Dialectologia 32 *(2024), 165-191.*
ISSN: 2013-2247

Received 10 November 2021.

Accepted 26 February 2022.

Published January 2024.

DOI: 10.1344/DIALECTOLOGIA2023.32.7

THE INFLUENCE OF GEOGRAPHIC VARIABLES IN LINGUISTIC VARIATION[1]

Javier Orlando FERNÁNDEZ[1] , Johnatan BONILLA[2] & Luz Ángela ROCHA[1] *

District University Francisco José de Caldas[1] / Caro and Cuervo Institute[2]

jofernandezc@correo.udistrital.edu.co / johnatan.bonilla@caroycuervo.gov.co /

lrocha@udistrital.edu.co

ORCID: 0000-0001-9239-6783 / 0000-0002-8166-3548 / 0000-0001-5274-4819

Abstract

Dialectology is a branch of linguistics that analyzes the geographical and sociolinguistic variation of languages in space. However, traditionally the spatial component in dialectology is usually limited to the location of data collection without considering other geographical aspects that can be measured. According to this article, it shows the results of the research to determine the existence of a relevant quantitative relationship between linguistic variants and geographic variables, through the design of a spatial model that incorporates lexical data from the Linguistic-Ethnographic Atlas of Colombia (*Atlas Lingüístico-Etnográfico de Colombia (ALEC)*), metadata and information of different geographical phenomena. Among the results, the evaluation explained by the spatial autocorrelation index suggested that the variables: agroclimatic suitability, precipitation, geographical distance and access roads show the highest bivariate spatial dependence in relation to the linguistic distance calculated with the Relative Identity Index (IRI). The treatment of this interaction within the mixed geographic autoregressive spatial regression model confirmed this dependence, thus corroborating the relationship between language and geography.

Keywords: dialect, econometrics, geographical data, geography, language change

[1] Result of the research project of the Master degree: Information and Communication Sciences at the Universidad Distrital "Francisco José de Caldas", Bogotá - Colombia.

* [1] Faculty of Engineering, District University Francisco José de Caldas, Bogotá-Colombia / [2] Instituto Caro y Cuervo, Bogotá, Colombia – Ghent University, Belgium.

LA INFLUÈNCIA DE LES VARIABLES GEOGRÀFIQUES EN LA VARIACIÓ LINGÜÍSTICA

Resum

La dialectologia és una branca de la lingüística que analitza la variació geogràfica i sociolingüística de les llengües en l'espai. No obstant això, tradicionalment el component espacial en dialectologia s'ha limitat a la localització de la recollida de dades sense tenir en compte altres aspectes geogràfics que es poden mesurar. Aquest article mostra els resultats de la investigació per determinar l'existència d'una relació quantitativa rellevant entre variants lingüístiques i variables geogràfiques, mitjançant el disseny d'un model espacial que incorpora dades lèxiques de l'Atles Lingüístic-Etnogràfic de Colòmbia (*Atlas Lingüístico-Etnográfico de Colombia* (ALEC)), metadades i informació sobre diferents fenòmens geogràfics. Entre els resultats, l'avaluació que s'explica per l'índex d'autocorrelació espacial va suggerir que les variables idoneïtat agroclimàtica, precipitació, distància geogràfica i vies d'accés tenen una major dependència espacial bivariada en relació amb la distància lingüística calculada amb l'Índex d'Identitat Relativa (IRI). El tractament d'aquesta interacció dins del model de regressió espacial autoregressiva geogràfica mixta va confirmar aquesta dependència, corroborant així la relació entre llengua i geografia.

Paraules clau: dialecte, econometria, dades geogràfiques, geografia, canvi lingüístic

LA INFLUENCIA DE LAS VARIABLES GEOGRÁFICAS EN LA VARIACIÓN LINGÜÍSTICA

Resumen

La dialectología es una rama de la lingüística que analiza la variación geográfica y sociolingüística de las lenguas en el espacio. Sin embargo, tradicionalmente el componente espacial en dialectología se ha limitado a la localización de la recogida de datos sin tener en cuenta otros aspectos geográficos que se pueden medir. Este artículo muestra los resultados de la investigación para determinar la existencia de una relación cuantitativa relevante entre variantes lingüísticas y variables geográficas, mediante el diseño de un modelo espacial que incorpora datos léxicos del *Atlas Lingüístico-Etnográfico de Colombia* (ALEC), metadatos e información sobre distintos fenómenos geográficos. Entre los resultados, la evaluación que se explica por el índice de autocorrelación espacial sugirió que las variables idoneidad agroclimática, precipitación, distancia geográfica y vías de acceso tienen una mayor dependencia espacial bivariada con relación a la distancia lingüística calculada con el Índice de Identidad Relativa (IRI). El tratamiento de esta interacción dentro del modelo de regresión espacial autorregresiva geográfica mixta confirmó esta dependencia, corroborando así la relación entre lengua y geografía.

Palabras clave: dialecto, econometría, datos geográficos, geografía, cambio lingüístico

1. Introduction

Linguistic variation is a complex phenomenon, its characteristics can be studied in a multidimensional context with analysis and description tools based on mathematical and statistical models (Nerbonne 2006). In recent years this type of analysis has had a particular relevance where the use of technological tools for the processing and analysis of large volumes of data has been fundamental, in such a way

that experts emphasize that dialect variation is not reduced to simple characterizations, but there are geographical, temporal, economic and social aspects (extralinguistic factors) that affect the variation of languages (Dubert-García & Sousa 2016).

The extralinguistic factors in essence do not bear any relationship to the linguistic phenomena of the system, but according to Campoy (1999) aspects such as social class, age, race, or religion have a primary importance in the differentiation of the language, its diffusion, and its change. In the spatial field, in the words of Labov (1994), geographic separation naturally and inevitably leads to linguistic separation, showing that generally the configuration of certain isoglosses in linguistic maps is highly related to certain physical barriers, in such a way that isolated sites separated by valleys or hills, generally maintain a considerable linguistic difference, differing with those sites whose geographic barrier is a main communication route of urban centers. In fact, a dialect barrier in England, it crosses a swampy area with difficult access called 'The Fens' between the counties of Cambridge and Lincoln (Campoy 1999).

According to Anselin (1988) the relationship between different geographical phenomena can be represented as dependency or autocorrelation, this is manifested as a consequence of a functional interaction between what happens in a certain point and its affectation in another place, being described and verified in dialectometry by Goebl (2006), who shows that as the geographical distance between two locations increases, their linguistic difference also increases; Likewise, the existence of this dependency, according to Anselin (1988), can be included within spatial regression models that allow us to know diffusion effects, dispersion processes, interactions, externalities, hierarchies, among others (Pérez-Pineda 2006). Based on the above, this article presents the design and implementation of a quantitative spatial model that incorporates geolinguistic data from volume III of the Linguistic-Ethnographic Atlas of Colombia (*Atlas Lingüístico-Etnográfico de Colombia* (ALEC)), including lexical and spatial information, to corroborate how dialect variation is intrinsically related to geographic factors.

2. Framework

2.1 Dialectology and dialectometry

Dialectology is the discipline in charge of the study of dialects in a social and geographical context (Heeringa 2004), from dialectology an analysis method called dialectometry emerges, which is a quantitative method developed by Séguy (1973) whose focus is measurement of dialect distances (Nerbonne 2006) using computational and statistical tools (Wieling & Nerbonne 2015).

Dialectometry in recent years has integrated additional data into linguistic analysis (Wieling 2012), from the use of regression designs in which geography is included as a measure of distance to the methodological approach with sociolinguistics, which includes social factors (Wieling 2012) such as age, sex, or socioeconomic status. This mixed regression model approach allows evaluating the importance of linguistic information against individual social and geographic approaches (Wieling & Nerbonne 2015).

Dialectometry will be the instrument and it will support the analysis of the information (lexical) contained in the ALEC, it will provide the initial methodology for the establishment of distances, differences and lexical similarities, the calculation of similarity measures (relative identity index) (Goebl 1987) and the rationale for finding the relationship with geographic data.

2.2 Linguistic Similarity Measures: The Relative Index of Identity (IRI)

According to Goebl (1987) there are two possibilities to measure the proximity or similarity between two linguistic data vectors of a matrix.

- An unweighted measure of similarity, also called isocratic.
- A weighted similarity measure, also called an anisocrat.

The Relative Index of Identity (IRI_{jk}) is an unweighted measure of similarity between vectors (j, k), known in German as "*Relativer Identitätswert* (RIW_{jk})", a name adopted by Goebl (1987) in dialectometric investigations carried out at the beginning

of the 1970s, according to Goebl (1987) the IRI_{jk} is based on the taxometric concept of co-identities use (COI_{jk}) and co-differences (COD_{jk}) between a pair of reference vectors (j, k), its formula is the following:

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

Where $\tilde{p}$ is the number of attributes in vector j as in vector k, $(COI_{jk})i$ is the co-identity between points j and k in attribute i, $(COD_{jk})i$ is the co-difference between points j and k in attribute i, j is the index of the reference vector, k is the index of the vector in comparison, and finally i is the index of the attribute.

2.3 Weights and spatial correlation

Spatial weights are a key element for modeling geographic data, they are defined as a structure based on neighborhoods, where they formally express the contiguous structure between the observations (in this case the number of locations) as a W matrix of *n x n* dimensions in where the elements W_{ij} are the spatial weights (Anselin & Smirnov 2006).

Likewise, for the measurement of spatial autocorrelation there is Moran's I index (Bohórquez & Ceballos 2008), a method similar to Pearson's correlation, where the values vary in the interval [1,-1], if the value $I = 1$, there is a positive relationship, the values are concentrated in a geographic space and there is correspondence, if the value $I = -1$ there is a negative autocorrelation and the values are perfectly dispersed, if the value $I = 0$, the values are spatially random (Vilalta 2005); for its calculation its mathematical formula is as follows:

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

Where *n* is the number of locations, or geographic units, W_{ij} is the matrix that defines the distances, or the contiguity matrix of spatial weights (in this case), i is the number of rows and j number of columns, values that explain the contiguity between two localities. The test of statistical significance is given with the assumption of a normal distribution (Goodchild 1987).

2.4 Spatial Econometrics: Spatial Models

Anselin (2003) defines spatial econometrics as a branch of econometrics that deals with the treatment of spatial interaction and the structure of spatial data for its analysis and observation within geographic models. There are different and varied models that allow the incorporation of spatial dependence in a formal way, below there is a brief introduction of the basic structure of the regression models to be used within this development:

2.4.1 Basic linear regression model (MBRL)

The basic linear regression model allows finding a (linear) relationship between a dependent variable and a set of explanatories and/or independent variables (Yrigoyen 2002), its expression is given by:

$$y = x\beta + \mu \qquad (3)$$
$$\mu \approx N(0, \sigma^2 I)$$

Where x is a matrix of size (K, N) of K variables and N observations, β the parameters vector of these variables with size $(K, 1)$ and μ the random disturbance (Yrigoyen 2002); the specification of a basic model (equation 3) would only be correct in spatial terms when the spatial effect (in this case the linguistic difference) is fully explained by values of one or more variables in that place (i). According to Yrigoyen (2002), generally the use of the MBLR model estimated by ordinary least squares (OLS)

produces a significant spatial dependence effect due to its poor estimation, since it fails to explain the structure of the response variable. When this occurs, the truest way to address these types of issues is with the use of spatial dependency models, such as the residual model (spatial error models) or the substantive spatial dependency model (spatial lag models).

2.4.2 Mixed autoregressive model of spatial regression or model of spatial lag

This model incorporates the influence of the variables through the spatially lagged dependent variable, that is, through the values that the variable adopts for each point i in the neighboring locations, this model is suitable when a process of spatial diffusion of the linguistic difference is being revealed, in such a way that the values of this variable *y* in a locality i would be increasing the probability of occurrence of values in neighboring places:

$$y = \rho W_y + X\beta + \mu \rightarrow y = (1 - \rho W)^{-1} X\beta + \mu \qquad (4)$$

$$\mu \sim N(0, \sigma^2 I)$$

Where y is a vector $(N, 1)$ of observations of the explained variable, W the matrix of spatial weights of the same variable, X a matrix of K exogenous variables, W_y the spatial lag, ρ the spatial autoregressive coefficient (a scalar value), value that takes the intensity of the interdependencies between the observations and finally μ the random disturbance (Yrigoyen 2002).

2.4.3 Regression model with spatial dependence on random disturbance or spatial error model

The model of spatial dependence in the random disturbance according to Yrigoyen (2002) is a structure that allows defining the existence of factors or variables not considered in the model and stating them in terms of the error, in this way the

dependency relationship between the response variable (y) is explained not only by the independent variables but by those that are absent as well (residual spatial dependence), its representation in general is as follows:

$$y = x\beta + \mu$$
$$\mu = \lambda W_\mu + \epsilon \qquad (5)$$
$$\epsilon \approx N(0, \sigma^2 I)$$

Where μ is the distributed random disturbance, λ the autoregressive parameter in this case associated with the spatial lag (W_μ) and finally ϵ, a vector that represents the random disturbances.

2.4.4 Spatial dependency contrasts

The model was selected from the detection contrasts and spatial dependence analysis, for this Anselin (2005) proposed a Lagrange multipliers test, these tests include five alternatives to find the best model for the input information, the first two tests are spatial lag (LM-Lag and Robust LM-Lag), the following two are alternative models referring to the use of spatial error (LM-Error and Robust LM-Error) and finally we have the LM-SARMA model, one last higher order alternative that combines spatial lag with spatial error.

3. Study area: Localities of the Linguistic-Ethnographic Atlas of Colombia (ALEC)

The Linguistic-Ethnographic Atlas of Colombia (*Atlas Lingüístico-Etnográfico de Colombia* (ALEC)) is a compendium of maps resulting from dialectological research applied to Colombian Spanish and initiated by Luis Flórez, researcher at the Caro and Cuervo Institute (ICC), around 1954; The survey began with a questionnaire that consisted of 8065 questions and reduced by 1961 to a total of 1500; questions related

to different topics (semantic fields), collecting traditional linguistic, folkloric, and ethnographic elements of Colombia (Flórez 1983).

The places (localities) selected to carry out the surveys were 262, the main selection criterion was their cartographic representation, trying to preserve within the map equidistance between the various survey places, in addition to this criterion other aspects were taken into account, one of the kind of chronological, referring to the founding date of a community and other about a geographical nature, related to temperature, height above mean sea level, predominant economic activity, population and access roads (Flórez 1983). However, the mapped variants correspond to 238 localities (Figure 1) of the total (264), selecting 26 due to their spatial proximity, these remaining localities (26) were nevertheless considered within the sections: *Other responses and Additions*.

The people who answered one or more linguistic surveys (informants) were a total of 2234, mostly native to the place surveyed (Flórez 1983), half of them were between thirty and sixty years old, for every two of the men a woman was questioned and about a fifth of the total was illiterate. As a result of the investigation, six volumes were published “each one measuring 50 x 35 cm (Figure 1 - Upper right corner), containing 1696 plates with 1523 maps of linguistic, ethnographic or mixed information, text additions, photographic material and illustrations” (Bonilla & Bernal Chávez 2020: 348).

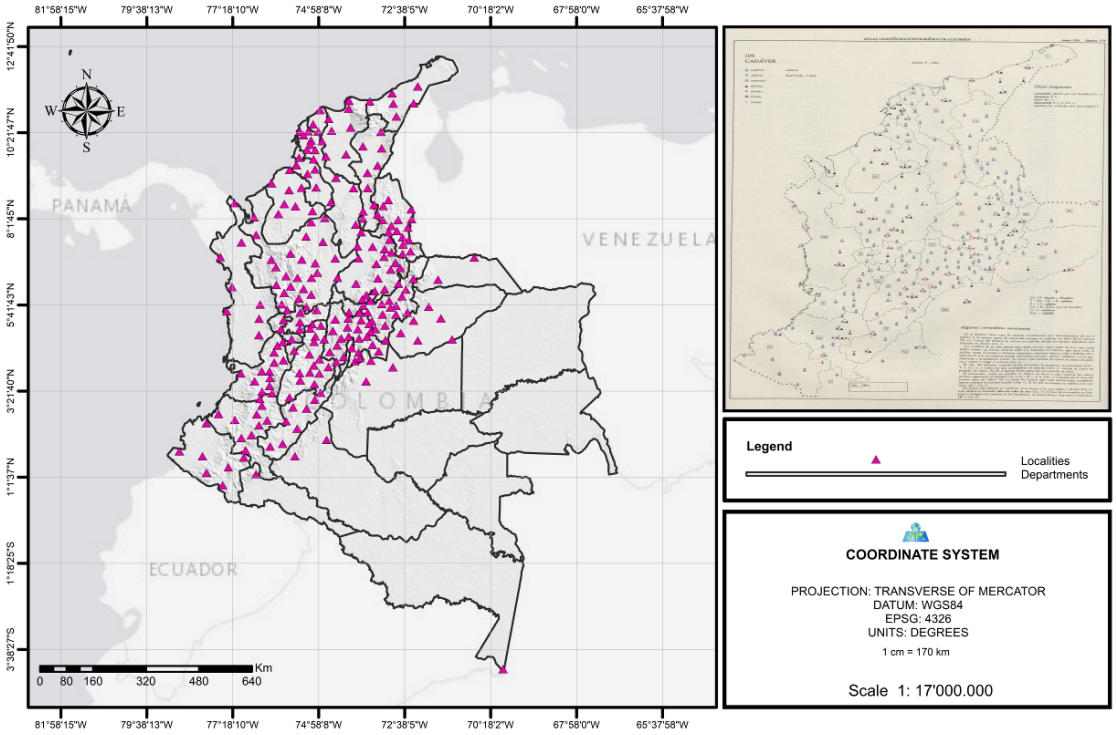

Figure 1. Study area - ALEC mapped localities
Data: Prepared from data from Instituto Caro y Cuervo

The most recent development of the ALEC was carried out around 2015 by the research group NIDE (Nucleus of Research in Spatial Data) of the Faculty of Engineering of the Francisco José de Caldas District University of Bogotá - Colombia, that was invited to participate in the Interactive ALEC project, through the Corpus and Computational Linguistic Research Group (LICC) of the Instituto Caro y Cuervo, this with the purpose of creating and implementing an online Geographic Information System as well as an ALEC Web[2] that would allow to conserve, systematize and disseminate the Linguistic-Ethnographic Atlas of Colombia (ALEC) based on robust technological development (Bonilla et al. 2020; Rocha et al. 2018). Currently both developments can be consulted online.

[2] Acces to digital tools: GIS ALEC (restricted to intranet): http://atlasweb.caroycuervo.gov.co Digital ALEC: http://alec.caroycuervo.gov.co/alec/.

4. Data and Methods

4.1 Data

The data are those natural limits referring to geographical and linguistic phenomena, as well as some metadata of information taken within the development of the ALEC surveys (Table 1), such information is:

Data	Type	Entity
Economic activities	Categorical	Metadata ALEC - Instituto Caro y Cuervo
Height above mean sea level	Numeric	Geographic Institute Agustín Codazzi (IGAC)
Access roads	Categorical	Metadata ALEC - Instituto Caro y Cuervo
Temperature	Categorical	Institute of Hydrology, Meteorology and Environmental Studies (IDEAM)
Agroclimatic suitability	Categorical	Institute of Hydrology, Meteorology and Environmental Studies (IDEAM)
Water availability index	Categorical	Institute of Hydrology, Meteorology and Environmental Studies (IDEAM)
Seismic Hazard Risk	Categorical	Colombian Geological Service
Precipitation	Categorical	Institute of Hydrology, Meteorology and Environmental Studies (IDEAM)
Mean Linguistic Distance	Numeric	Result IRI - Instituto Caro y Cuervo
Geographic Distance	Numeric	Vincenty's formula result - Instituto Caro y Cuervo

Table 1. Data to be used within the development of the model

In principle, the localities or places visited by the ICC researchers were selected, having as a principle that they had the greatest number of responses, those localities initially mapped in the ALEC printed volumes (238 localities) were chosen for this reason. Among the metadata for each locality, those related to economic activities and height above mean sea level were collected. Likewise, with respect to geographic information, open data from state entities were selected (Table 1), taking among them the agroclimatic aptitude, defined as the capacity of a certain place to produce a specific crop (Salvatore et al. 2009), the levels used to this research were taken from the official IDEAM classification, data in the time interval 1981 - 2010, close to the

completion of the survey under which the ALEC emerged. Similarly, precipitation data were incorporated, representing the spatial distribution of the annual mean over the Colombian territory in the same time interval (1981-2010). It should be noted that the information of these geographical phenomena of a Categorical type (Table 1) was reclassified to numerical information using dichotomous variables, this with the purpose of being quantitatively integrated into the spatial model.

Regarding linguistic similarity, in this development the Relative Index of Identity was used, the linguistic data for its calculation were taken from the lexical maps of volume III of the ALEC. Information that after being refined from ethnographic, phonetic elements and supplements allowed to select *100 maps*, in this investigation the mean value of similarity was determined to carry out the subsequent calculations, it is noteworthy that the use of the mean is used for visualization and grouping by Goebl (1987), however, in this development it was used to reduce the number of values to the number of localities (Table 2). Finally, we have the mean geographical distance, for its calculation the Vincenty algorithm (Esenbuğa et al. 2016) was used, an algorithm that allowed obtaining a square matrix that relates the mean distance value for each of the localities.

Data	**Greater appearance**		**Lesser appearance**	
Economic activities	Farming (198)		Industry (5), Commerce (5)	
Access roads	Road (208)		Railway (1), Aerial (1)	
Temperature	Between 26 and 28 degrees (64)		Between 20 and 22 degrees (9)	
Agroclimatic suitability	Extremely dry (112)		Moderately humid (12)	
Water availability index	Semi-dry (66)		Humid (4)	
Seismic Hazard Risk	High (143)		Extremely low (1)	
Precipitation	Between 1000 and 1500 (73)		Between 7000 and 9000 (3)	
Data	**Mean**	**Maximum**	**Minimum**	**Standard deviation**
Mean Linguistic Distance	0,465078	0,617272	0,348222	0,048371
Geographic Distance	385,78	1278,7	224,79	132,83
Height above mean sea level	1266,24	3833,12	80	1064,5

Table 2. Characteristics of the data to be used within the developmen

4.2 Methods

4.2.1 Spatial model design

To generate an optimal design of the spatial model, it is important to know the geographical relevance that each of the localities will have, that is why the matrix of spatial weights W was generated, a matrix that represents the contiguity of the localities with shared vertices and arcs (Anselin & Rey 2014) (Figure 2), bearing in mind the relationship of linguistic variation with geographical distance (Figure 2).

The connectivity histogram (Figure 2) generated from the spatial weight matrix shows the number of observations for each cardinality value in relation to each location, with a general symmetric pattern and center at 7, this represents 31.9% of the total, the minimum value of neighboring observations at a point is two, and a maximum of nine. Once the spatial weights were defined, in this stage the concentration or dispersion of the values of each variable within the territory was determined, as well as its unique relationship with the mean linguistic distance, for which the spatial autocorrelation index I of Moran was used, for its calculation a total of 999 permutations were made, in such a way that they had greater technical reliability, additionally the relationship between each of the variables and the mean linguistic difference was made, with the I index of bivariate Moran, making it clear how there is a consistent relationship between geographic variables and metadata about the territory under study.

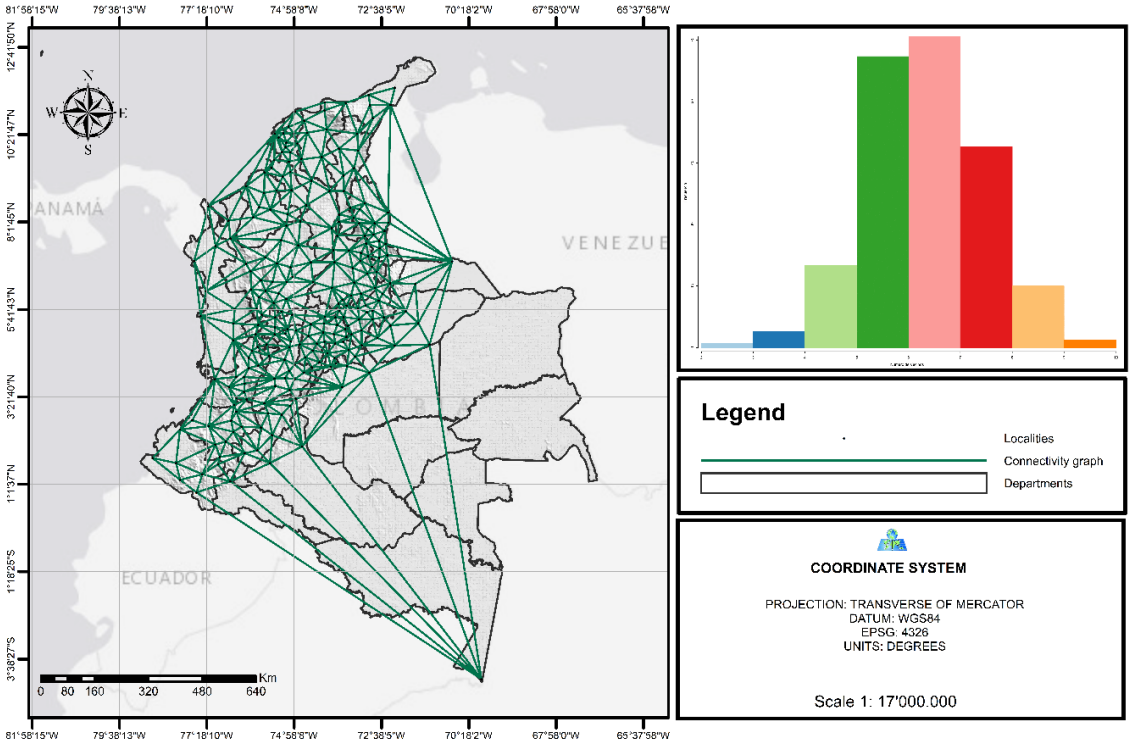

Figure 2. Connectivity map and histogram of spatial weights

After analyzing the spatial dependence for the different geographic variables, the next step was about identifying the causes of it, this with the purpose of integrating and analyzing it within a spatial model; according to Anselin (1988) this spatial relationship can be identified from two facts: the existence of measurement errors in the observations of neighboring geographic units and the existence of several spatial interaction phenomena, for the selection of one or the other Anselin (2005) proposed different tests of Lagrange multipliers, these tests include five alternatives to find the best model for the input information, the process results (Table 3) showed that the values are significant and robust for all the tests being the p-value less than α (α = 0.05) rejecting the null hypothesis of spatial non-dependence; in this case, the LM-LAG test is more significant than LM-ERR as well as the most significant robust test is for the spatial lag, the results allow to consider this model as the most optimal, however, the construction and implementation of additional models, to verify such assumption:

Test	Value	Degrees of freedom	p-value
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

Table 3. Values of spatial dependency tests

4.2.2 Construction and implementation of the spatial model

The construction and implementation of different spatial models (Table 4) were done in the R statistical software, the purpose of this activity is to contrast the hypothesis product of the spatial dependence tests (Table 3), which states that the optimal model in geographic terms is the spatial lag model and/or spatial lag model.

Generated models
Basic regression model estimated by ordinary least squares (MBRL)
Mixed autoregressive model of spatial regression or model of spatial lag.
Spatial regression or spatial lag model and variables with greater significance (VMS).
Estimation of the mixed spatial regression model with random autoregressive perturbations. (SARAR).
Mixed spatial regression model with autoregressive random disturbances (SARAR) for significant variables (VMS).

Table 4. Built models

These spatial models allow to systematically identify levels of dependence for a variable from the different values that it takes in neighboring geographic units, the models built were based on the generation of multiple iterations of integration and discarding of independent variables, considering both the matrix of spatial weights and the Spatial Correlation Index (Moran's I).

4.2.3 Selection and validation of the spatial model

The selection of the optimal model was made based on the coefficient of determination, the analysis of the error terms: heteroscedasticity, normality, and spatial autocorrelation, and finally the analysis of the estimated terms (Table 5):

	Model MBRL	**Spatial Lag Model**	**Spatial Lag Model VMS**	**Model SARAR**	**Model SARAR VMS**
Determination coefficient					
R^2	0.525				
R^2 **adjusted**	0.4227				
Pseudo R^2		0.6109	0.58062	0.62926	
ρ (spatial dependence)		0.54529	0.61495	0.77318	
Normality test - Residuals					
Shapiro-Wilk	0.986	0.990	0.993	0.989	0.990
p-value	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-value	0.018	0.052	0.167	0.042	0.096
Heteroscedasticity - Residuals					
Breusch-Pagan	65.187	69.634	25.583	62.611	27.053
p-value	0.0124	0.0046	0.0602	0.0212	0.0409
Spatial correlation - Residuals					
Moran's I	43.583	-1.631	-12.771	-0.25413	-0.051
p-value	1E-05	0.2448	0.2016	0.7994	0.9587

Table 5. Test statistics of generated models

Table 5 shows how the coefficient of determination, classical R^2, adjusted R^2, and pseudo R^2 have values that represent a medium-high prediction success, with an explained variance that mostly exceeds 50% ($R^2 >$ 0.5). Likewise, the analysis of residuals for each model when executing the normality, correlation and heteroscedasticity tests allowed selecting the model with the greatest statistical cohesion: *the VMS spatial lag model*. For this model, the Shapiro-Wilk test has a calculated statistic of 0.993 with a p-value of 0.35 (significance level α equal to 0.05), in this case the null hypothesis is that the distribution of the data is normal, as the p-value greater than alpha (p > α) does not reject the null hypothesis of normality, therefore it is probable that the residuals have this distribution, the D'Agostino's K^2

test reaffirmed this hypothesis, thus can confirm that according to both tests the residuals come from a normal distribution (Figure 3).

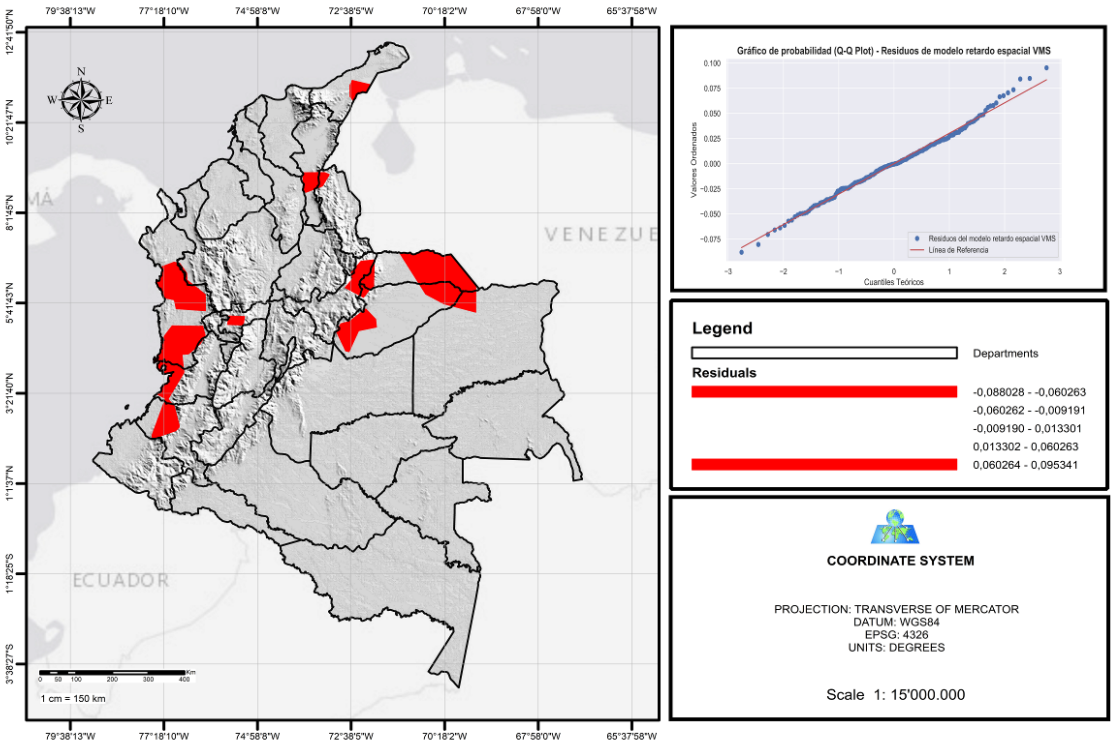

Figure 3. VMS spatial lag model residuals

For the detection of homogeneity or spatial heteroscedasticity within the residuals, the Breusch-Pagan test was used, whose statistic calculated for the selected model was 25.583 with a p-value of 0.06018, the null hypothesis states that there is homoscedasticity, being the value of α greater than 0.05 (α = 0.05) the null hypothesis is accepted, therefore, there is homoscedasticity or homogeneity in the selected estimated model.

Finally, in order to contrast the hypothesis that the residuals are located randomly in space, there is the I Moran's statistic, which calculates the association of similar or dissimilar values between neighboring regions (Vayá & Moreno 2000). Given that in the generated spatial models the p-value is greater than alpha (α = 0.05) the null hypothesis cannot be rejected, it is likely that the residuals for these models are

randomly distributed, that is, they are not spatially correlated. The visualization of the residuals (Figure 3) was done with the use of Voronoi or Thiessen polygons, it shows the residuals considering their level of confidence, for this model this confidence level is equal to ± 2σ, where sigma is σ = 0.03013, that is ± 2σ = 0.06026.

5. Results

The univariate and bivariate spatial correlation was presented as an indicator of geographic dependence between the variables that represent the observations within each locality, Moran's I index showed how the Agroclimatic suitability (Table 6) variable and its categories: Extremely dry and Slightly humid have a significant value; according to the bivariate spatial correlation test, there is a negative dependence when it comes to the relationship between the slightly humid category and the mean linguistic distance, this category includes a total of 33 localities located in 10 departments, most of them in the central zone (Andean region).

Variable	Global Moran's I	p-value	Bivariate Moran's I (Mean Linguistic Distance)	p-value
Agroclimatic suitability				
Slightly humid	0.105	0.006	-0.113	0.001
Extremely dry	0.659	0.001	0.063	0.015
Moderately humid	0.217	0.001	-0.121	0.001
Precipitation				
Between 2000 and 2500	0.118	-0.013	-0.091	0.002
Between 500 and 1000	0.492	-0.128	-0.038	0.078
Temperature				
Between 20 and 22 degrees	0.090	0.018	-0.058	0.025
Between 26 and 28 degrees	0.510	0.001	0.242	0.001
IDH				
Half humid	0.262	0.001	0.0033	0.454
Semi-dry	0.444	0.001	-0.261	0.001

Access roads				
Bridle path	0.065	0.090	0.074	0.006
Highway and Railway	0.084	0.024	0.023	0.182
Cartable roads	0.420	0.001	-0.206	0.001
Highway	0.147	0.003	-0.041	0.070
Geographic Distance				
Geographic Distance	0.798	0.001	0.269	0.001

Table 6. Global and bivariate Moran's I Index of spatial objects

In relation to the water availability index variable (IDH), the highest spatial correlation is presented in the Semi-dry category (Table 6), with a p-value = 0.001, which implies a significant relationship, within the model selected in an equivalent way, it arises that a locality with these conditions may present less linguistic distance with respect to the total.

A notable case corresponds to the variable Economic activities (Table 7), this because the Mining category, according to the results, exerts a great influence when the mean linguistic difference is high. It is noteworthy that the localities with this characteristic are located in the northwest, center-west and southwest of the territory, in places where there is a high average linguistic distance or close to these regions (Colombian Pacific), in such a way that the hypothesis of the relationship between high linguistic distance and areas with high Afro-descendant influence can be raised (Alba 1979, Barbary & Urrea 2002, Romero 1991, Sharp 1970), likewise it can be stated that the model responds and verifies such hypothesis in its results, however, it should be mentioned that the sample is small to take this result as decisive.

The models generated throughout this research followed the theory proposed in spatial econometrics by Anselin (1988) where the causes of spatial autocorrelation for the mean linguistic distance are the product of the existence of the interaction of spatial phenomena in contiguous geographic units, phenomena that were integrated into geographic models, evidencing the existence of interdependence and multidirectionality in spatial data (Pérez-Pineda 2006). Additionally, the construction methodology of the estimated models was carried out considering the proposal by

Anselin & Rey (2014), in such a way that initially the linear model (MBRL) was generated by ordinary least squares (OLS), the respective tests of spatial dependence were applied, and finally statistical coherence tests were used for the selection and validation of the best model (Table 5).

Variable	**Estimated**	**Std. Error**	**z-value**	**Pr(>|z|)**
Intercept	-0,0041	0,049	-0,0836	0,9334
Geographic Distance				
Log (Geographic Distance)	0,0236	0,0082	2,8789	0,0039
Economic activities				
Mining	0,0254	0,012	1,9709	0,048
Agroclimatic suitability				
Extremely dry	0,0091	0,0052	1,7447	0,081
Slightly humid	-0,0068	0,0063	-1,084	0,2783
Moderately humid	0,0259	0,0125	2,0755	0,0379
Precipitation				
Between 2000 and 2500	-0,011	0,0062	-1,77	0,0767
Between 500 and 1000	0,0064	0,0077	0,8244	0,4097
Seismic hazard				
Moderately high	0,0068	0,0061	1,1166	0,2641
Access roads				
Bridle path	0,0711	0,0202	3,515	0,0004
Cartable roads	0,0413	0,013	3,1595	0,0015
Highway	0,0443	0,0165	2,6722	0,0075
Highway and Railway	0,0683	0,0162	4,1977	0,00002
Temperature				
Between 20 and 22 degrees	-0,0058	0,0098	-0,5972	0,5503
Between 26 and 28 degrees	0,0021	0,005	0,4324	0,6654
IDH				
Half humid	-0,0203	0,0108	-1,8735	0,0609
Semi-dry	-0,0086	0,005	-1,7133	0,0866
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				

Table 7. Spatial lag model for the variables with greater significance

Likewise, within the spatial model construction stage, additional models were also generated and implemented, such as the mixed autoregressive model of spatial regression and the mixed model of spatial regression with autoregressive random

disturbances (SARAR) for both the total of variables and for those of greater significance (VMS), however, the selection of the optimal model was made based on different test statistics (Table 5).

The selected model was the spatial lag model (Table 7), this shows how the coefficient ρ, the one that reflects the spatial dependence inherent in the data is equal to 0.61, in such a way that there is a medium-high positive effect between the information object of study; similarly, the explanatory variables were subjected to the Wald test (Table 7), allowing to contrast the hypothesis about the coherence of the model and the value that the variables add to it, in this case a significant test, discarding the null hypothesis of non-coherence.

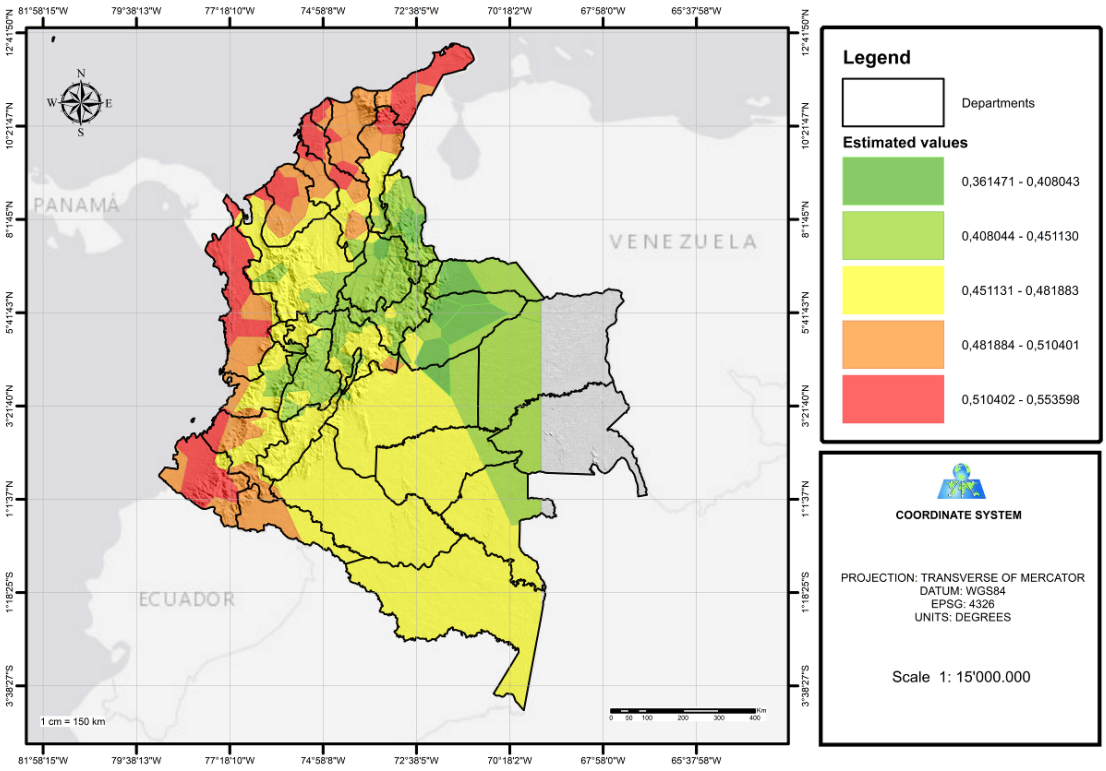

Figure 4. Estimated values with the VMS spatial lag model (model selected)

It is noteworthy that the selected model includes the geographic variables, Geographic distance, Agroclimatic suitability, Precipitation, Threat due to seismic hazard, Temperature and finally the Water Availability Index (HDI), geographic

variables that explain linguistic distance, one of the highest influence is the agroclimatic suitability, and its category Moderately humid, with a positive β value and a medium level of significance (Table 7), in such a way that it can be hypothesized that areas located in the southwest of the Colombian territory, between the departments Caldas, Cauca, Chocó, Huila, Tolima and Valle del Cauca, which present this geographical condition, may have a greater mean linguistic distance in relation to the rest of the territory. Likewise, the HDI variable and its Semi-dry category greatly influence the selected model, its relationship with the linguistic distance is inverse, where the values are concentrated in the central part of the territory (Andean Region), a region whose mean linguistic distance is less with respect to the rest of the geographic space under study.

According to the results obtained, it can be concluded that the area with the lowest mean linguistic distance with respect to the total is the central area (Andean region), followed by the southeast and the western (Pacific), as mentioned, there are both geographic variables and groups of metadata that can influence the dialect variation of the Colombian territory (Figure 4), at a historical level it can be argued that economic activities and the ease or difficulty of mobilization (access routes) can influence this phenomenon, however there are additional social phenomena involved, such is the case of the demographic phenomenon and internal migration registered in the 20th century (Colmenares et al. 1982; LaRosa & Mejía 2013; Osorio Baquero 2019). According to LaRosa & Mejía (2013) the main population, mestizo and white, today is mostly socially and culturally urban, however at the time of the survey, the family affiliation with the countryside was probably greater, a notable difference with the indigenous and Afro-Colombian population, the majority of which live in rural areas that, according to LaRosa & Mejía (2013), are recently colonized, and finally, it should be noted that since the time of the conquest, the region of greatest preference for the settlement of the population was the Andean region, followed by the Caribbean region, whose cultural differences are marked, in such a way that the selected model adjusts to it.

6. Discussion and conclusions

The main evaluation explained by the spatial autocorrelation index (Moran's I) suggested that the variables: economic activity, agroclimatic suitability, precipitation, water availability index and access roads show the highest bivariate spatial dependence in relation to the calculated linguistic distance, that is, there is correspondence in the values of a variable (geographic variable or variable product of metadata) in a geographic space that can be partially explained based on the value of another variable (in this case the linguistic variable) in neighboring spaces (Goodchild 1987), the result of this geographic dependence diagnosis was quantitatively verified within the multiple spatial models generated.

The model with the highest statistical consistency in the estimated values, as well as in the error terms, was the mixed autoregressive spatial regression model or the spatial lag model for the variables with greater significance (VMS), a model that according to the proposed spatial dependency contrasts by Anselin (2005) allowed to reject the null hypothesis of non-existence of geographic correspondence, it includes a total of eight variables, being the most influential, the categories of the variable Access roads, partially validating the hypothesis of Bonilla (2019) about these routes as a medium of linguistic diffusion, in this sense it should be noted that the concentration of national road infrastructure was in the Andean region, and only towards the 1960s the proportion of highways increased in departments that had not been taken into account (Pachón 2006), in this way both the proposed model and historical facts are proof of the concentration not only of language in the center of the territory but of its inherent spatial relationship with economic activities and mobilization and/or road communication processes.

Although it was mentioned how the geographical variables have a relationship of spatial dependence with the linguistic distance expressed quantitatively through Moran's I index, the proposed model verified this hypothesis, allowing to conclude that there is a measurable and/or quantifiable relationship between lexical and geographic

variables, whose main categories correspond to variables such as agroclimatic suitability, precipitation, water availability index and geographical distance.

Finally, the developed spatial model adjusts to economic and sociodemographic phenomena presented at the beginning of the 19th century, the incorporation of geospatial mechanisms and computer science algorithms made possible the generation of new knowledge, demonstrating the quantitative relationship between geographic space and linguistic phenomena, however it is necessary for further research to collect and analyze information in unexplored regions by the ALEC researchers, as well as updating the information in the places visited for the generation of a more in-depth and recent analysis.

References

ALBA, J. G. M. de (1979) "Estudios sobre un área dialectal hispanoamericana de población negra. Las tierras bajas occidentales de Colombia", Bogotá: Instituto Caro y Cuervo, 17, 352-357.

ANSELIN, L. (1988) *Spatial Econometrics: Methods and Models*, Nueva York: Springer Dordrecht. <https://doi.org/10.1007/978-94-015-7799-1>

ANSELIN, L. (2003) "Spatial Econometrics", in Badi H. Baltagi (ed.), *A Companion to Theoretical Econometrics*, Oxford: Blackwell Publishing Ltd., 310-330 <https://doi.org/10.1002/9780470996249.ch15>

ANSELIN, L. (2005) *Exploring Spatial Data with GeoDa: AA Workbook*, Center for Spatially Integrated Social Science, Urbana-Champaign: University of Illinois.

ANSELIN, L. & S. J. REY (2014) *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*, Chicago: GeoDa Press llc.

ANSELIN, L. & O. SMIRNOV (2006) "Efficient Algorithms for Constructing Proper Higher Order Spatial Lag Operators", *Journal of Regional Science*, *36,* 67-89. <https://doi.org/10.1111/j.1467-9787.1996.tb01101.x>

BARBARY, O. & F. URREA (2002) "La población negra en la Colombia de hoy: Dinámicas sociodemográficas, culturales y políticas", *Estudos Afro-Asiáticos*, 25 (1), 9-21. <https://doi.org/10.1590/S0101-546X2003000100002>

BOHÓRQUEZ, I. A. & H. V. CEBALLOS (2008) "Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales", *Ecos de Economía: A Latin American Journal of Applied Economics*, 12(27), 9-2.

BONILLA, J. (2023) "Superdialectos, dialectos y subdialectos del español de Colombia", *Lexis*, 47 (2), 536-564.

BONILLA HUÉRFANO, J. S. (2019) *Propuesta de división dialectal del español de Panamá desde el análisis dialectométrico del nivel fonético del ALPEP y Propuesta de división dialectal del español de Colombia con base en el análisis dialectométrico de datos léxicos del ALEC*, Colombia: Instituto Caro y Cuervo, Facultad Seminario Andrés Bello.

BONILLA, J. E., R. Y. RUBIO LÓPEZ, A. L. LLANOS CHÁVEZ, D. E. BEJARANO BEJARANO & J. A. BERNAL CHÁVEZ (2020) "Proyecto de digitalización y nuevas perspectivas del Atlas Lingüístico-Etnográfico de Colombia", in Á. Gallego & F. Roca (eds.), *Dialectología digital del español*. *Verba Anuario Galego de Filoloxía*. Anexos, 80, 13-28. <https://dx.doi.org/10.15304/9788418445316>

CAMPOY, J. M. H. (1999) *Geolingüística: Modelos de interpretación geográfica para lingüistas*, Murcia: Universidad de Murcia.

COLMENARES, G., A. CORRADINE ANGULO, J. FRIEDE, F. GIL TOVAR, J. JARAMILLO URIBE, J. PALACIOS PRECIADO, G. REICHEL-DOLMATOFF & M. T. C. CRISTINA (eds.) (1982[2]) *Manual de historia de Colombia*, Bogotá, Colombia: Instituto Colombiano de Cultura; Procultura.

DUBERT-GARCÍA, F. & X. SOUSA (2016) "On quantitative Geolinguistics: An illustration from Galician Dialectology", *Dialectologia*, *Special Issue VI*, 191-221. <https://doi.org/10.1344/DIALECTOLOGIA2016.2016.11>

ESENBUĞA, Ö. & E. ÇOLAK (2016) "Comparison of Principal Geodetic Distance Calculation Methods for Automated Province Assignment in Turkey", presented at the conference *16th International Multidisciplinary Scientific GeoConference SGEM 2016*, Albena, Bulgaria, June 30 to July 6, 2016.

FLÓREZ, L. (1983) *Manual del atlas lingüístico-etnográfico de Colombia*, Bogotá: Instituto Caro y Cuervo.

GOEBL, H. (1987) "Points chauds de l'analyse dialectométrique: Pondération et visualisation", *Revue de linguistique romane*, 51, 64-118 <http://doi.org/10.5169/seals-399808>

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

GOODCHILD, M. F. (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>

HEERINGA, W. J. (2004) *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, Doctoral thesis in Humanities Computing, Groningen: University of Groningen Library.

LABOV, W. (1994) *Principles of Linguistic Change, Vol. 1, Internal Factors*, Oxford: Blackwell.

LAROSA, M. J. & G. R. MEJÍA (2013) *Historia concisa de Colombia (1810-2013)*, Bogotá: Universidad del Rosario.

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

OSORIO BAQUERO, I. (2014) "Breve reseña histórica de las vías en Colombia", *Ingeniería Solidaria*, Vol. 10, No. 17, 183-187 <https://doi.org/10.16925/in.v10i17.880>

PACHÓN, A. (2006) *La infraestructura de transporte en Colombia durante el siglo XX*, Bogotá: Fondo de Cultura Económica.

PÉREZ-PINEDA, J. (2006) "Econometría espacial y ciencia regional", *Investigación Económica*, *65* (258): 129-160.

ROCHA, S. L. A., J. E. BONILLA, J. BERNAL, C. DUARTE & A. RODRÍGUEZ (2018) "Design and implementation of the web linguistic and ethnographic atlas of Colombia", in *Proceedings of the ICA*, 96(1), 1-4. <https://doi.org/10.5194/ica-proc-1-96-2018>

ROMERO, M. D. (1991) "Procesos de poblamiento y organización social en la costa pacífica colombiana", *Anuario Colombiano de Historia Social y de la Cultura*, 18-19 (January), 9-31.

SALVATORE, M., A. Kassam & A. C. Gutiérrez (2009) *Metodología de Evaluación de Aptitud de Tierras*, 9, 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 dialectométrie dans l'Atlas linguistique de la Gascogne", *Revue de linguistique romane*, 37, 1-24.

SHARP, W. F. (1970) *Forsaken but for Gold: An Economic Study of Slavery and Mining in the Colombian Choco, 1680-1810*, Chapel Hill, North Carolina: University Microfilms Internacional.

VAYÁ, E. & R. MORENO (2000) *La Utilidad de la Econometría Espacial en el Ámbito de la Ciencia Regional*, Barcelona: Ediciones Universidad de Barcelona.

VILALTA, C. (2005) "Cómo enseñar autocorrelación espacial", *Economía, Sociedad y Territorio*, 5 (18), 323-333.

WIELING, M. (2012) *A Quantitative Approach to Social and Geographical Dialect Variation*, Groningen: University of Groningen Library.

WIELING, M. & J. Nerbonne (2015) "Advances in Dialectometry", *Annual Review of Linguistics*, *1*, 243-264. <https://doi.org/10.1146/annurev-linguist-030514-124930>

YRIGOYEN, C. C. (2002) *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*, Madrid: Consejería de Economía e Innovación Tecnológica, Comunidad de Madrid.