KU LEUVEN

FACULTY OF SCIENCE

Identification and Agreggation of Regional Linguistic Variation in Colombia

Methods of Corpus Linguistics

Daniel Gerardo GIL SANCHEZ
daniel.gilsanchez@student.kuleuven.be
MSc Statistics

Supervisor: Prof. Dirk Speelman

Academic year 2018-2019

Contents

1 Introduction

Dialects are varieties of a language that are characteristic of specific groups of speakers. These groups can be defined by social class, ethnicity, geographical areas, among others, but also by different combinations of intrinsic variables that affect the way people usually talk.

Studies on Linguistic variation, which purpose is to identify dialects, are often conducted with surveys or interviews, where the probability that a person expresses himself/herself in a natural way is actually really small, producing results that can be somewhat biased. For this reason, different kinds of methodologies have been used to collect raw linguistic data without the need of having an interviewer or someone (something) that may possibly change the behavior and the way a person talk.

One of them is exploiting the use of social media, which are basically different platforms where people interact with each other and express themselves in the most natural way. Twitter in particular, has been chosen in different studies because it allows retrieving a large amount of data easily with the bonus that is possible to localize geographically where the information was produced (see Huang et al., 2016; Goncalves and Sánchez, 2014)

The purpose of this study is to identify regional linguistic variation in Colombia using a statistical methodology proposed by Grieve et al. (2011), that consists in a combination of spatial and multivariate analysis. This methodology is applied in 34, continuously measured, alternation variables collected from a 20 million-words corpus of tweets representing the most important cities in the country.

In the following, a description of how the data was retrieved and organized is presented. Then the statistical analysis is conducted and discussed to finally get conclusions about the number and location of dialect regions found in the country. It is important to mention that the person who did this document is not an expert on Linguistics and therefore some results are explained in a statistical way instead of a linguistic way.

2 Corpus Compilation

The corpus analyzed in this study consists of 1.6 million geo-tagged tweets and approximately 21 million words in Colombia, representing the capitals of each department and the cities which population is 100.000 inhabitants or more. The corpus has been compiled over the third week of December 2018.

This kind of corpus was selected for several reasons. First, it almost offers real time information in specific geographical locations. Second, it is relatively easy and cheap to get via web-scrapping algorithms. Third, data can be retrieved from almost any place in the country, including those cities that are located in the Amazonas as well in ”Los Llanos” (the plains), which is an ecoregion of the flooded grasslands and savannas biome. Finally, the retrieving process leads to control temporal linguistic variation by getting tweets of a certain time period.

On the other hand, there are some limitations with the use of this kind of data. For instance, it is not possible to know whether a tweet located in a specific city corresponds to a person who lives there or to a person that is visiting the place (tourists, business trips, etc.). Additionally, Twitter data do not allow analyzing demographic

background of informants such as gender, age and socioeconomic status. In some cases, the gender of the informant can be obtained by analyzing the name or username of each tweet, but given that there are over 1.6 million tweets in the corpus, this labor could take days if not more time and there is evidence that this variable may not be significant (see Grieve et al., 2011). Furthermore, since Twitter platform only allows users to write in 140 characters, there is a trend to use contractions of words and emojis to express feelings. Since these contractions are not standardized, it is difficult to know to what word is the user referring to without reading the context. When this happens, the frequency of these words may be underrepresented. Finally, there are criticisms about the uncertainty in socio-demographic representativeness. Huang et al. (2016) and Goncalves and Sánchez (2014) mentioned that on average Twitter users are young people living in urban areas. So, even though such studies identifying user population in Colombia have not been done, it is plausible to say these results also hold in this country.

The retrieving process was performed in R using the function `search_tweets` from the package `rtweets`, because it allows to look for tweets near a specific location given the longitude and latitude coordinates, and a certain radius around it. This method was chosen, instead of using the exact location of each tweet, because it increases the number of tweets collected. It is important to mention that even though the Twitter API offers the information of exact location, only a few number of tweets, approximately 1%, has this information available.

In this sense, the first step was to get the coordinates of each city as well as a good approximation of the radius coverage for the retrieving process. The coordinates for each city were obtained from the web-page `https://www.geodatos.net`, a portal that is known to give precise information. Then, the respective radius was obtained by searching the coordinates (longitude and latitude) of each city on Google Maps and calculating the distance to the farthest point in the urban area.

The retrieving process took place when this information was available, but several problems were found in the data collected. First of all, since the radius considered in all cities only took into account urban areas, there were some cities where the number of tweets collected did not reach the thousand threshold. When this happened, the radius of the buffer considered was extended to get more coverage of rural areas and increase the likelihood of getting more data. Usually, the radius was increased about 10 to 20 kilometers, but there were places, in particular in "Los llanos", where the radius had to be extended up to 150 kilometers (see cities in the east part of the country in Figure 1). The increase in the radius considered in such cities was motivated by the lack of cities close to these locations, so the probability of getting duplicated tweets with other cities was minimal if not zero. Second, there were still some cities with a low number of tweets even when the radius was increased. In these cases, the corresponding information was merged with the closest city (within a 30 km range) under the assumption that people that live in close locations share most of the words they use.

As it is clear in Figure 1, the distribution of the cities in the corpus is not even. There are more data sampled from the west side of the country that on the east side. As a result of this distribution, dialect patterns can be identified with greater confidence in those places with better coverage.

Once the retrieving process finished, a cleaning process was conducted on the

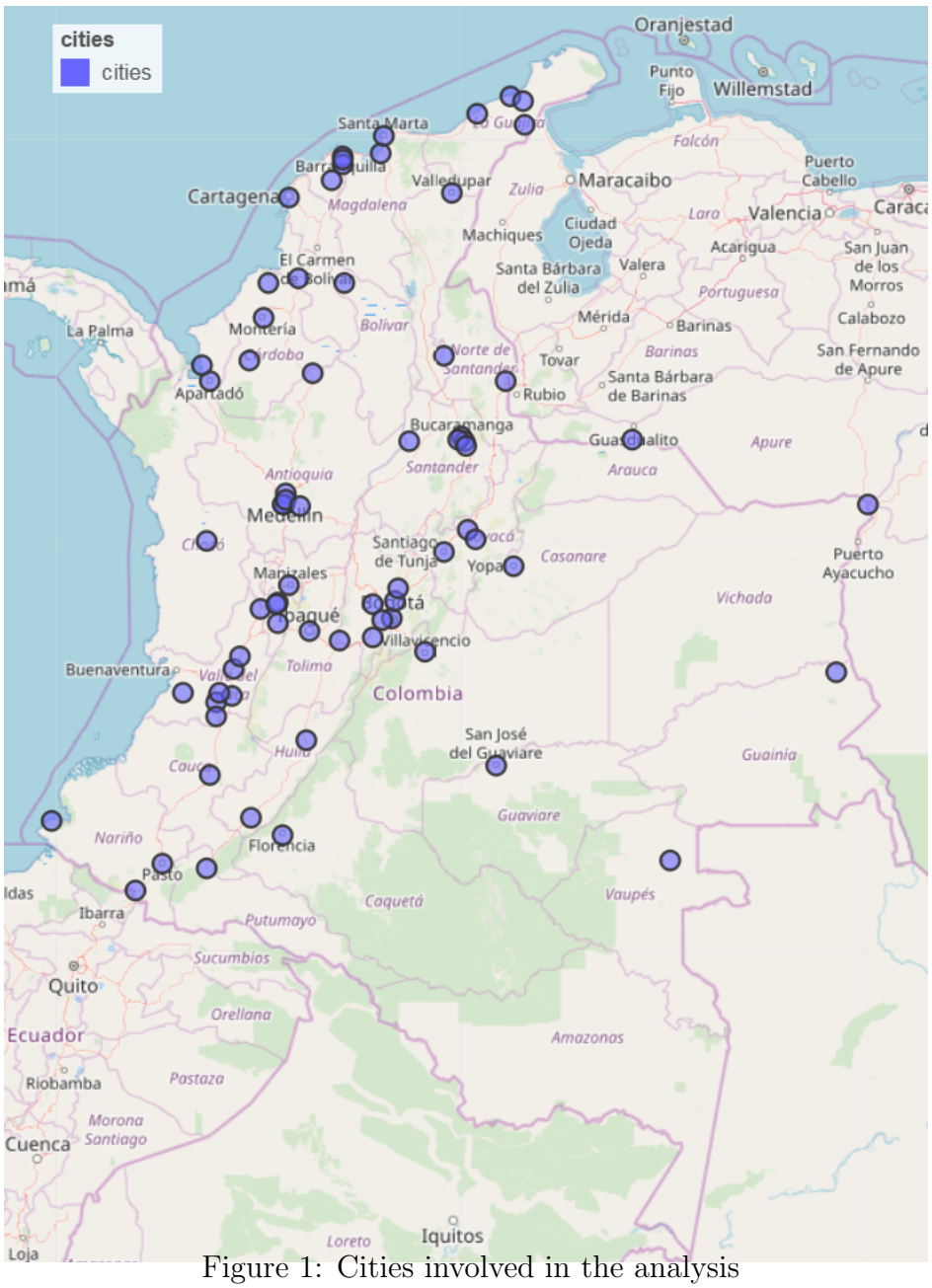


Figure 1: Cities involved in the analysis

data. First, all URLs were removed by using a regular expression that represents the pattern *http://* or *https://*. Then, all mentions were also removed by using a regular expression that identifies words that begin with @. Finally, all emoticons were removed by using a regular expression that represents the pattern $< U + 00000000 >$, where all characters after the plus (+) are alphanumeric (for more detail see scripts in R). Additionally, given that several attempts were conducted to get more data and that each time all tweets were saved, the likelihood of getting duplicated tweets increased each time a new attempt was made. Therefore, within each city, all duplicated tweets were removed.

After the tweets were cleaned, each city subcorpus contained, on average, three hundred ten thousand $(310,000)$ words. As it is mentioned in Grieve et al. (2011), a cutoff of $35,000$ words was selected to obtain reasonable estimates of the alternation

variables. In this sense, there is a lot of variation in the size of city subcorpora where *Mitú* and *Medellin* have the minimum and maximum number of words with $36,880$ and $1,056,395$ respectively. In total, the final corpus has $20,594,975$ words from $1,675,667$ tweets representing 63 cities in Colombia (see Figure 1).

3 Corpus Analysis

As it is mentioned in Grieve et al. (2011), in regional dialectology the linguistic alternation variables most often analyzed are synonymous words or words with equivalent pronunciations. In this study, special focus has been set in synonymous words measured with a quantitative value, representing the relative frequency of one of the variants with respect to all of the variants considered for a specific word.

As it is suggested in Huang et al. (2016) and Grieve et al. (2011), to compute the value of the alternation variable with two variants in a specific location, the proportion of one variant is computed by dividing the frequency of that variant (f_a) by the sum of the frequencies of both variants ($f_a + f_b$), see equation 1. Note that it does not matter the variant chosen to compute the relative frequency, in this study the variant with higher frequency was chosen to be in the numerator.

$$f = \frac{f_a}{f_a + f_b} \tag{1}$$

This formula is the basis for all alternation variables considered in this study and it can be extended to more that two variants by summing the respective frequencies to the denominator.

In regard to the selection of the alternation variables, a list of suitable lexical alternations was compiled using different approaches. First, a list of total synonyms was considered and some variants with high frequencies in the corpus were chosen. Since in the literature the concept of total synonyms is not standardized, in this study a total synonym is when two words have the same meaning in all contexts. Then, some adverbs and prepositions were considered. Finally, the Varilex database, which provides a list of possible words representing the same concepts (Ueda and Takagaki, 1993), was used to complement the list of alternation variables.

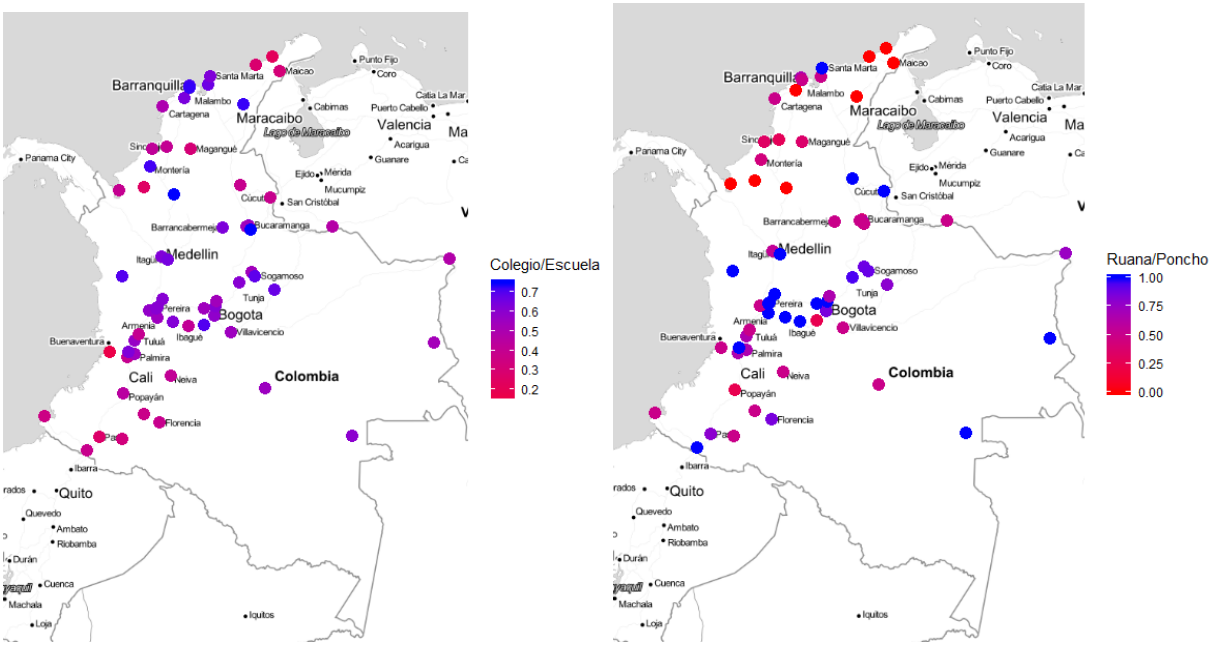
In total 34 linguistic alternations were considered, see Table 1. The function used to count these variables is `freqlist` from the package `mclm` and the file encoding used is "*Latin1*" because of the Spanish accents and the letter *ñ*.

Finally, before conducting the statistical analysis, the alternation variables were mapped individually in the whole country. Figures 2a and 2b shows the proportion of Colegio/Escuela alternation and Ruana/Poncho alternation in all cities considered in the analysis. The first variant means school and people tend use it indistinctly, whereas the second variant is a garment and the use of one variant or the other is mostly influenced by the weather. Figure 2a shows that people in the north and the south of the country used more the second variant: Escuela, while people in the center of the country tend to use the first variant. In most of the variables though, these patterns are difficult to obtain because the proportion is closer to 50% than any bound limit (0% or 100%).

Classification	Variant 1	Variant 2
Adjective	Estimado	Apreciado
	Mejor	Superior
	Lindo	Bello or Bonito
	Inteligente	Pilo
Adverb	Ahí	Allí or Allá
	Aquí	Acá
	Adelante	DelanteEnfrente
	Cerca	Próximo or Cerquita
	Lejos	Apartado
	Ya	Ahora
	Después	Luego
	Antes	Anteriormente
	Aún	Todavía
	Mientras	Durante
	Quizás	Quizá
Noun	Cariño	Afecto
	Ambición	Codicia
	Casa	Hogar
	Colegio	Escuela
	Estudiante	Alumno
	Dinero	Plata
	Impuesto	Gravamen
	Saco	Buzo
	Ruana	Poncho
	Diadema	Balaca
	Bolso	Cartera
	Frasco	Envase
	Esposo	Marido
	Regalo	Obsequio
	Ave	Pájaro or Pájarito
Preposition	Bajo	Debajo
	Hasta	Incluso
Pronoun	Tú	Usted or Vos
	Ustedes	Vosotros

Table 1: Alternation variables

It is important to mention that in some cases the frequency of both variants was zero, meaning that none of the variants were found in the city, which leads to a mathematical indetermination ($0/0$). In these cases, the corresponding variable was manually set to 0.5, giving the same importance to both of the variants.



(a) Colegio/Escuela alternation (b) Ruana/Poncho alternation

Figure 2: Map of two lexical variables in raw values

4 Statistical Analysis

The statistical approach considered in this study for the analysis of regional linguistic variation corresponds to three steps introduced in Grieve et al. (2011). First, each variable is spatially analyzed by measuring spatial autocorrelation to identify significant patterns in regional linguistic variation. Second, the results of the spatial autocorrelation analysis are used to conduct a factor analysis to identify and describe patterns in regional linguistic variation. Finally, the results of the factor analysis are used to identify dialect regions.

4.1 Spatial autocorrelation analysis

Each of the 34 variants is subjected to global and local autocorrelation analysis to identify significant patterns of regional variation. Spatial autocorrelation is a measure of similarity between nearby observations located in geographical areas. Two items are needed to compute it: locations, represented in this study as cities in Colombia and variables, represented by the alternation variables. To test whether there is geographical clustering in the values of each variable, the global Moran's I test statistic is used and to identify clusters with high or low values in the distribution of each variable, the local Getis-Ord $Gi*$ is conducted. See Grieve et al. (2011) for technical detail.

Both test statistics depend on the specification of a *spatial weight matrix* that introduces the spatial information in the calculation of the correlation. This matrix is computed as it is suggested in Grieve et al. (2011), which assigns the value of 1 to all pairs of cities that are within a certain distance and 0 to the remaining cities. This type

of spatial weights is called binary, since it only takes values 0 or 1. To set the cutoff distance for each variable, different cutoff values were tried, from 325 to 1000 kilometers, together with their respective Moran's I statistic. Then, the selection of the final cutoff distance was based on the most significant spatial clustering for each variable, i.e., the distance that produces the highest test statistic value or lowest p-value.

To interpret the significance of spatial clustering, a standardized z-score is obtained from the Moran's I statistic under the assumption of randomization. This z-score is compared to a critical value from the normal distribution given by a significance level with the Bonferroni correction because there are several variables that are simultaneously analyzed. This corrected significance level is 0.00147 because there are 34 alternation variables and the regular significance level is 0.05 ($0.05/34 = 0.00147$). This is a one-tail test because the goal of the analysis is to detect spatial clustering via positive autocorrelation measures. It is important to mention that even though two-tail hypothesis testing can be conducted, the default hypothesis in R is one-tail.

The final results of the global autocorrelation analysis are shown in Table 2, where the final cutoff distance, the Moran's I statistic, the z-score and the p-value are listed. It is important to note that the z-score does not come from a standard normal distribution, but from a normal distribution with specific mean and variance. For this reason, it is recommended to focus on the p-value when looking this table. There are in total 15 variables that present significant global spatial autocorrelation, including Colegio/Escuela (Figure 2a) and Ruana/Poncho (Figure 2b).

Regarding local spatial autocorrelation, Getis-Ord Gi* statistic is a measure of the degree to which a particular location is part of a high or low value cluster. Therefore, the statistic is not only one value to measure the extend of the spatial autocorrelation, but each city has its own value, meaning that for each variable there are 63 Gi* values. These values are also called z-scores and correspond to a normal distribution where a significant negative score indicates that the city is part of a low value cluster, and a significant positive score indicates that the city is part of a high value cluster. Once again, the Bonferroni correction is used, implying that the corrected significance level is 0.00147. Note that in this case it is of interest to judge whether there are positive or negative significant values, so the null hypothesis is two-tailed.

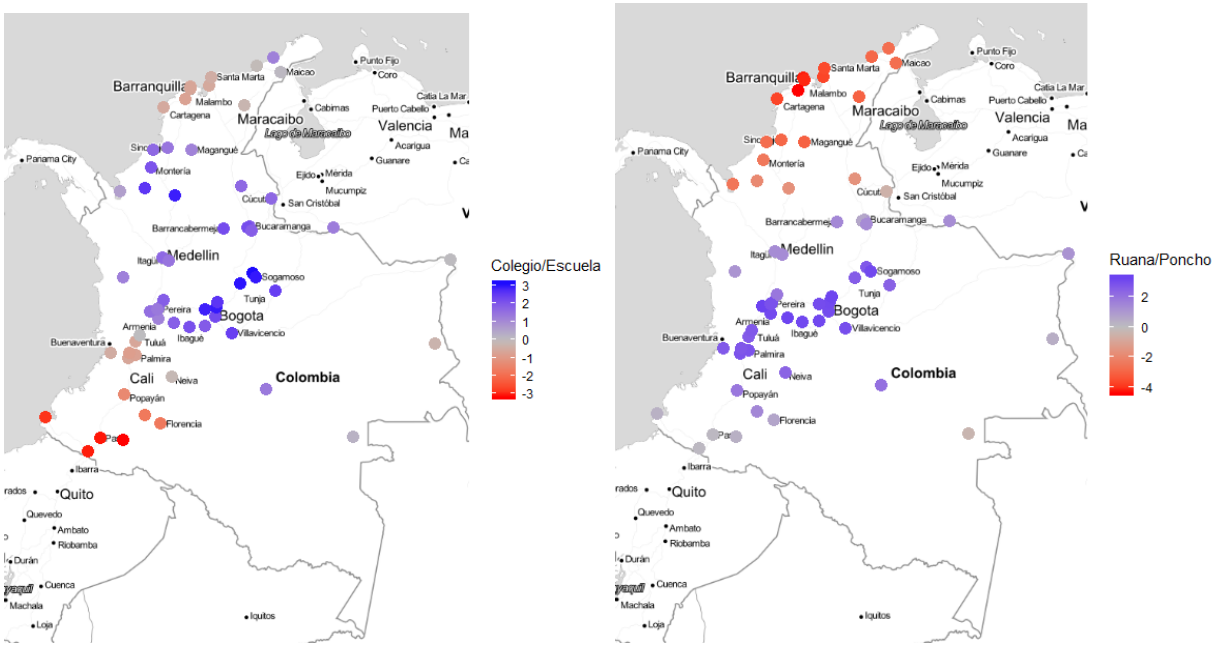
Given the magnitude of values to analyze ($63\ cities * 34\ variants = 2142\ values$), the easiest way to derive conclusions is through a map. Figure 3a and Figure 3b show the Getis-Ord z-scores for Colegio/Escuela alternation and Ruana/Poncho alternation. In both maps, a positive z-score (blue dots) indicates that the location is part of a high value cluster, meaning that the first variant occurs more frequently, whereas a negative z-score (red dots) indicates that the location is part of a low value cluster, where the second variant occurs more frequently. Figure 3a shows that *Colegio* is relatively common in the center of the country, whereas *Escuela* is relatively common in the Southwest. On the other hand, Figure 5 shows that *Ruana* is relatively common in the North, whereas *Poncho* is relatively common in the center together with some cities in the East. Note that these maps are similar to the maps with raw values (Figure 2) but the underlying regional signals are identified by the local autocorrelation.

Variable	Cutoff	Moran's I	z-Score	p-value
Tú/Usted-Vos	425	0.2571	12.5571	0
Ruana/Poncho	350	0.1926	7.4317	0
Inteligente/Pilo	325	0.1633	5.78	0
Adelante/Delante-Enfrente	825	0.026	5.4949	0
Ave/Pájaro-Pájarito	1000	0.0092	5.0498	0
Quizás/Quizá	325	0.1343	4.8357	0
Frasco/Envase	500	0.0654	4.6866	0
Estimado/Apreciado	425	0.083	4.5127	0
Lindo/Bello-Bonito	375	0.0956	4.3886	0
Regalo/Obsequio	325	0.0997	3.8649	0.0001
Colegio/Escuela	375	0.0811	3.761	0.0001
Bajo/Debajo	1000	0.0024	3.7303	0.0001
Dinero/Plata	600	0.0326	3.6398	0.0001
Cerca/Próximo-Cerquita	900	0.0037	3.1217	0.0009
Mejor/Superior	650	0.0185	3.0238	0.0012
Después/Luego	700	0.0125	2.8594	0.0021
Bolso/Cartera	375	0.0524	2.6638	0.0039
Antes/Anteriormente	550	0.0216	2.6154	0.0045
Aún/Todavía	350	0.0376	1.9402	0.0262
Diadema/Balaca	800	−0.0007	1.9182	0.0275
Impuesto/Gravamen	1000	−0.0071	1.8718	0.0306
Ahí/Allí-Allá	475	0.014	1.6588	0.0486
Esposo/Marido	625	0.0037	1.5456	0.0611
Ya/Ahora	375	0.0204	1.448	0.0738
Saco/Buzo	800	−0.0047	1.4356	0.0756
Aquí/Acá	350	0.0161	1.3582	0.0872
Cariño/Afecto	350	0.0174	1.3203	0.0934
Lejos/Apartado	325	0.0094	1.1918	0.1167
Estudiante/Alumno	900	−0.0094	1.1222	0.1309
Hasta/Incluso	350	0.0017	0.6462	0.2591
Ambición/Codicia	850	−0.012	0.5762	0.2822
Mientras/Durante	675	−0.0115	0.423	0.3361
Ustedes/Vosotros	975	−0.0145	0.3893	0.3485
Casa/Hogar	925	−0.0148	0.2698	0.3937

Table 2: Global spatial correlation results

4.2 Factor analysis

The z-scores from the local autocorrelation analysis were used to conduct a factor analysis with the goal of identifying common patterns of regional linguistic variation. According to Grieve et al. (2011), if the raw values were used to conduct the factor analysis, then many patterns of spatial clustering identified would be lost. Varimax rotation or variance maximization was conducted to improve the interpretability of each factor by making variable loadings redistribute over the factors such as each variable



(a) Colegio/Escuela alternation (b) Ruana/Poncho alternation

Figure 3: Local autocorrelation

is correlated with the minimum number of factors possible.

A three-factor solution was chosen because the first three factors explained 75% of the variance in the set of 34 alternation variables. The first factor explained 38.5%, the second factor explained 30.1% and the last one explained 6.3%. As stated by Grieve et al. (2011), the fact that three factors account for $3/4$ of the regional variation shows that there are consistent regional patterns in the dataset.

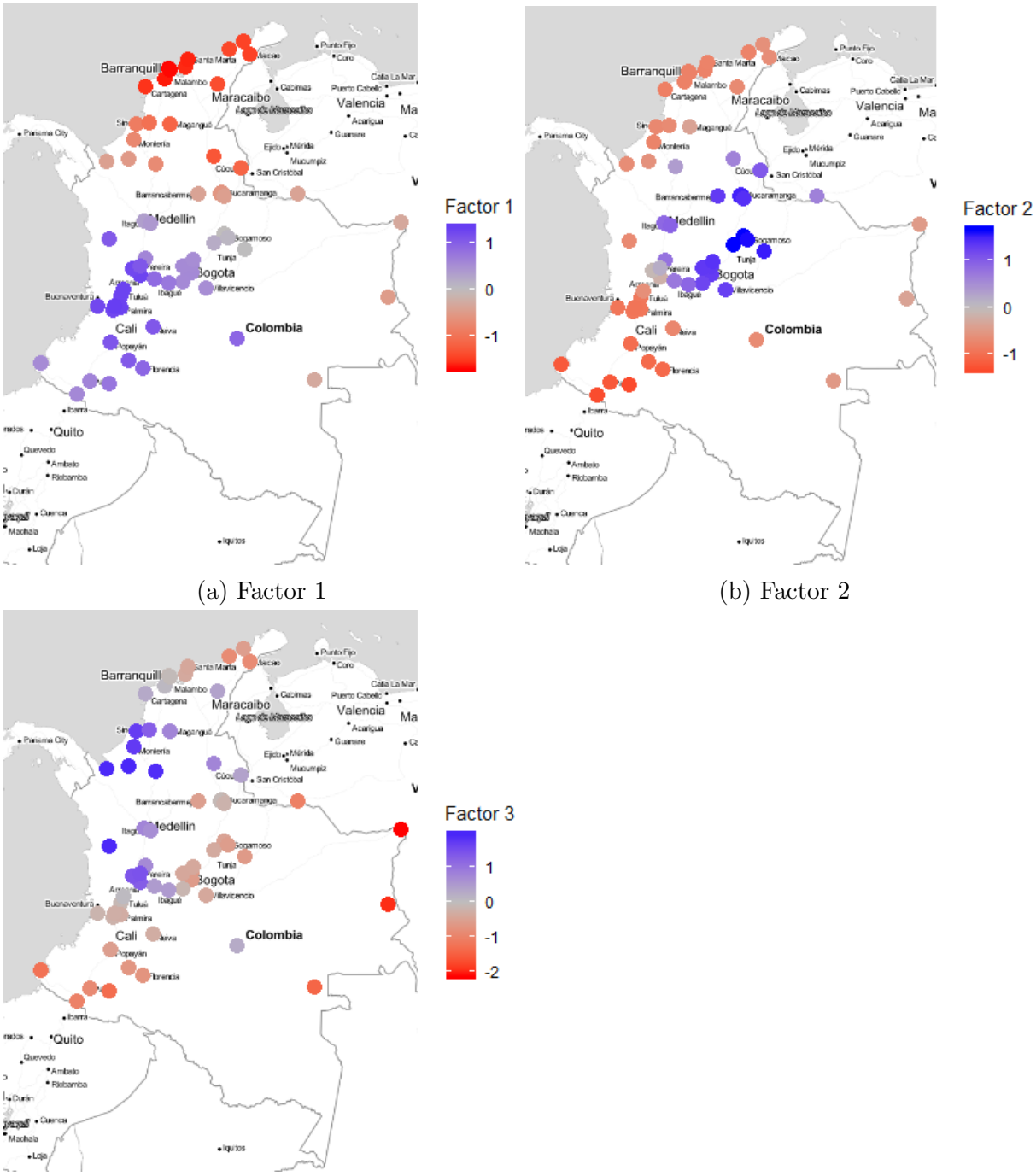
Table 3 shows the loadings of each variable in each factor after the rotation. These loadings represent the degree to which the regional pattern showed by each variable is regarded by each of the factors. In this table, loadings that were smaller that 0.3 were suppressed and the sign of the loadings reveals which variant characterizes the clusters identified by the factor analysis. The uniqueness value of each variable is also listed, where a value higher than 0.8 indicates that the pattern revealed by that variable is not well represented by the three factor solution. In this analysis, all variables are well represented by the three factor solution.

In addition to the table and to visually identify regional patterns, Figures 4a to 4c show the factor scores mapped across the country. Factor 1 contrasts the North part of the country with Southwest. Almost half of the variables present high loadings in this factor and, they are characterized by being a comparison between formal and colloquial language such as Inteligente/Pilo, Dinero/Plata, Esposo/Marido, Después/Luego, Saco/Buzo, Frasco/Envase, among others. Factor 2 contrasts the cities located in the interior of the country with the cities located towards the borders. This factor is characterized by those variants that are common in a daily conversation, such as Estimado/Apreciado, Cerca/Próximo-Cerquita, Aqui/Acá, Ustedes/Vosotros. In fact, words like Casa/Hogar, Colegio/Escuela, Ave/Pajaro, Estudiante/Alumno are words that appear frequently in a conversation between members of the same household. Finally, Factor 3 contrasts the (North-)West part of the county with those cities outside this

region. In this case there are not as many variables that could help the interpretation of the factor and it is actually a somewhat expected result because of the amount of variance that contributes.

Variable	Uniqueness	Factor 1	Factor 2	Factor 3
Tú/Usted-Vos	0.023	−0.92	−0.355	
Ruana/Poncho	0.036	0.863	0.446	
Inteligente/Pilo	0.078	0.93		
Adelante/Delante-Enfrente	0.385	0.681		−0.317
Ave/Pájaro-Pájarito	0.171	0.551	0.718	
Quizás/Quizá	0.088	−0.739	−0.604	
Frasco/Envase	0.105	−0.917		
Estimado/Apreciado	0.021	−0.522	−0.722	−0.43
Lindo/Bello-Bonito	0.107	0.786	0.523	
Regalo/Obsequio	0.162	−0.442	−0.772	
Colegio/Escuela	0.191		0.806	0.391
Bajo/Debajo	0.164	0.535	0.736	
Dinero/Plata	0.219	0.883		
Cerca/Próximo-Cerquita	0.237		−0.822	
Mejor/Superior	0.218	0.826		0.31
Después/Luego	0.354	−0.793		
Bolso/Cartera	0.108	0.783	0.526	
Antes/Anteriormente	0.152	0.85	0.321	
Aún/Todavía	0.299	0.695	0.424	
Diadema/Balaca	0.475	−0.522		0.488
Impuesto/Gravamen	0.482		0.624	0.338
Ahí/Allí-Allá	0.247	−0.68	−0.52	
Esposo/Marido	0.399	0.745		
Ya/Ahora	0.324		0.648	0.421
Saco/Buzo	0.271	−0.828		
Aquí/Acá	0.202		−0.841	
Cariño/Afecto	0.273		−0.801	
Lejos/Apartado	0.635			−0.574
Estudiante/Alumno	0.174	−0.341	−0.818	
Hasta/Incluso	0.305	−0.822		
Ambición/Codicia	0.494		0.655	
Mientras/Durante	0.513	0.335		0.599
Ustedes/Vosotros	0.173		−0.889	
Casa/Hogar	0.31		0.82	

Table 3: Factor analysis uniqueness values and loadings



(a) Factor 1

(b) Factor 2

(c) Factor 3

Figure 4: Factor scores

4.3 Cluster analysis

As it is suggested by Grieve et al. (2011), since each factor represents a different common pattern of spatial clustering in the set of alternation variables, it is necessary to combine these results to determine a classification of each location and identify dialect regions. This is conducted by using a hierarchical classification method on the factor scores retrieved after the factor analysis. The difference between hierarchical and non-

hierarchical classification methods is that in the former each city is considered a cluster and in subsequent steps, cities that are closer (in terms of any distance measure) are merged to establish a larger cluster. The process stops when all cities are merged into one big cluster.

The first step to conduct this analysis is to decide what kind of distance is going to be used. Since the data used are the factor scores, which by definition are projections of the original coordinates mapped in a 34-dimensional space into a 3-dimensional space, Euclidean distance is chosen. Then, Ward's method is used to measure the similarity between clusters because it is based on an analysis of variance and leads to get clear and compact agglomerations (Grieve et al., 2011). The results can be seen in a dendrogram, where the history of each step done in the algorithm is showed, easing the decision of the number of final clusters and the identification of possible subclusters.

Figure 5 shows the dendrogram obtained using the factor scores from the three factor solution. Four clusters are clearly visible and they are mapped in Figure 6a. These four clusters can be labelled by their location as North, West, East and South and are clearly derived from the common patterns of regional variation found in the factor analysis. By analyzing the internal structure within each cluster, it is also possible to identify subregions that share linguistic properties. In the southern region there are two regions that stand out by their location. The first subregion is characterized by those cities located in "Los Llanos" (labelled as six in Figure 6b). And the second subregion is characterized by cities that are located near Valle del Cauca (Cali in the map) and near the border with Ecuador. In the sake of illustration, Figure 6b mapped all subregions across the country.

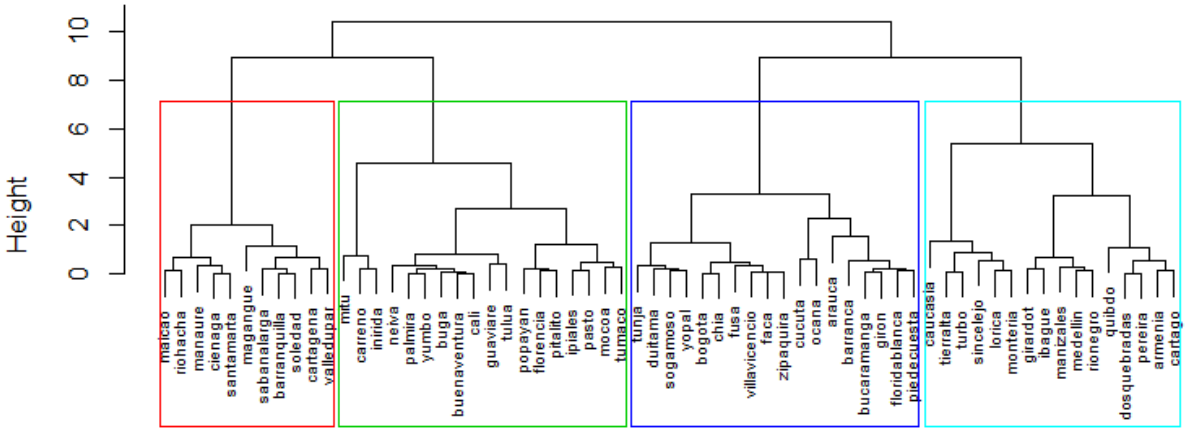
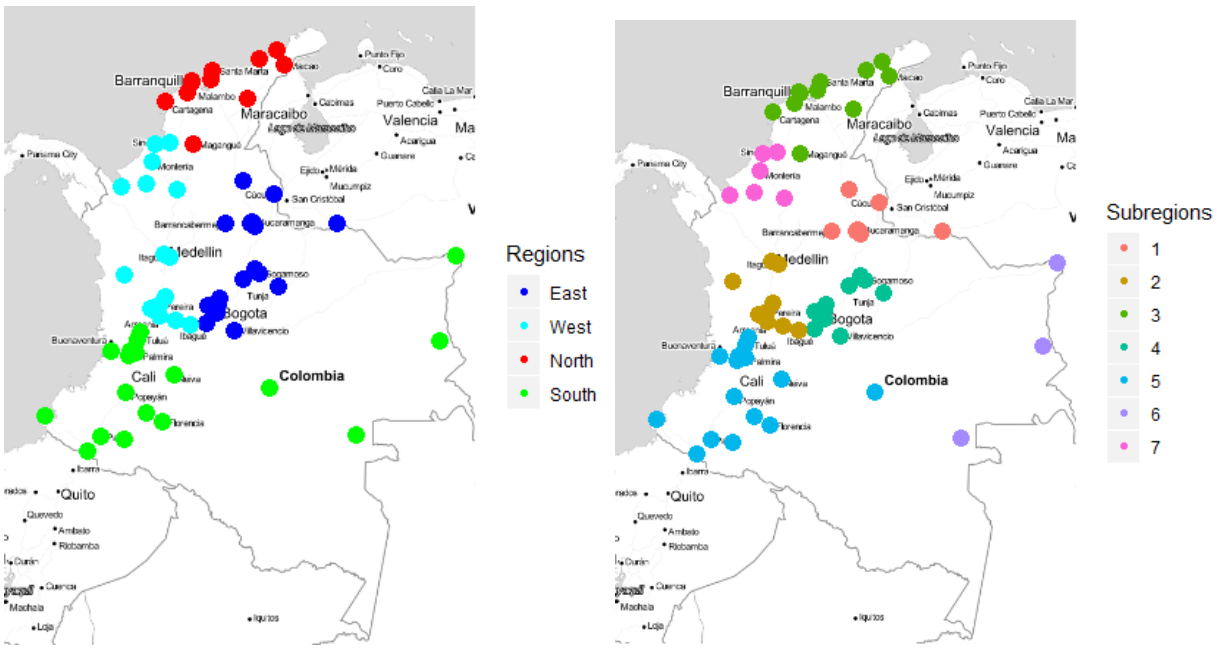


Figure 5: Dendrogram based on three factor solution



(a) Dialect regions (b) Dialect Subregions

Figure 6: Dialects

It is important to mention that this process was also performed with a two factor solution and a five factor solution. The motivation for a two factor solution was because it explained 69.1% of the variance, and the motivation of the five factor solution was because up to that point the eigenvalues, which is the same as the sum of squares of the loadings in each factor, were larger than 1. The dialectic regions conformed with the two factor solution are a little bit different and somewhat harder to explain, illustrating that the third factor is important. On the other hand, the regions conformed by the five factor solution were similar to those presented in this document, giving some robustness to the results found here.

5 Conclusion

This paper analyzed Twitter data from the most important cities, in terms of population, of Colombia. In the beginning a total of 72 cities were considered to be in the study, but in the retrieving process some changes had to be made because of the small number of tweets that some cities presented. The study successfully retrieved information on 63 cities of them.

To identify linguistic variation across the country, 34 continuously measured alternation variables were considered in the analysis from a corpus with approximately 20.5 million of words corresponding to 1.6 million of tweets. These alternation variables were chosen manually by checking the frequency of each word in the corpus and looking total synonyms for them.

This information was then analyzed using the statistical methodology for linguistic variation proposed by Grieve et al. (2011), resulting in four big regions and seven subregions that share particularities of the language with respect to formality and variability

in the day to day conversation.

Although these results cannot be generalized for different reasons, such as the under-representation of the population in some parts of the country and the socio-demographic profile of twitter users, it is important to adopt this kind of methodology in the analysis of linguistic variants in a country as diverse as Colombia. There are more benefits that disadvantages, and is always a good idea to be curious about how people live and what makes them to be how they are.

Now, from the perspective of the author of this document that does not have any experience in the field in Linguistics, this experience has been enriching in several ways, but two in specific: First, the retrieving process and the use of regular expressions (for me knowing how regular expressions work is a game changer in my further analyses as a statistician). Second, the way how statistical methodologies have gained acceptance in social sciences.

References

Goncalves, B. and Sánchez, D. (2014). Crowdsourcing dialect characterization through twitter. *PLoS ONE*, 9.

Grieve, J., Speelman, D., and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23:193–221.

Huang, Y., Guo, D., Kasakoff, A., and Grieve, J. (2016). Understanding u.s. regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Ueda, H. and Takagaki, T. (1993). Varilex, variación léxica del español del mundo. `https://lecture.ecc.u-tokyo.ac.jp/~cueda/varilex/`. Last accessed on 28-12-2018.

6 Appendix

In the folder there are to R-scripts:

- `DataRetrieval.R`: Is the script used to collect the data from Twitter. It is there in the sake of illustration, it does not have to be run.
- `Analysis.R`: Is where the whole analysis is conducted, including maps. It requires the file "Alternation_var.csv" (dataset) to run correctly.