Asymptotic standard error: 0,061078, z-value: 10.068, p-value: 2,22e-16 - Wald statistic: 101,37, p-value: 2,22e-16.

El modelo seleccionado, modelo de rezago espacial, muestra cómo el coeficiente *p* es igual a 0,61 y, por tanto, evidencia un efecto positivo medio alto en la información objeto de estudio. Igualmente, las variables explicativas se someten a la prueba de Wald, lo que permite contrastar la hipótesis acerca de la coherencia del modelo y el valor que las variables agregan al mismo, en este caso una prueba significativa con un *p*-valor igual a 2,22e-16; así, se descarta la hipótesis nula de no coherencia.

Es de resaltar que dentro de las variables geográficas una de las de mayor influencia es la aptitud agroclimática, y su categoría moderadamente húmedo, con un valor β positivo y un nivel de significancia medio (Tabla 7).

De tal manera que se puede plantear la hipótesis de que zonas ubicadas en el suroccidente del territorio colombiano, entre los departamentos de Caldas, Cauca, Chocó, Huila, Tolima y Valle del Cauca, que presentan esta condición geográfica, pueden tener una distancia lingüística media mayor con relación al resto del territorio. Así mismo, la variable IDH y su categoría semiseco influye en el modelo seleccionado, como se encontró también en el cálculo de la correlación espacial. Su relación con la distancia lingüística es inversa; esto se puede ver en la Figura 8 en la que los valores se concentran en el territorio central (Región Andina), región cuya distancia lingüística media es menor respecto al resto del espacio geográfico objeto de estudio.

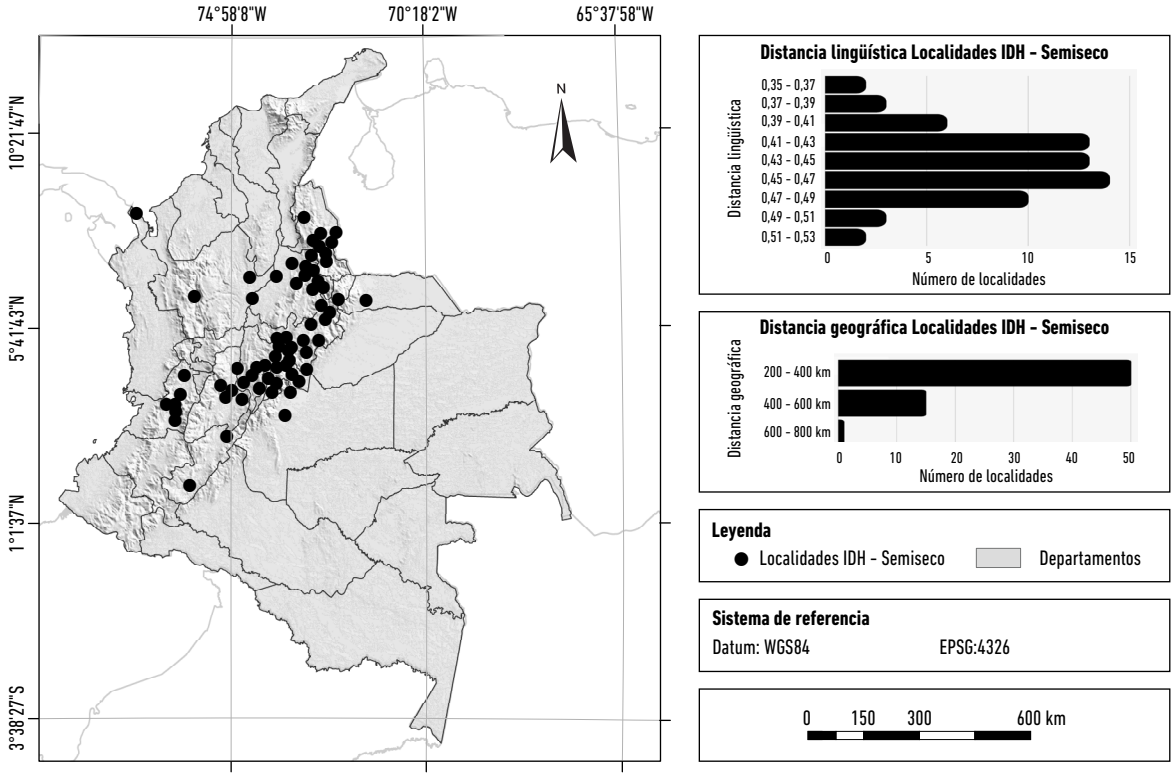

Figura 8. Localidades IDH Semiseco – Distancia geográfica y lingüística medias.
Datos: elaborada a partir de datos ICC (2020); IGAC (2020).

Finalmente, la variable denominada vías de acceso tiene dentro del modelo cuatro categorías significativas: caminos de herradura, caminos carreteables, carreteras y carreteras y vía férrea. Conforme al modelo seleccionado todas aumentan significativamente la distancia lingüística, con influencia especial de caminos de herradura. En general, las vías de acceso con mayor influencia corresponden al 94,95 % del total y aunque puede pensarse que una variable con mayor número de localidades puede generar un mayor cambio, el modelo ha determinado en sus resultados que es contraria esta hipótesis, de tal manera que la categoría camino de herradura es la más influyente sobre las demás, con solo un 1,68 % de localidades.

Por los resultados obtenidos se puede concluir que la zona con menor distancia lingüística media con respecto al total es la zona central (región Andina), seguida de la suroriental y la occidental (Pacífico). Como se mencionó, existen tanto variables geográficas como grupos de metadatos que pueden influir en la variación dialectal del territorio colombiano; a nivel histórico se ha mencionado la manera en que las actividades económicas y la facilidad o dificultad de movilización (vías de acceso) pueden influir en dicho fenómeno; sin embargo existen fenómenos sociales adicionales implicados, como es el caso del fenómeno demográfico y de migración interna registrada en el siglo XX (Jaramillo Uribe 1979; LaRosa y Mejía 2013; Osorio 2014; Kalmanovitz 2015).

De acuerdo con LaRosa y Mejía (2013) la población principal, la mestiza y blanca, en su mayoría es hoy social y culturalmente urbana; sin embargo, en el momento de realizarse la encuesta, probablemente la filiación de las familias con el campo era mayor, una diferencia notable

con la población indígena y afrocolombiana cuya mayoría habita las zonas rurales, que según LaRosa y Mejía (2013) son de reciente colonización. Por último, debe precisarse que desde tiempos de la conquista la región de mayor preferencia para el asentamiento de la población fue la región Andina, seguida de la región Caribe, cuyas diferencias lingüísticas y culturales son marcadas; de tal manera que el modelo se ajusta a ello.

Conclusiones

La evaluación principal explicada por el Índice de Autocorrelación Espacial (I de Moran) sugirió que las variables actividad económica, aptitud agroclimática, precipitación, IDH y vías de acceso muestran la mayor dependencia espacial bivariante en relación con la distancia lingüística calculada; es decir, existe correspondencia en los valores de una variable (variable geográfica o variable producto de los metadatos) en un espacio geográfico que pueden ser explicados parcialmente en función del valor de otra. El resultado de este diagnóstico de dependencia geográfica se verificó de forma cuantitativa dentro de los múltiples modelos espaciales generados.

El modelo de mayor consistencia estadística en los valores estimados, así como en los términos del error, fue el modelo mixto autorregresivo de regresión espacial, o modelo del retardo espacial, para las variables con mayor significancia (VMS). Modelo que conforme a los contrastes de dependencia espacial propuestos por Anselin (2005) permitió rechazar la hipótesis nula de inexistencia de correspondencia geográfica. En el modelo se incluyeron ocho variables en total, siendo las de mayor influencia las categorías de la variable vías de acceso, lo que valida parcialmente la hipótesis de Bonilla (2023) acerca de estas rutas como medio principal de difusión lingüística. En este sentido cabe resaltar que la concentración de infraestructura nacional vial se dio en la región Andina, y solo hacia los años sesenta la proporción de carreteras aumentó en departamentos que no se habían tenido en cuenta (Pachón y Ramírez 2006). De esta manera tanto el modelo propuesto como hechos históricos son prueba de la concentración, no solo de fenómenos lingüísticos en el centro del territorio, sino de su relación espacial inherente a actividades económicas y procesos de movilización o comunicación vial.

Si bien se mencionó cómo las variables geográficas guardan una relación de dependencia espacial con la distancia lingüística expresada en forma cuantitativa a través del Índice I de Moran, el modelo propuesto verificó dicha hipótesis; ello permitió concluir que existe una relación medible o cuantificable entre variables léxicas y variables geográficas, cuyas principales categorías corresponden a las variables ya mencionadas; se posibilitó así la estimación con un modelo óptimo para la generación de isoglosas, y la delimitación precisa y detallada de zonas dialectales.

Finalmente, el modelo espacial desarrollado se ajusta a fenómenos de tipo económico y sociodemográficos presentados a principios del siglo XIX; la incorporación de mecanismos geoespaciales y algoritmos de ciencias de la computación posibilitó la generación de nuevo conocimiento, con lo cual se demostró la relación cuantitativa entre el espacio geográfico y fenómenos lingüísticos. No obstante, es menester que investigaciones posteriores realicen la toma y análisis de información en regiones inexploradas por los investigadores del ALEC, así como la actualización de la información en los lugares visitados para la generación de un análisis más profundo y reciente.

Referencias

Acevedo Bohórquez, Ingrid, y Ermilson Velásquez Ceballos. 2008. "Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales". *Ecos de Economía* 12 (27): 9-34.

Anselin, Luc. 1988. *Spatial Econometrics: Methods and Models*. Nueva York: Springer Dordrecht. https://doi.org/10.1007/978-94-015-7799-1

Anselin, Luc y Oleg Smirnov. 1996. "Efficient Algorithms for Constructing Proper Higher Order Spatial Lag Operators". *Journal of Regional Science* 36 (1): 67-89. https://doi.org/10.1111/j.1467-9787.1996.tb01101.x

Anselin, Luc. 2003. "Spatial Econometrics". En *A Companion to Theoretical Econometrics*, editado por Badi H. Baltagi, 310 - 330. https://doi.org/10.1002/9780470996249.ch15

Anselin, Luc. 2005. *Exploring Spatial Data with GeoDa: A Workbook*. Champaign: Center for Spatially Integrated Social Science.

Anselin, Luc y Sergio J. Rey. 2014. *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and* PySAL. Chicago: GeoDa Press LLC.

Ávila, Youlín, Freddy Mendieta, Ana Constanza Álvarez, Carlos Alberto Rodríguez y Fabio Silva. 2015. "Análisis dialectométrico de las variedades del español de Colombia: nuevas aproximaciones al ALEC". En *Armonía y contrastes: estudios sobre variación dialectal, histórica y sociolingüística del español*. Madrid: Editorial Axac.

Barbary, Olivier y Fernando Urrea. 2003. "La población negra en la Colombia de hoy: dinámicas sociodemográficas, culturales y políticas". *Estudios Afro-Asiáticos* 25 (1): 9-21. https://doi.org/10.1590/S0101-546X2003000100002

Bernal Chávez, Julio Alexander y Camilo Enrique Díaz-Romero (2017). "Caracterización panorámica del español hablado en Colombia: fonología y gramática". *Cuadernos de Lingüística Hispánica* (29): 19-37.

Bonilla, Johnatan E. y Daniel Eduardo Bejarano. 2022. "Representación de la distribución diatópica de algunos procesos fonológicos del español de Colombia según el ALEC". *Literatura y Lingüística* (45): 299-332. https://doi.org/10.29344/0717621X.45.2824

Bonilla, Johnatan E. y Julio Alexander Bernal Chávez. 2020. "Modelamiento de una base de datos espacial para el Atlas Lingüístico-Etnográfico de Colombia". *Revista Signos* 53 (103): 346-368. https://doi.org/10.4067/S0718-09342020000200346

Bonilla, Johnatan E., Ruth Yanira Rubio López, Andrea Lizeth Llanos Chávez, Daniel Eduardo Bejarano y Julio Alexander Bernal Chávez. 2020. "Proyecto de digitalización y nuevas perspectivas del Atlas Lingüístico-Etnográfico de Colombia". En *Dialectología digital del español,* editado por Ángel Gallego Bartolomé y Francesc Roca Urgell, 13-28, Santiago de Compostela. https://doi.org/10.15304/9788418445316

Bonilla, Johnatan E. 2023. "Superdialectos, dialectos y subdialectos del español de Colombia". *Lexis* 47 (2).

Chasco, Coro. 2003. *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Madrid, Consejería de Economía e Innovación Tecnológica, Comunidad de Madrid.

Dubert-García, Francisco y Xulio Sousa. 2016. "On Quantitative Geolinguistics: An Illustration from Galician Dialectology". *Dialectologia*, special Issue VI, 191-221. http://doi.org/10.1344/DIALECTOLOGIA2016.2016.11

Esenbuga, Özge y Emre Colak. 2016. "Comparison of Principal Geodetic Distance Calculation Methods for Automated Province Assignment in Turkey". Presentado en la conferencia: *16th International Multidisciplinary Scientific GeoConference* SGEM 2016. Albena, Bulgaria. Del 30 de junio al 6 de julio de 2016.

Flórez, Luis. 1983. *Manual del atlas lingüístico-etnográfico de Colombia*. Bogotá: Instituto Caro y Cuervo.

Goebl, Hans. 1987. "Points chauds de l'analyse dialectométrique: pondération et visualisation". *Revue de linguistique romane* 51: 63-118.

Goebl, Hans. 2006. "Recent Advances in Salzburg Dialectometry". *Literary and Linguistic Computing* 21 (4): 411-435. https://doi.org/10.1093/llc/fql042

Goodchild, Michael. 1987. "A spatial analytical perspective on geographical information systems". *International Journal of Geographical Information Systems* 1 (4): 327-334. https://doi.org/10.1080/02693798708927820

Granda, German. 1977. *Estudios sobre un área dialectal hispanoamericana de población negra. Las tierras bajas occidentales de Colombia*. Bogotá: Instituto Caro y Cuervo.

Heeringa, Wilbert. 2004. "Measuring Dialect Pronunciation Differences Using Levenshtein Distance". Tesis doctoral en Humanities Computing, University of Groningen Library, Groningen.

Hernández Campoy, Juan Manuel. 1999. *Geolingüística: modelos de interpretación geográfica para lingüistas*. Murcia: Editum, Ediciones de la Universidad de Murcia.

ICC (Instituto Caro y Cuervo). 1983. *Atlas Lingüístico-Etnográfico de Colombia (ALEC)*. Bogotá: Instituto Caro y Cuervo.

Jaramillo Uribe, Jaime. 1979. *Manual de historia de Colombia: siglo* XIX. Biblioteca Colombiana de Cultura - Instituto Colombiano de Cultura.

Kalmanovitz, Salomón. 2015. *Breve historia económica de Colombia*. Bogotá: Universidad Jorge Tadeo Lozano.

Labov, William. 1994. *Principles of Linguistic Change. Volume I: Internal Factors (Language in Society 20)*. Oxford: Blackwell.

Lancheros Redondo, Hugo Fernando. 2018. "Los indigenismos léxicos en las variedades diatópicas del español colombiano". *Forma y Función* 31 (2): 9-29. https://doi.org/10.15446/fyf.v31n2.74652

LaRosa, Michael y Germán R. Mejía. 2013. *Historia concisa de Colombia (1810-2013)*. Bogotá: Universidad del Rosario.

Montes Giraldo, José Joaquín. 1982. "El español de Colombia: propuesta de clasificación dialectal". *Thesavrvs: Boletín del Instituto Caro y Cuervo* 37 (1): 23-92.

Mora Monroy, Siervo Custodio, Mariano Lozano, Ricardo Aparicio Ramírez Caro, María Bernarda Espejo Olaya y Gloria Esperanza Duarte Huertas. 2004. *Caracterización léxica de los dialectos del español de Colombia según el ALEC*. Bogotá: Publicaciones del Instituto Caro y Cuervo.

Nerbonne, John. 2006. "Identifying Linguistic Structure in Aggregate Comparison". *Literary and Linguistic Computing* 21 (4): 463-475. https://doi.org/10.1093/llc/fql041

Nerbonne, John. 2010. "Measuring the Diffusion of Linguistic Change". *Philosophical Transactions of the Royal Society B: Biological sciences* 365 (1559): 3821-3828. https://doi.org/10.1098/rstb.2010.0048

Nerbonne, John y Peter Kleiweg. 2007. "Toward a Dialectological Yardstick". *Journal of Quantitative Linguistics* 14 (2-3): 148-166. http://doi.org/10.1080/09296170701379260

Osorio Baquero, Ismael. 2014. "Breve reseña histórica de las vías en Colombia". *Revista Ingeniería Solidaria* 10 (17): 183-187. http://doi.org/10.16925/in.v10i17.880

Pachón, Álvaro y María Teresa Ramírez. 2006. *La infraestructura de transporte en Colombia durante el siglo XX*. Bogotá: Fondo de Cultura Económica.

Pérez Pineda, Jorge A. 2006. "Econometría espacial y ciencia regional". *Investigación Económica* 65 (258):129-160.

Rocha Salamanca, Luz Angela, Johnatan Bonilla, Julio Alexander Bernal Chávez, Catherine Duarte y Alejandro Rodriguez. 2018. "Design and Implementation of the Web Linguistic and Ethnographic Atlas of Colombia". *Proceedings of the International Cartographic Association* 1: 1-4. https://doi.org/10.5194/ica-proc-1-96-2017

Rodríguez de Montes, María Luisa. 1987. "Algunos quechuismos en el "ALEC": posibles quechuismos en el muisca y en el español de la primitiva zona de asentamiento muisca". *Thesaurus: Boletín del Instituto Caro y Cuervo* 42 (1): 95-121.

Rodríguez-Diaz, Carlos Andrés, Sergio Jiménez, George Dueñas, Johnatan Estiven Bonilla y Alexander Gelbukh. 2018. "Dialectones: Finding Statistically Significant Dialectal Boundaries Using Twitter Data". *Computación y Sistemas* 22 (4): 1213-1222. https://doi.org/10.13053/cys-22-4-3104

Romero, Mario Diego. 1991. "Procesos de poblamiento y organización social en la costa pacífica colombiana". *Anuario Colombiano de Historia Social y de la Cultura* (18-19) (enero): 9-31.

Ruiz Vásquez, Néstor Fabián. 2014. *Léxico de la muerte en el español hablado en Colombia, según el ALEC: estudio dialectológico y lexicológico*. Bogotá: Instituto Caro y Cuervo.

Salvatore, Mirella, Amir Kassam, Ana Cecilia Gutiérrez, Mario Bloise y Michela Marinelli. 2009. *Metodología de Evaluación de Aptitud de Tierras*. Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO). Consultado el 10 de enero de 2020. http://www.fao.org/3/i1708s/i1708s02.pdf

Séguy, J. 1973. "La dialectometrie dans l'Atlas linguistique de Gascogne". *Revue de Linguistique Romane* (37): 1-24. https://doi.org/10.5169/seals-658403

Sharp, William Frederick. 1970. "Forsaken but for Gold: An Economic Study of Slavery and Mining in the Colombian Chocó, 1680-1810". Chapel Hill, North Carolina: University Microfilms Internacional.

Vayá Valcarce, Esther y Rosina Moreno Serrano. 2000. "La utilidad de la econometría espacial en el ámbito de la ciencia regional". Barcelona: Ediciones Universidad de Barcelona.

Vilalta y Perdomo, Carlos Javier. 2005. "Cómo enseñar autocorrelación espacial". *Economía, Sociedad y Territorio* 5 (18): 323-333.

Wieling, Martijn. 2012. *A Quantitative Approach to Social and Geographical Dialect Variation*. Groningen: University of Groningen Library.

Wieling, Martijn y John Nerbonne. 2015. "Advances in Dialectometry". *Annual Review of Linguistics* 1: 243-264.

Javier Orlando Fernández Campos

Magíster en Ciencias de la Información y las Comunicaciones con énfasis en Geomática de la Universidad Distrital Francisco José de Caldas; miembro destacado del grupo de investigación Núcleo de Investigación en Datos Espaciales (NIDE). Ha acumulado una amplia experiencia laboral, desempeñando funciones en el Instituto Geográfico Agustín Codazzi (IGAC), en la Subdirección de Geografía y Cartografía. Durante su tiempo en el IGAC, se destacó por la implementación de proyectos y procesos orientados a la producción y mantenimiento de información geodésica y cartográfica. En la actualidad, está encargado de la gestión de la información espacial en la Corporación Regional del Centro de Antioquia (CORANTIOQUIA).

Johnatan E. Bonilla

Experto en Lingüística de Corpus y Computacional, con maestría del Instituto Caro y Cuervo (ICC) y doctorado en curso en la Gent Universiteit y la Humboldt-Universität zu Berlin. Destacado en el diseño de sistemas de gestión de corpus, diccionarios y atlas lingüísticos, con un enfoque especial en dialectología digital y adaptación de modelos de lenguaje a dialectos colombianos. Sus intereses de investigación abarcan geolingüística, métodos computacionales para la dialectología y el análisis morfosintáctico del español rural/oral.

Luz Angela Rocha Salamanca

Profesora titular de la Facultad de Ingeniería de la Universidad Distrital Francisco José de Caldas. Doctora en Geografía por la Universidad Nacional de Colombia. Magíster en Ciencias de la Geo-información en el ITC, ahora Facultad de Ciencias de la Geo-información y Observación de la Tierra de la Universidad de Twente (Países Bajos). Miembro del Comité Técnico Asesor para la Gestión Catastral de Colombia como experta en cartografía, teledetección y geodesia. Presidenta de la Asociación de Especialistas en Percepción Remota y Sistemas de Información Geográfica SELPER Capítulo Colombia.