

Modelo para la determinación de isoglosas y clasificación dialectal a partir de datos del tomo III del Atlas Lingüístico Etnográfico de Colombia

Autor
Javier Orlando Fernández Campos

Tutor
PhD. Luz Ángela Rocha Salamanca

Codirector
MSc. Johnatan Estiven Bonilla Huerfano

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS
MAESTRÍA EN CIENCIAS DE LA INFORMACIÓN Y LAS COMUNICACIONES
ÉNFASIS EN GEOMÁTICA
BOGOTÁ D.C, 2022

Tabla de Contenido

I Propuesta de Investigación **1**

1. Introducción **1**

2. Problema de Investigación **3**
2.1. Formulación del Problema 4
2.2. Preguntas de la investigación 4

3. Hipótesis **5**

4. Objetivos **6**
4.1. Objetivo General 6
4.2. Objetivos Específicos 6

5. Justificación **7**

6. Marco de Referencia **8**
6.1. Dialectología, geolingüística y dialectometría 8
6.1.1. Dialectología y geolingüística 8
6.1.2. Dialectometría 8
6.1.3. Medidas de similitud Lingüistica 11
6.2. Atlas Lingüistico Etnográfico de Colombia (ALEC) 13
6.2.1. Investigación 13
6.2.2. Encuestadores 13
6.2.3. Localidades 14
6.2.4. Informantes 14
6.2.5. Estructura de los mapas 14
6.2.6. ALEC digital 16
6.3. Propuestas de clasificación dialectal del español de Colombia con base en datos del ALEC 17
6.3.1. Luis Flórez, 1961 17
6.3.2. José Joaquín Montes, 1982 19
6.3.3. Mora et al., 2004 22
6.3.4. Avila et al., 2015 23
6.3.5. Johnatan Bonilla, 2019 24
6.3.6. Néstor Ruíz, 2020 24
6.4. Aplicaciones informáticas 26
6.4.1. Gabmap 26
6.4.2. GeoDa 27

6.4.3. SpaceStat 28
6.4.4. Python 28
6.4.5. R 29
6.4.6. ArcGIS 29

II Desarrollo de la investigación 31

7. Metodología 31
7.1. Exploración y recopilación de información 32
7.2. Análisis de datos 35
7.2.1. Localidades seleccionadas 36
7.2.2. Actividades económicas predominantes 37
7.2.3. Atura media sobre el nivel medio del mar 39
7.2.4. Vías de acceso 41
7.2.5. Temperatura media 42
7.2.6. Aptitud agroclimática 43
7.2.7. Índice de disponibilidad hídrica (IDH) 46
7.2.8. Riesgo por amenaza sísmica 48
7.2.9. Precipitación media multianual 50
7.2.10. Similitud lingüistica 52
7.2.11. Distancia geográfica 57
7.3. Diseño del Modelo 58
7.3.1. Pesos Espaciales 58
7.3.2. Correlación Espacial 63
7.4. Construcción del Modelo 68
7.5. Selección del Modelo 86
7.5.1. Coeficiente de determinación 86
7.5.2. Análisis de términos del error 88
7.5.3. Sumario modelo seleccionado: 103
7.6. Generación de limites lingüísticos 105
7.6.1. Superficie dialectal continua 105
7.6.2. Construcción de Isoglosas 111
7.7. Clasificación dialectal 113
7.7.1. Regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (REDCAP) . 113
7.7.2. Propuesta de clasificación dialectal 121

III Resultados **127**

8. Resultados **127**
8.1. Similitud lingüistica - IRI 127
8.2. Distancias Geográficas 128
8.3. Correlación espacial 129
8.4. Modelo estimado 131
8.4.1. Modelo Lineal (Estimado por MCO) 131
8.4.2. Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial 133
8.5. Limites Lingüísticos 136
8.5.1. Superficie dialectal continua 136
8.5.2. Isoglosas 136
8.6. Clasificación dialectal 137

9. Análisis de Resultados **139**
9.1. Distancia lingüistica media 139
9.2. Distancia geográfica 142
9.3. Correlación espacial 145
9.4. Modelo estimado 152
9.4.1. Modelo lineal no espacial 152
9.4.2. Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial (VMS), protocolo de pruebas y validación 154
9.5. Limites Lingüísticos 163
9.6. Clasificación dialectal 166

IV Conclusiones **177**

Lista de Figuras

1. Matriz de Similitud (Índice Relativo de Identidad) 11
2. Mapa (265) ALEC análogo, *Cometa* 16
3. SIG ALEC y ALEC Web 17
4. Áreas dialectales propuestas 18
5. División dialectal propuesta por Montes 20
6. División dialectal: superzonas propuesta por Montes 21
7. Superdialectos del español en Colombia propuestos por Ruíz . 25
8. Python para el análisis de información lingüística 28
9. Metodología de la investigación 31
10. Localidades seleccionadas 37
11. Actividades Económicas predominantes 38
12. Información básica IGAC, altura sobre el nivel medio del mar 39
13. Distribución espacial, altura sobre el nivel medio del mar . . . 40
14. Histograma de las vías de acceso de localidades 42
15. Temperatura promedio multianual - fuente IDEAM 42
16. Temperatura promedio multianual - Histograma 43
17. Aptitud agroclimática promedio multianual - fuente IDEAM . 44
18. Aptitud agroclimática promedio multianual - Histograma . . . 45
19. IDH promedio multianual 47
20. IDH promedio multianual por localidades - Histograma 48
21. Riesgo por Amenaza Sísmica - Fuente: Servicio Geológico . . . 49
22. Amenaza Sísmica Localidades- Histograma 50
23. Precipitación media multianual - Fuente IDEAM 51
24. Precipitación Localidades - Histograma 51
25. Similitud lingüística (Az)- Similitud lingüística media (Narj) . 54
26. Gráfico de probabilidad QQ similitud lingüistica media 55
27. Distancia lingüística por localidad 56
28. Distancia Geográfica por Localidad 58
29. Polígonos de thiessen de las localidades 59
30. Contigüidad de Torre (Iz) vs Reina (Der) 60
31. Histograma de conectividad - pesos espaciales 61
32. Valores de contigüidad visualización 62
33. Grafo de conectividad 62
34. Índice de Moran para la Distancia Lingüistica 64
35. Ilustración Modelo SAR(1) 70

36. Ilustración Modelo Retardo Espacial (Yrigoyen, 2003) 72
37. Ilustración generación del modelo espacial (Anselin y Rey, 2014) 77
38. Ilustración prueba normalidad modelo MCO 89
39. Residuos del modelo estimado por MCO 90
40. Valores estimados MCO 91
41. Valores distancia lingüistica media 91
42. Ilustración prueba normalidad modelo retardo espacial 92
43. Residuos del modelo retardo espacial 94
44. Estimados modelo retardo espacial 94
45. Ilustración prueba normalidad modelo retardo espacial VMS . 95
46. Residuos del modelo retardo espacial VMS 97
47. Valores estimados modelo retardo espacial VMS 97
48. Ilustración prueba normalidad modelo SARAR 98
49. Residuos del modelo SARAR 99
50. Valores estimados modelo SARAR 100
51. Ilustración prueba normalidad modelo SARAR VMS 101
52. Residuos del modelo SARAR VMS 102
53. Valores estimados modelo SARAR VMS 102
54. Modelo seleccionado: valores estimados 104
55. Variograma con valores estimados 106
56. Mapa variográfico y nube variográfica 107
57. Kriging ordinario valores estimados 109
58. Kriging ordinario valores estimados - Valores medidos 110
59. Kriging ordinario gráfico QQ para errores estandarizados . . . 111
60. Isoglosas estimadas 112
61. Clasificación método REDCAP (8 Grupos) 116
62. Clasificación método REDCAP (2 grupos) 118
63. Clasificación método REDCAP (5 grupos) 118
64. Clasificación método REDCAP (12 grupos) 119
65. Clasificación método REDCAP (16 grupos) 119
66. Clasificación método REDCAP (20 grupos) 120
67. Clasificación método REDCAP (24 grupos) 120
68. Clasificación dialectal propuesta 123
69. Clasificación dialectal propuesta - dialectos 125
70. Resultado Superficie dialectal continua 136
71. Resultado Isoglosas 137
72. Resultado clasificación dialectal propuesta 138
73. Localidades mayor distancia lingüística media 139

74. Localidades mayor distancia lingüística - Actividades Económicas . . . 140
75. Temperatura y precipitación mayor distancia lingüística 140
76. Localidades menor distancia lingüística media 141
77. Actividades Económicas e IDH menor distancia lingüística . . 142
78. Localidades mayor distancia lingüística media - distancia geográfica . . . 144
79. Localidades menor distancia lingüística - distancia geográfica . 145
80. Actividad económica minería y superficie dialectal continua . . 146
81. Zonas de población afrodescendiente (Barbary y Urrea, 2003) . 147
82. Aptitud agroclimática - Ligeramente húmedo 148
83. Aptitud agroclimática - Extremadamente seco 148
84. Precipitación - Rango 2000- 2500 149
85. IDH - Semiseco 150
86. Vías - Camino de herradura 151
87. Vías de acceso - modelo estimado por MCO 153
88. Aptitud agroclimática - moderadamente húmedo 156
89. Amenaza Sísmica - Moderadamente Alto 157
90. Temperatura - Entre 20 y 22 grados 158
91. Temperatura - Entre 26 y 28 grados 158
92. IDH - Semihúmedo 159
93. IDH - Semiseco 159
94. Vías de acceso - Camino de herradura 161
95. Vías de acceso - Carretera y férrea 161
96. Vías de acceso - Caminos carreteables 162
97. Vías de acceso - Carretera 162
98. Red Vial (INVIAS) - superficie dialectal continua 163
99. Movimientos de población siglo XIX (Mejía, 2013) 165
100. Superdialectos 169
101. Tres divisiones mediante agrupamiento REDCAP 173
102. Tres divisiones mediante agrupamiento difuso (Bonilla, 2019, pg. 19) 174

Lista de Tablas

1. Propuesta de clasificación dialectal según Mora et al., 2004 . . 22
2. Clasificación dialectal propuesta por Ruíz (2020) 26
3. Métodos de análisis lingüistico Gabmap 27
4. Metadatos de cada localidad ALEC 34
5. Datos a utilizar dentro del desarrollo 36
6. Actividades Económicas principales 38
7. Actividades Económicas - Variables reclasificadas 38
8. Valores referentes para la altura sobre el nivel medio del mar . 40
9. Vías de acceso a localidades 41
10. Vías de acceso a localidades - Variables ficticias o reclasificadas 41
11. Temperatura en variables ficticias o reclasificadas 43
12. Niveles agroclimáticos según el IDEAM 44
13. Diferencia en valores de aptitud agroclimática 45
14. Categorías IDEAM - IDH 47
15. IDH - Variables Ficticias o reclasificadas 48
16. Amenaza Sísmica - Categorías 49
17. Amenaza Sísmica - Variables reclasificadas ejemplo 50
18. Precipitación - Variables reclasificadas 52
19. Valores nulos - Determinación de mapas léxicos 53
20. Datos finales - Ejemplo de información subida y procesada en Gabmap 53
21. Matriz diferencia lingüistica - Relación por Localidad 54
22. Similitud lingüística media por Localidad 54
23. Resultados de pruebas de Normalidad 56
24. Matriz distancia geográfica 57
25. I de Moran para actividades económicas global y bivariante . . 65
26. I de Moran para aptitud agroclimática global y bivariante . . 65
27. I de Moran global y bivariante para la altura 66
28. I de Moran global y bivariante para la precipitación 66
29. I de Moran global y bivariante para la temperatura 67
30. I de Moran global y bivariante para el IDH 67
31. I de Moran global y bivariante para las vías de acceso 68
32. I de Moran global y bivariante para la distancia geográfica . . 68
33. Modelo de regresión no espacial 75
34. Valores de pruebas de dependencia espacial 80
35. Estimación del modelo de rezago espacial 81

36. Estimación del modelo de retardo espacial para variables significativas 83
37. Estimación del modelo SARAR 84
38. Estimación del modelo SARAR para variables significativas . . 85
39. Estimación del coeficiente de determinación 87
40. Autocorrelación residuos modelos 103
41. Sumario modelo seleccionado 104
42. Clasificación dialectal propuesta 123
43. Similitud Lingüística media por localidad 128
44. Estadísticos resumen de similitud lingüistica media 128
45. I de Moran para Aptitud Agroclimática univariante y bivariante 129
46. I de Moran univariante y bivariante - Precipitación 129
47. I de Moran para las Vías de Acceso - Resultados 130
48. Modelo lineal (estimado por MCO) variables mayor significancia 131
49. Resultados contraste de dependencia espacial 132
50. Modelo de retardo espacial para variables significativas 133
51. Localidades con mayor error en la diferencia Valor estimado - Valor Real 135
52. Localidades con mayor diferencia lingüística media 139
53. Localidades con mayor diferencia lingüística - distancia geográfica 143
54. Localidades con actividad económica - Minería 154
55. Localidades con actividad económica - Minería (año encuesta) 155
56. Localidades con vías de acceso - mayor influencia 160
57. SSD mayor y menor por particiones REDCAP 167

Capítulo I
Propuesta de Investigación

1. Introducción

El Instituto Caro y Cuervo (ICC) es una institución de educación superior de carácter nacional encargada del resguardo del patrimonio lingüístico del país (Instituto Caro y Cuervo, 2021), razón por la cual realizó una investigación acerca del uso del español de Colombia en los niveles dialectal y subdialectal (Florez, 1961) cuyo resultado fue el Atlas Lingüístico Etnográfico de Colombia (ALEC) en formato análogo. Como parte de este trabajo se publicaron seis tomos con 1523 mapas del territorio explorado, con el fin de determinar la variación de distintas palabras y expresiones (Florez, 1983). Dada su reciente divulgación mediante el uso de herramientas geoespaciales el ALEC [1] brinda la posibilidad de realizar investigaciones novedosas y diferentes aplicaciones tanto en las ciencias humanas como en otras disciplinas, principalmente en dialectología y geolingüística, áreas del saber que a lo largo de esta investigación permitirán resolver diferentes interrogantes de carácter científico.

Esta investigación plantea llevar a cabo el análisis cualitativo y cuantitativo de la información lingüística (léxica) contenida en el tomo III del ALEC a partir del uso de técnicas dialectométricas, de econometría espacial, geoestadisticas y minería de datos. Recientemente la dialectología no se ha enfocado únicamente en el análisis de variantes léxicas, fonéticos o gramaticales sino en cambio ha podido desarrollar diferentes técnicas para identificar y clasificar áreas geográficas con características lingüísticas similares, el análisis de

[1] Acceso herramientas digitales haciendo uso de los links:
SIG ALEC: http://atlasweb.caroycuervo.gov.co/
ALEC Digital: http://alec.caroycuervo.gov.co/alec/

tipo geoestadístico y econométrico es el más reciente, este enfoque ha posibilitado que la explicación a la variación lingüística pueda ser determinada por factores geográficos, lingüísticos y sociales subyacentes en cada región (Wieling y Nerbonne, 2015). Es entonces el objetivo principal de esta investigación diseñar un modelo para la estimación de Isoglosas y clasificación dialectal a partir de datos geolingüísticos (léxicos) del tomo III del ALEC, partiendo de una perspectiva cuantitativa basada en análisis geoestadistico y econométrico, teniendo como propósito demostrar que la variación dialectal es producto y/o está relacionada intrínsecamente con información geográfica.

2. Problema de Investigación

La dialectología tiene como propósito el estudio de la variación lingüística en el contexto espacial donde los resultados y su representación se dan generalmente con la ayuda de técnicas cartográficas (Dubert y Sousa, 2016). A lo largo del tiempo los métodos de análisis de información geolingüística se han centrado en la evaluación de rasgos en distintos niveles de la lengua considerados relevantes, ya sea de forma individual (aislada) o grupal. El análisis cuantitativo es uno de estos, fue desarrollado a principios de los años setenta y en los últimos años ha tenido una relevancia particular en geolingüística y dialectología donde el uso de herramientas tecnológicas e informáticas para el procesamiento y análisis de grandes volúmenes de datos ha sido fundamental, actualmente los expertos resaltan que la variación dialectal no se reduce a caracterizaciones simples sino que en cambio existen aspectos geográficos, temporales, económicos y sociales que inciden en la clasificación lingüística y/o división dialectal (Dubert y Sousa, 2016), tal es el caso de la investigación de Burridge (2017), quien mostró como las características geográficas en las fronteras (costeras) tienden a generar limites dialectales perpendiculares debido a la relativa probabilidad de contacto social, de esta manera determina predicciones sobre los límites dialectales a partir del estudio de las propiedades físicas del espacio donde se hablan.

El estudio de la relación existente entre el lenguaje y la geografía se remonta a los inicios de la dialectología como disciplina, esta conexión no solo es evidente por la distancia posicional de los diferentes hablantes, sino también por los accidentes geográficos (ríos, montañas, valles, lagos, etc.), las divisiones culturales, administrativas y los aspectos sociales (Dubert y Sousa, 2016). Por lo tanto la variación lingüística es un fenómeno complejo cuyas características se pueden estudiar en un contexto multidimensional con herramientas de análisis y descripción basadas en modelos matemáticos y estadísticos (Nerbonne, 2006). Es de resaltar que la dialectometría ha enfatizado en medir diferencias en las variaciones del lenguaje sin un compromiso con la información geográfica, esta es relevada a un dato más sobre la información sin jugar un papel importante sobre la medición misma (Wieling y Nerbonne, 2015).

En Colombia a la fecha se han realizado seis propuestas de clasificación y/o división dialectal a partir del uso de información contenida en el ALEC, estas propuestas se basan en la identificación de distintos rasgos y fenómenos lingüísticos junto con el trazado de isoglosas, líneas imaginarias que permiten crear subconjuntos de variables lingüísticas mostrando patrones semejantes de variación espacial (Dubert y Sousa, 2016); una de las clasificaciones dialectales más recientes es la de Bonilla (2019), en ella se hace uso de herramientas dialectométricas alternativas permitiendo el análisis estadístico y cuantitativo de información léxica, sin embargo no incorpora elementos geográficos importantes dentro de su geomodelado, elementos que al ser tomados en cuenta pueden mostrar patrones ocultos dentro de la información y darle una mayor precisión y valor a la clasificación dialectal propuesta (Wieling y Nerbonne, 2015).

A partir de lo anterior el objetivo principal de esta investigación comprende el diseño e implementación de un modelo espacial para la estimación de Isoglosas y clasificación dialectal a partir de datos geolingüísticos del tomo III del ALEC, este modelo busca incluir información léxica y geográfica, definiendo de forma cuantitativa áreas con características lingüísticas semejantes con objetividad y precisión, demostrando así que la variación dialectal está relacionada intrínsecamente con la información geográfica.

2.1. Formulación del Problema

¿Existe una relación cuantitativa (modelo espacial) entre variables geográficas y variables léxicas que explique la variación dialectal subyacente en los datos léxicos del tomo III del Atlas Lingüístico Etnográfico de Colombia?

2.2. Preguntas de la investigación

- ¿Cuál ha sido la manera como se ha clasificado dialectalmente el español de Colombia?
- ¿Qué variables geográficas se correlacionan con la distribución dialectal de los datos léxicos del tomo III del Atlas Lingüístico Etnográfico de

Colombia?

- ¿Que herramienta informática es la más óptima para el desarrollo e implementación del modelo espacial para la estimación de isoglosas y clasificación dialectal a partir de datos del tomo III del Atlas Lingüístico Etnográfico de Colombia?
- ¿Que protocolo de pruebas se debe seguir para hacer la evaluación del modelo propuesto?

3. Hipótesis

Existe relación entre variables léxicas y variables geográficas que permite la estimación de un modelo cuantitativo (espacial) para la delimitación y clasificación precisa y detallada de zonas dialectales a partir de datos (léxicos) del tomo III del ALEC.

4. Objetivos

4.1. Objetivo General

Diseñar un modelo espacial para la estimación de isoglosas y clasificación dialectal a partir de datos geolingüísticos del tomo III del Atlas Lingüístico Etnográfico de Colombia.

4.2. Objetivos Específicos

- Evaluar la información contenida en el tomo III del ALEC para determinar la importancia de los metadatos a utilizar y cada una de las variables del modelo en la estimación de isoglosas y clasificación dialectal a partir de datos del tomo III del Atlas Lingüístico Etnográfico de Colombia.
- Diseñar un modelo para la estimación cuantitativa de isoglosas y clasificación dialectal a partir de datos del tomo III del Atlas Lingüístico Etnográfico de Colombia.
- Seleccionar herramientas para el desarrollo e implementación del modelo espacial para la estimación de isoglosas y clasificación dialectal a partir de datos del tomo III del Atlas Lingüístico Etnográfico de Colombia.
- Diseñar e implementar un protocolo de pruebas para la evaluación del modelo propuesto.

5. Justificación

El desarrollo de un modelo para la clasificación dialectal y la determinación de isoglosas a partir de datos del tomo III del ALEC es de gran importancia para nuestro país, intenta resolver una discusión que se ha hecho a lo largo de los años en la búsqueda de determinar y definir de manera cuantitativa la interrelación de un dialecto con el espacio geográfico y las diversas hipótesis que existen alrededor de este tema, el modelo desarrollado adicionalmente intentará contrastar y validar las diferentes propuestas de clasificación dialectal (Florez, 1961; Montes, 1982; Mora et al., 2004; Avila et al., 2015; Bonilla, 2019; Ruíz, 2020) realizadas hasta la fecha.

Esta investigación es importante para el grupo de investigación NIDE dado que posibilita la opción de seguir incorporándose en un campo del conocimiento que no había sido abarcado en investigaciones anteriores (ciencias humanas), en esta oportunidad haciendo uso de herramientas dialectométricas, geolingüísticas, geoestadisticas y computacionales se busca desarrollar un modelo espacial que permita representar y delimitar áreas con un rasgo dialectal concreto, lo cual implica un avance significativo hacia el entendimiento de fenómenos de tipo social, cultural, económico, entre otros.

Para la Maestría en Ciencias de la Información y las Comunicaciones la importancia de esta investigación radica en lo valioso que resulta el tratamiento y análisis de los datos lingüísticos a nivel estadístico y espacial, un análisis preciso permitirá no solo integrar dicha información sino adicionalmente incorporar mecanismos geoespaciales y algoritmos de ciencias de la computación (minería de datos) para automatizar tareas y descubrir patrones ocultos, posibilitando la generación de nuevo conocimiento.

6. Marco de Referencia

6.1. Dialectología, geolingüística y dialectometría

6.1.1. Dialectología y geolingüística

La Dialectología es la disciplina encargada del estudio de los dialectos y su variación dentro en un contexto social y geográfico (Heeringa, 2004), la geolingüística es uno de los métodos que permite especializar dicha variación de la lengua dentro de cartografía (Córdoba, 2011), en el desarrollo de la presente investigación se ahondara en ellas haciendo uso del concepto de dialectometría, un método cuantitativo desarrollado por Séguy (1973) cuyo enfoque es la medición de distancias dialectales (Nerbonne, 2006) utilizando herramientas computacionales y estadísticas (Wieling y Nerbonne, 2015), este método permitirá la generación de límites dialectales, ya sea como lineas imaginarias (isoglosas) que delimitan y separan las ocurrencias de diferentes variantes producto de fenómenos lingüísticos, o de forma continua, bajo la hipótesis que plantea la variación dialectal como una distribución gradual en el espacio geográfico (Jeszenszky y Weibel, 2015).

6.1.2. Dialectometría

La dialectometría nació a principios de los años setenta, anterior a nacimiento había surgido el problema alrededor de la distinción precisa de áreas dialectales, problema que en la actualidad es atribuido a la forma continua o dispersa como se distribuye el lenguaje en el espacio y sus limites territoriales (con base en las características dialectales predominantes) (Wieling y Nerbonne, 2015). A Séguy (1973) se le atribuye el examinar dichas características de forma agregada, es decir, analizar grandes conjuntos de características lingüísticas, cambiando de manera radical lo que se hacía hasta entonces, un estudio de rasgos (fonéticos, gramaticales y/o léxicos, entre otros) individuales (Nerbonne y Wieling, 2017), estas agregaciones permitieron indicar de manera más confiable las áreas en que se encuentran patrones lingüísticos comunes, Seguy no solo caracterizó la lengua de una manera distinta sino que también descubrió cuantitativamente la relación hipotética descrita en la geografía lingüística (Wieling y Nerbonne, 2015), aquella que existe entre las diferencias lingüísticas agregadas y la distancia geográfica (Séguy, 1973).

Seguy alcanzó a publicar solo dos documentos con sus teorías antes de sufrir un accidente inesperado que le causaría la muerte (Nerbonne et al., 2019), Hans Goebl lingüista de la Universidad de Salzburgo (Goebl, 1982) adoptó sus ideas y trabajo en su desarrollo (Goebl, 1982), Goebl centró sus estudios principalmente en las lenguas romances, haciendo grandes avances en el desarrollo de técnicas de medición dialectal, entre las contribuciones más destacadas de la escuela de Salzburgo (orientada por Goebl) a la Dialectometria se encuentra el uso de polígonos de Thiessen, un medio para dividir mapas en áreas relacionadas con el lugar de recolección de la información (Wieling y Nerbonne, 2015), la introducción a una amplia variedad de estadísticas descriptivas (Dubert y Sousa, 2016), la ponderación de frecuencias inversas para el calculo de diferencias agregadas (Wieling y Nerbonne, 2015) y el abogo de procedimientos computacionales complejos. La versión dialectometrica de Salzburgo representa una herramienta heurística del análisis de datos exploratorio y de uso universal, hace uso de instrumentos de análisis que van desde la cartografía lingüística de similitud hasta mapas dialectometricos correlativos (Goebl, 2006).

Hacia finales del siglo XX y principios del siglo XXI se hicieron grandes aportes por parte de otros lingüistas investigadores, entre los más relevantes se encuentra la introducción de la distancia de Levenstein (distancia de edición) utilizada por Kessler (1995) y adoptada por John Nerbonne y Wilbert Heeringa quienes publicaron una serie de estudios dialectométricos que representaron finalmente lo que sería lo que se conoce actualmente como la Escuela Dialectométrica de Groningen, siguiendo la ruta iniciada por Goebl desarrollaron nuevas técnicas y procedimientos estadísticos tales como el uso de escalamiento multidimensional, el agrupamiento jerárquico aglomerado y el análisis por componentes principales (Dubert y Sousa, 2016), sin embargo esta escuela ha enfatizado en que la dialectometría mide las diferencias del lenguaje y no tiene un compromiso relevante con la interpretación de la información geográfica (Wieling y Nerbonne, 2015).

La dialectometría sin embargo en los últimos años ha enfatizado en integrar datos adicionales en el análisis lingüístico (Wieling, 2012), desde el uso de diseños de modelos de regresión en los que se incluye la geografía como una medida de distancia hasta el acercamiento metodológico con la socio-

lingüística en los que se incluyen factores sociales (Wieling, 2012), para ello se hace uso de modelos de regresión aditivos generalizados (GAM) que permiten evaluar la importancia de la información lingüística frente a enfoques individuales sociales y geográficos (Wieling y Nerbonne, 2015), la diferencia notable entre los modelos de regresión y los modelos GAM es que los primeros simplifican la distancia geográfica como un dato predictor más, mientras que los segundos (GAM) representan patrones geográficos complejos a través del uso de coordenadas teniendo presente que la relación entre la variable respuesta y los predictores puede no ser lineal (Nerbonne y Wieling, 2017), uno de los estudios que utilizan tanto GAM como modelos de regresión es el realizado por Wieling et al. (2018), su conjunto de datos consistió en distancias de pronunciación para 562 palabras (comparadas con el holandés estándar) en 424 ubicaciones de los Países Bajos, los resultados muestran la existencia de un patrón geográfico complejo respecto a la distancia de pronunciación entre el holandés estándar y el holandés en las zonas periféricas, identificando simultáneamente la importancia de varias características sociales y léxicas, una de ellas, el uso más cercano al holandés estándar por las comunidades de mayor tamaño demográfico (Nerbonne y Wieling, 2017).

Es de destacar finalmente a Grieve et al. (2011), quienes han sido los principales exponentes del uso de una perspectiva geoestadística sobre la distribución espacial del dialecto. La geoestadística es un campo de la geografía matemática en la que se busca encontrar y tener en cuenta dependencias entre observaciones disponibles (de diferentes niveles y tipos) y su asociación con el espacio de forma cuantitativa (Emery, 2013), prestándole una gran atención a áreas como la geografía, la geología, la ecología, entre otros campos, Grieve (2018) recientemente ha estudiado y utilizado varios tipos de muestreo geográfico así como interpolaciones simples, observando limites dialectales a partir de las propiedades físicas del espacio donde se hablan (Nerbonne et al., 2019).

La dialectometría finalmente será el instrumento que servirá para el análisis de la información (léxica) contenida en el ALEC, proporcionará la metodología inicial para el establecimiento de distancias, diferencias y similitudes léxicas, el cálculo de las medidas de similitud (índice relativo de identidad, índice ponderado de identidad y distancia media euclidiana) (Goebl, 2006) y

los fundamentos para encontrar la relación con datos geográficos agregados.

6.1.3. Medidas de similitud Lingüistica

Según Goebl (1987) existen dos posibilidades de medir la proximidad o semejanza entre dos vectores de una matriz de datos.

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

Su elección depende de las intenciones de clasificación del investigador y su experticia. Dado que existen múltiples enfoques taxonómicos, en el presente documento se trataran los aspectos más importantes de dichas mediciones.

El Indice relativo de Identidad (IRI_{jk}) es una medida de similitud no ponderada entre vectores j , k, es conocida en alemán como "Relativer Identitätswert (RlW_{jk})", nombre adoptado por Goebl (1987) en investigaciones dialectométricas realizadas a principios de los años setenta, su esquema de medición se puede observar en la figura 1:

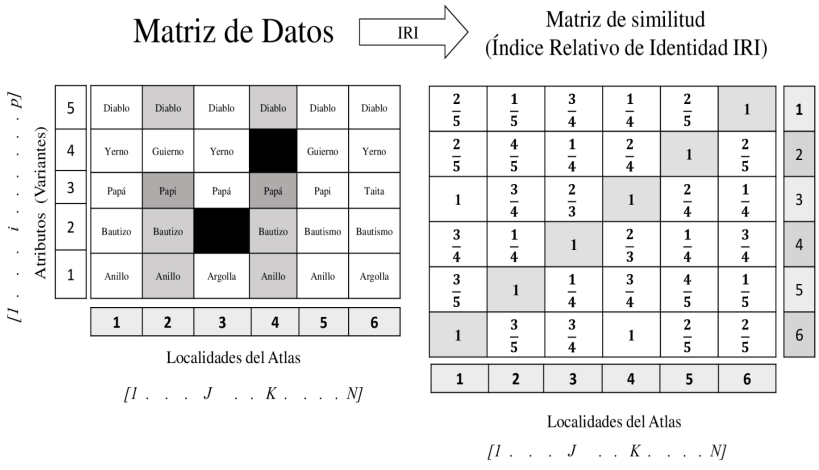


Figura 1: Matriz de Similitud (Índice Relativo de Identidad)
Cálculos a partir de datos del tomo III del ALEC

La generación de la matriz de similitud con el índice relativo de identidad IRI_{jk} se puede observar en la Figura 1, donde:

- En la parte izquierda se encuentra la matriz de datos, en color gris claro se resaltan las co-identidades (COI), es decir, para el atributo $i = 1$ tenemos $(Anillo, Anillo)$ para el atributo $i = 2$ se tiene $(Bautizo, Bautizo)$ y para el atributo $i = 5$ se tiene $(Diablo, Diablo)$.
- Dentro de la misma matriz de datos en color gris oscuro se resaltan las co-diferencias (COD), en el atributo 3 en este caso se tienen los valores $(Papi, Papa)$.
- Finalmente en la parte derecha se encuentra la matriz de similitud, una matriz cuadrada que depende del número de localidades del Atlas, para el ejemplo se tiene que el $IRI_{2,4} = \frac{3}{4} = 0,75$

Según Goebl (1987) el IRI_{jk} se basa en el concepto taxométrico del uso de co-identidades (COI) y co-diferencias (COD_{jk}) entre un par de vectores (j, k) de referencia, la formula es la siguiente (Ecuación 1):

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

En la formula:

- IRI_{jk} es el índice de identidad relativa.
- $\tilde{p}$ es el número de atributos en el vector j como en el vector k.
- (COI_{jk}) es la co-identidad entre los puntos j y k en el atributo i.
- (COD_{jk}) es una co-diferencia entre los puntos j y k en el atributo i.
- j es el índice del vector de referencia.
- k es el índice del vector en comparación.
- i es el índice del atributo.

Finalmente Goebl (1987) manifiesta que la ventaja del IRI como medida isocrática consiste en otorgarle a los valores de co-identidades (COI) el mismo que a las co-diferencias (COD), discrepancia notable con el IPI (Índice Ponderado de Identidad), una medida anisocrática ponderada de similitud lingüistica que permite elegir un factor (w) de ponderación para cada medición.

6.2. Atlas Lingüistico Etnográfico de Colombia (ALEC)

6.2.1. Investigación

Luis Flórez, investigador del Instituto Caro y Cuervo (ICC), cerca del año 1954 realizó el anteproyecto del ALEC, una investigación que pretendía realizar estudios de dialectología aplicados al español de Colombia a partir de distintos estudios lingüísticos realizados en diferentes lugares de Europa (Andalucía, Aragón, Portugal, Francia, etc.), la encuesta inició con un cuestionario que constaba de 8065 preguntas que hacia el año 1961 fue reducido a un total de 1500, relacionadas con distintos temas (campos semánticos) y recogiendo elementos lingüísticos, folclóricos, etnográficos tradicionales de nuestro país (Florez, 1983).

6.2.2. Encuestadores

Puesto que la extensión del territorio Colombiano es amplia el ALEC fue planeado para que distintos exploradores hicieran las encuestas, cada uno de ellos estaba vinculado al departamento de Dialectología del ICC, quien le daba instrucción fonética y dialectológica para realizar su labor, además de ello en la ciudad de Bogotá realizaron ejercicios de transcripción con el propósito de que existiera uniformidad en el registro de las respuestas y acompañamiento en campo de los más experimentados para detallar la forma en que se estaba realizando dicho trabajo (Florez, 1961).

Finalmente el número de encuestadores se extendió a 23 esto a causa de la poca frecuencia de las salidas en el tiempo, las circunstancias económicas y el número de personas capacitadas para ello (Florez, 1983).

6.2.3. Localidades

Las localidades seleccionadas, lugares geográficos (municipios o veredas) donde se llevó a cabo la encuesta, fueron en total 262, del ellas se entresacaron 24, esto debido a su cercanía en el momento de ser cartografiadas, no obstante aunque la información de esos 24 puntos no esta representada geográficamente, fueron tenidas en cuenta en las secciones de *Otras Respuestas* y de *Adiciones*, además de este criterio geométrico de equidistancia entre los diversos lugares de encuesta se tuvieron en cuenta otros aspectos para tal selección, 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 (Florez, 1983).

6.2.4. Informantes

Los informantes seleccionados, personas que hacían parte de una localidad y respondieron una mo más encuestas, fueron en total 2234 (Florez, 1983), estos para participar en la encuesta debían cumplir con los siguientes criterios:

- Ser nativos de la localidad o haber vivido en ella la mayor parte de su vida, al igual que sus padres y cónyuge.
- Ser analfabetos o tener poca instrucción escolar.
- Tener preferiblemente una edad comprendida entre los cuarenta y sesenta años.
- Ser en lo posible campesino o de origen campesino (Florez, 1983).

Finalmente de los seleccionados más de la mitad tenían entre treinta y sesenta años, por cada uno de los hombres se interrogó a una mujer y aproximadamente una quinta parte de los hombres era analfabeta (Florez, 1983).

6.2.5. Estructura de los mapas

La estructura de los mapas del ALEC tiene como principal componente la cuadricula con la división político administrativa del país (Figura 2), en la parte superior en el costado izquierdo se encuentra el número del mapa, en

cada lamina (página) pueden existir texto o imágenes, por lo tanto el número de mapa con lamina no coincide. Sin embargo la numeración es continua desde el primer tomo y el primer mapa hasta el último tomo y el último mapa; en la parte superior derecha está ubicado el número del mapa, abajo de él se encuentra el nombre del mapa y en forma descendente los símbolos de las variantes que se encuentran cartografiadas, variantes cuya alta frecuencia de pronunciación las hace relevantes para su representación cartográfica.

Al lado derecho del recuadro están *Otras respuestas*, aquellas localidades entresacadas debido a su alta cercanía en el momento de ser cartografiadas y el poco espacia disponible para los símbolos indicadores de los distintos fenómenos (Florez, 1983), los mapas en algunos casos tienen información adicional, estos datos complementarios se encuentran en la parte inferior del mapa junto con la localidad de la que hacen parte.

Dentro del mapa cartográfico base las variantes aparecen identificadas por símbolos convencionales que indican información de tipo lingüístico, si se encuentra una coma entre dos variantes esto significa que estas dos formas son equivalentes, es decir, existe polimorfismo, sinonimia, si en cambio existe un punto y coma, este significa que se están separando objetos o palabras de naturaleza diferentes, pero próximos (Florez, 1983), de esta manera se asignó una connotación para cada signo convencional.

Como resultado se publicaron seis tomos con 1523 mapas divididos en diferentes campos semánticos, de la siguiente manera:

- Tomo I: Tiempo y espacio, Campo. Cultivos. Otros vegetales, Algunas industrias relacionadas con la agricultura.
- Tomo II: Ganadería, Animales domésticos, Animales silvestres.
- Tomo III: La familia, ciclo de vida, Instituciones, vida religiosa, Festividades y distracciones.
- Tomo IV: El vestido, La vivienda.
- Tomo V: La alimentación, El cuerpo humano.

- Tomo VI: Oficios y empleos, Embarcaciones y pesca, Transportes, Fonética, Gramática.

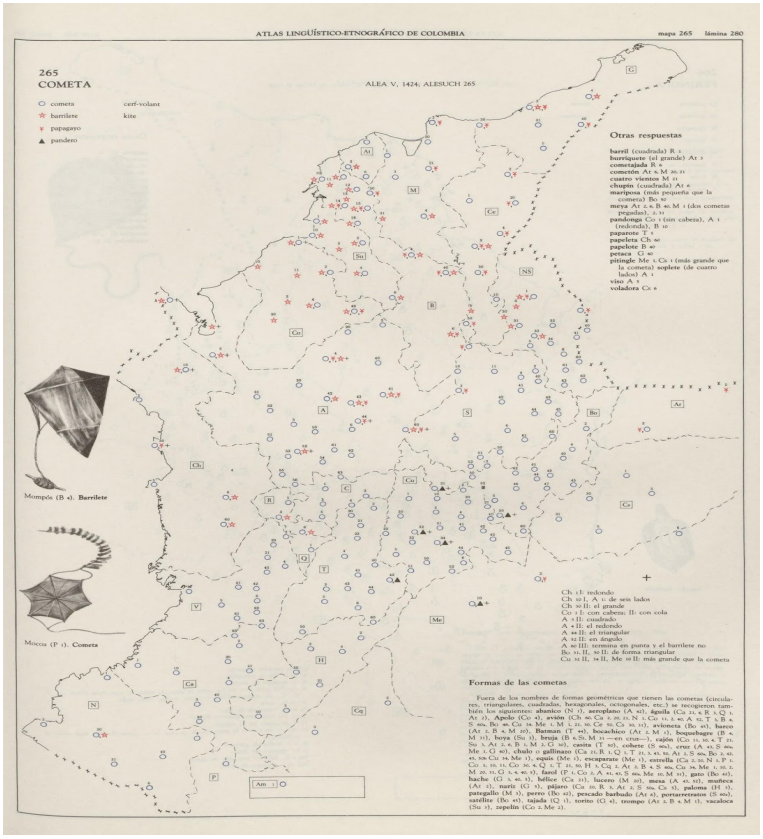


Figura 2: Mapa (265) ALEC análogo, *Cometa*

6.2.6. ALEC digital

Dado que el ALEC es una obra importante para los estudios lingüísticos del país, la línea de investigación en lingüística de corpus y computacional inició en el 2013 el macroproyecto: *Un ALEC para el siglo XXI*, esto con el propósito de crear e implementar un Sistema de Información Geográfica en línea así como un ALEC Web que permitiera conservar, sistematizar y divulgar el Atlas Lingüístico-Etnográfico de Colombia (ALEC) a partir de un desarrollo tecnológicos robusto (Instituto Caro y Cuervo, 2020).

En este momento el sistema de información geográfica y el ALEC Digital pueden ser consultados en línea, su estado actual integra la información de los seis tomos del ALEC, es decir, más de 1500 mapas con variaciones léxicas, fonéticas, entre otras, así mismo imágenes fotográficas e ilustraciones, un manual, un suplemento, archivos de audios con muestras de habla espontanea, entre otros, constituyendo así una herramienta indispensable para el análisis y estudio del español de Colombia (Instituto Caro y Cuervo, 2020).



(a) ALEC WEB (b) SIG ALEC

Figura 3: SIG ALEC y ALEC Web
(Rodriguez y Duarte, 2017) Iz y (Fernandez, 2017) Der

De esta manera, estos desarrollos informáticos (Figura 3) permiten estudios de investigación, así como análisis profundos y visualización en temas relacionados con dialectometría, tal es el caso de este desarrollo, donde la información léxica del ALEC así como algunos metadatos serán tomados de la base de datos espacial que hace parte del Sistema de Información Geográfica y del ALEC WEB, de tal manera que serán incluidos en el modelo espacial a generar así como en la determinación de isoglosas y clasificación dialectal propuesta.

6.3. Propuestas de clasificación dialectal del español de Colombia con base en datos del ALEC

6.3.1. Luis Flórez, 1961

Luis Flórez en el año 1961 realizó la primera propuesta de clasificación dialectal del español de Colombia (Florez, 1961) a partir del análisis de los

materiales recogidos de cincuenta y tres localidades, correspondientes a cinco departamentos del país, Antioquia, Bolívar, Cundinamarca, Nariño y Santander, Florez (1961) consideró diferentes aspectos lingüísticos para su elaboración, sin embargo menciona reiteradamente que esta es una propuesta de clasificación dialectal hipotética, la probabilidad de que fuera así estaba a merced del análisis de los datos de las regiones aun no visitadas y los existentes.

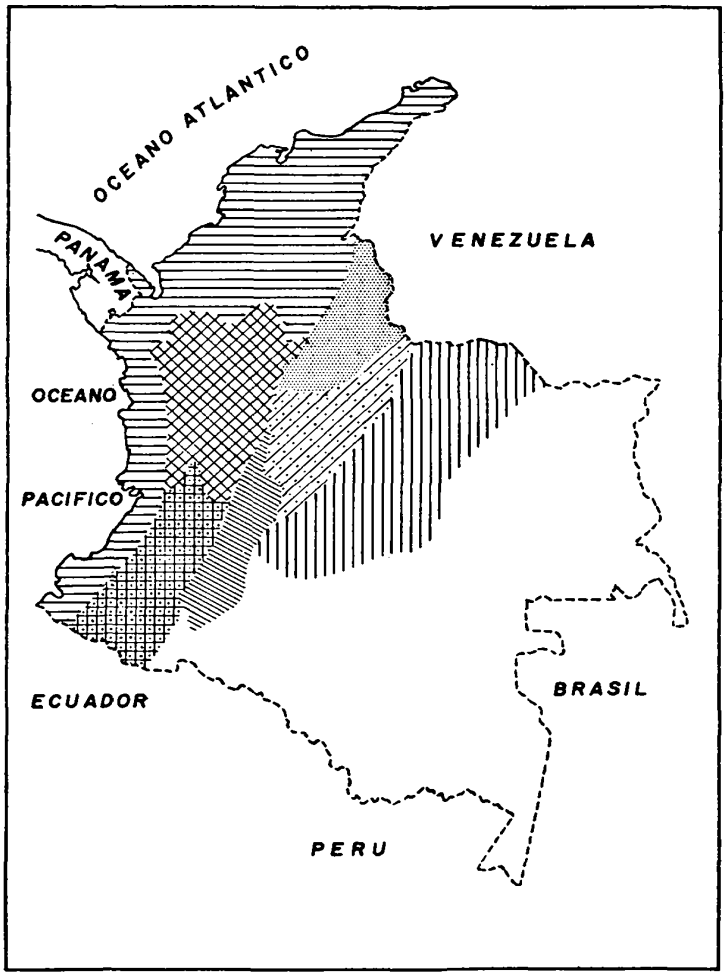


Figura 4: Áreas dialectales propuestas
(Florez, 1961)

La clasificación hipotética de Flórez tiene en cuenta seis zonas dialectales

(Figura 4), la primera es denominada región andina, en ella se incluye a Cundinamarca y Boyacá, la segunda y tercer zona hacen parte de la región del sur correspondiente a Nariño, entre las selvas del río Putumayo, la República del Ecuador y la Costa Pacífica, la cuarta zona se encuentra el occidente, compuesta por Antioquia, Caldas y partes de departamentos colindantes, la quinta zona está ubicada en el norte en la costa Atlántica. de ella hacen parte los departamento de Bolívar, Magdalena, la Guajira y Córdoba, finalmente la sexta zona es referida a la región llanera, de ella hacen parte los departamentos de Arauca, Meta y Casanare, cabe aclarar que la zona en mención no presenta información, esto debido a que para entonces la investigación no había llegado a estos lugares.

6.3.2. José Joaquín Montes, 1982

José Joaquín Montes realizó la segunda propuesta de clasificación dialectal cerca del año 1982 (Montes, 1982), su propuesta constituye el primer estudio sistemático del idioma español de Colombia enfocada al planteamiento de una división dialectal, en ella se incluyen diferentes fenómenos lingüísticos (léxicos, fonéticos, entre otros), algunos términos cruciales para entender su investigación son los siguientes:

- Norma: patrón de realización de las oposiciones del sistema lingüístico, es el criterio objetivo de delimitación de dialectos.
- Superdialecto: zonas geográficas extensas o superzonas que comparten algunos rasgos o normas lingüísticas.
- Dialecto: variante delimitada por una norma o conjunto de normas.
- Subdialecto: división del dialecto.
- Habla regional: pequeñas áreas territoriales con subconjuntos de normas idiomaticas.(Montes, 1982)

Montes divide el español de Colombia en dos superzonas o superdialectos (Figura 5): costeño y central (andino), el fenómeno lingüístico más importante en el que se justifica es el uso de la -s, puede ser pronunciada como aspiración (-h) o no serlo en la superzona "meridional", este fenómeno interfiere

en el sistema fonológico y morfológico del habla, además de este factor Montes añade el cronológico, el comercial y por último el factor administrativo-cortesano, sin embargo su descripción es exclusivamente cualitativa.

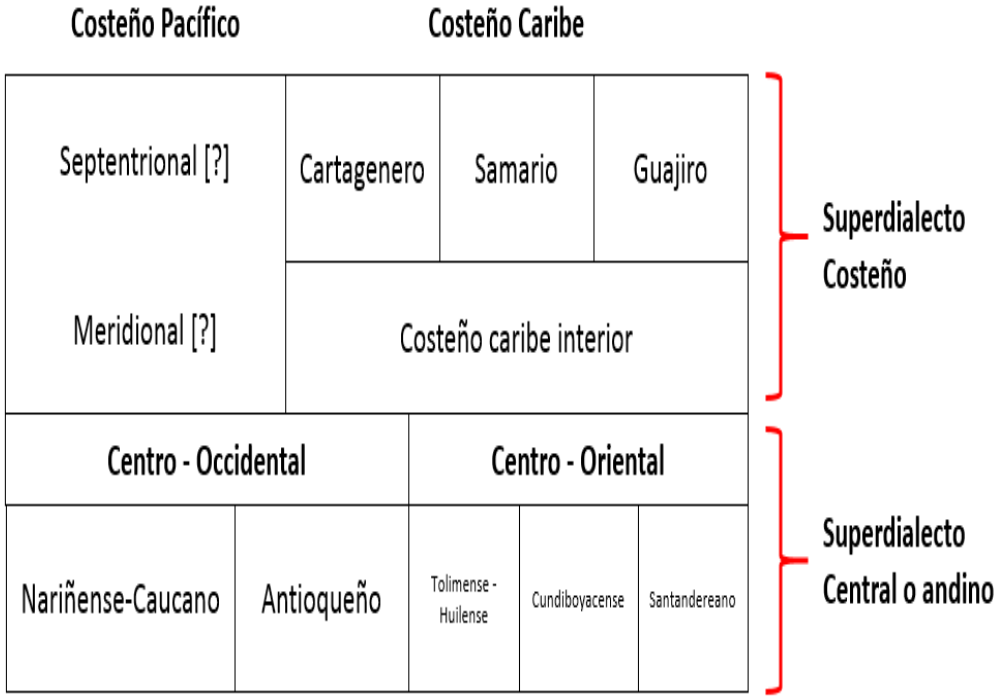


Figura 5: División dialectal propuesta por Montes (Montes, 1982)

Como se observa en la figura 5 los superdialectos son divididos en diferentes dialectos, el costeño se divide en dialecto Caribe y dialecto Pacífico y el superdialecto Central en occidental y oriental, zonas que tienen relaciones geográficas evidentes, cabe destacar que la propuesta de Montes incluye los datos adicionales que no tenia Flórez, aquellos referente a la visita de los departamentos de Arauca, Casanare, Meta y Caquetá, algo a resaltar en la propuesta de Montes (1982), es que no incluye zonas de transición dialectal, por lo que describe la superficie como un objeto discreta con limites dialectales concretos.

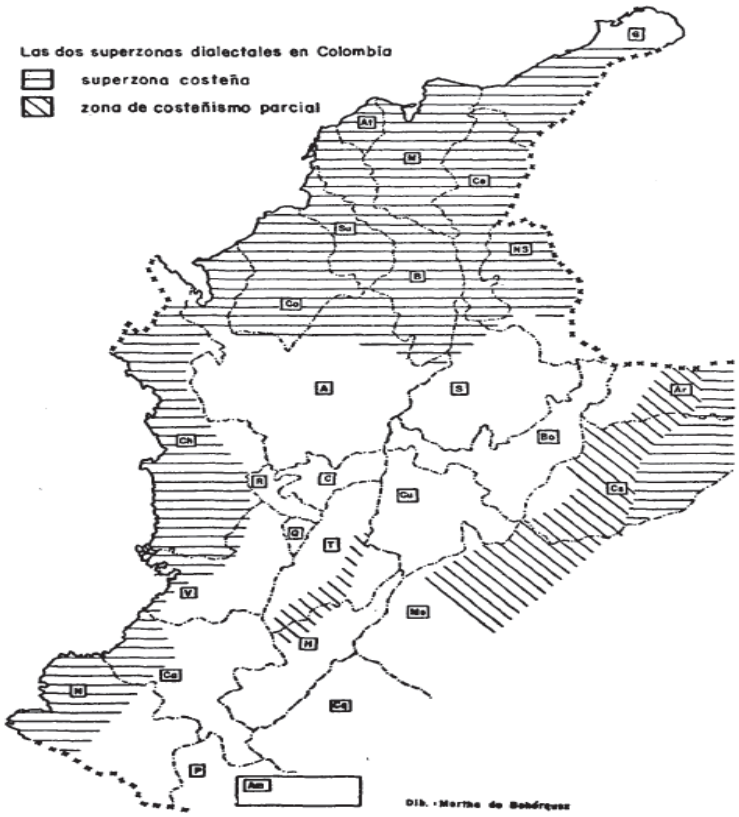


Figura 6: División dialectal: superzonas propuesta por Montes (Montes, 1982)

En conclusión la propuesta de Montes (1982) se caracteriza por el uso de normas precisas que permiten explicar fenómenos lingüísticos con un alcance diatópico según la geografía lingüística. Con el mayor uso de normas la determinación de zonas y subzonas dialectales pasa a ser más exacta, determinando así una clasificación dialectal de mayor rigurosidad lingüística, Montes sin embargo en el momento de su investigación no contaba con herramientas que le permitiera una fidelidad de tal magnitud en su propuesta, por lo que aunque su desarrollo sigue los pasos dados por Florez (1961) no hay análisis cuantitativo y/o dialectométrico que pueda repetirse y/o verificarse, de esta manera posteriores propuestas como la de Mora et al. (2004), Avila et al. (2015) y Bonilla (2019), encaminaron sus estudios de clasificación

dialectal a resultados provenientes de distintos métodos (cualitativos y cuantitativos) junto con el uso de herramientas informáticas de alta rigurosidad cuantitativa.

6.3.3. Mora et al., 2004

Mora et al. (2004) realizó una nueva propuesta de clasificación dialectal en el año 2004, mediante la identificación de isoglosas léxicas en los 1523 mapas del ALEC determinó superdialectos, dialectos y subdialectos tal como lo había hecho Montes (1982). La metodología de Mora et al. (2004) se basó en la agrupación de isoglosas y la determinación de frecuencia de aparición de variantes léxicas, lo cual constituyó el primer acercamiento estadístico al análisis de los datos del ALEC.

Finalmente la clasificación de Mora et al. (2004) se resume en la siguiente tabla:

Superdialecto	**Dialectos**	**Subdialectos**	**Léxico Regional**
Costeño	Costeño Atlántico	Cartagenero	Bolívar
		Samario	Cesar
		Guajiro	Guajiro
		Atlántico Interior	Córdoba
			Sucre
	Costeño Pacífico	Pacífico Norte	
		Pacífico Sur	
Andino	Andino Occidental	Antioqueño - Caldense	Antioqueño
			Caldense
		Caucano - Valluno	Valluno
			Caucano
		Andino - Sureño	Nariñense
	Andino Oriental	Santandereano	Nortesantandereano
			Santandereano
		Cundiboyacense	Cundinamarqués
			Boyacense
		Tolimense - Huilense	Tolimense
			Huilense
	Llanero	Llanero Norte	Araucano
			Casanareño
		Llanero Sur	

Cuadro 1: Propuesta de clasificación dialectal según Mora et al., 2004

Mora et al. (2004) al igual que Montes (1982) divide los superdialectos en Costeño y Andino, los dialectos predominantes en el superdialecto Costeño

siguen siendo el Costeño Atlántico y el Costeño Pacifico en cuanto al superdialecto Andino incluye los dialectos Andino Occidental y Andino Oriental al igual que Montes, sin embargo adiciona el dialecto Llanero lo que hace que la propuesta sea un poco más amplia que la de Montes, adicional a ello desglosa de forma detallada en subdialectos y léxicos regionales cada composición dialectal lo que supone un avance en este tipo de clasificación y análisis dialectal.

La propuesta de Mora et al. (2004) es el primer intento de análisis cuantitativo de datos lingüísticos enfocado a la clasificación dialectal del español de Colombia con datos del ALEC, según Lancheros Redondo (2018) además de incluir datos léxicos incluye datos etnográficos lo que puede influir de manera negativa en su clasificación o sor objeto de discusión dado que estos elementos no son aptos para el trazado de isoglosas.

6.3.4. Avila et al., 2015

Avila et al. (2015) realizaron una nueva propuesta de clasificación dialectal del español de Colombia con base en datos del ALEC en el año 2015, su propuesta incluye un análisis dialectométrico de cien mapas con información léxica del ALEC, su metodología se basó en el análisis de diferencias lingüísticas y valores de similitud lo que le permitió de alguna forma validar lo ya propuesto por Mora et al. (2004), Montes (1982) y Florez (1983), sin embargo no comparte la totalidad de los resultados con los resultados de los mencionados.

La propuesta de Avila et al. (2015) determina la división del superdialecto Costeño y Andino, la diferencia predominante con la clasificación de los otros autores radica en la inclusión de la zona pacifica dentro de la zona andina, su justificación se basa en la información referente a los metadatos de los informantes encuestados (lugar de proveniencia, edad, oficio. etc), Avila et al. (2015) considera a esta zona como un área de transición dialectal, en cuanto a los subdialectos Avila et al. (2015) los considera equivalentes a aquellos mencionados por Mora et al. (2004), finalmente la propuesta provee un mayor nivel de detalle en los resultados y un análisis cuantitativo (estadístico) preciso lo cual garantiza un alto nivel de confianza en la exactitud de la delimitación dialectal.

6.3.5. Johnatan Bonilla, 2019

Una de las propuestas más recientes es la de Bonilla (2019), su investigación fue realizada sobre doscientos mapas léxicos del ALEC escogidos de forma aleatoria y procesados en el software dialectométrico Gabmap, la similitud lingüística entre las variables léxicas la halló utilizando el índice relativo de identidad y finalmente plasmó los resultados cartográficamente empleando métodos de agrupamiento difuso.

Entre los resultados más relevantes se encuentra la división del español de Colombia en tres superdialectos, andino centro-oriental, costeño y andino centro-occidental, a su vez define ocho dialectos teniendo en cuenta las zonas de transición que se distribuyen desde la parte norte de Colombia hacia el sur y orientadas de forma vertical, finalmente presenta veintidós subdialectos con sus respectivas localidades lo que hace que en contraste con lo propuesto de Avila et al. (2015) y las demás propuestas sea una clasificación más precisa, actual y detallada del español de Colombia con datos del ALEC.

6.3.6. Néstor Ruíz, 2020

El estudio más reciente es el de Ruíz (2020), publicado en el 2020, su investigación ofrece una nueva división dialectal del español de Colombia basada en la integración de datos lingüísticos y dialectológicos, lo más llamativo de su propuesta es la elaboración e incorporación de un modelo espacial basado en el concepto de escala que surge como referencia a la necesidad de especificación de un superdialecto, dialecto y subdialecto en dos sentidos, en términos de su extensión geográfica y en relación a su configuración como una unidad dialectal.

El modelo según Ruíz (2020) se creó a partir de una distribución normal, en el intervalo $(0, 100)$ donde 100 expresado en porcentaje equivale al total de la superficie del territorio colombiano, a partir de ello hace una ponderación del peso y dimensión de cada unidad dialectal, dividiendo el intervalo en cuartiles, considerando cada cuartil como una escala dialectal y reconociendo que cada rasgo lingüístico detallado es subordinado del nivel superior o general, de tal manera generó el modelo en el que las hablas locales ocupan entre el 1 y 3.12 % del total del territorio, los subdialectos entre el 3 y

12.5 %, los dialectos entre el 12.5 y 50 % y los superdialectos entre el 50 y 100 %, la ventaja más llamativa sobre las otras propuestas es que el criterio de distinción entre rasgos de cada unidad dialectal principal pasa a ser espacial sin dejar de lado las características lingüísticos presentes en dichas áreas.

La clasificación propuesta por Ruíz (2020) presenta tres superdialectos, el español antillano, como una unidad dialectal que se propaga por las costas caribeñas entre los paises de Panamá, Colombia y Venezuela, el español andino que según Ruíz (2020) es el producto del contacto lingüístico Quechua-Español ubicado en el sur del territorio y finalmente el español neogranadino, en una ubicación intermedia entre ambos, el cual articula regiones distintas y extensas con desarrollos históricos diferentes.

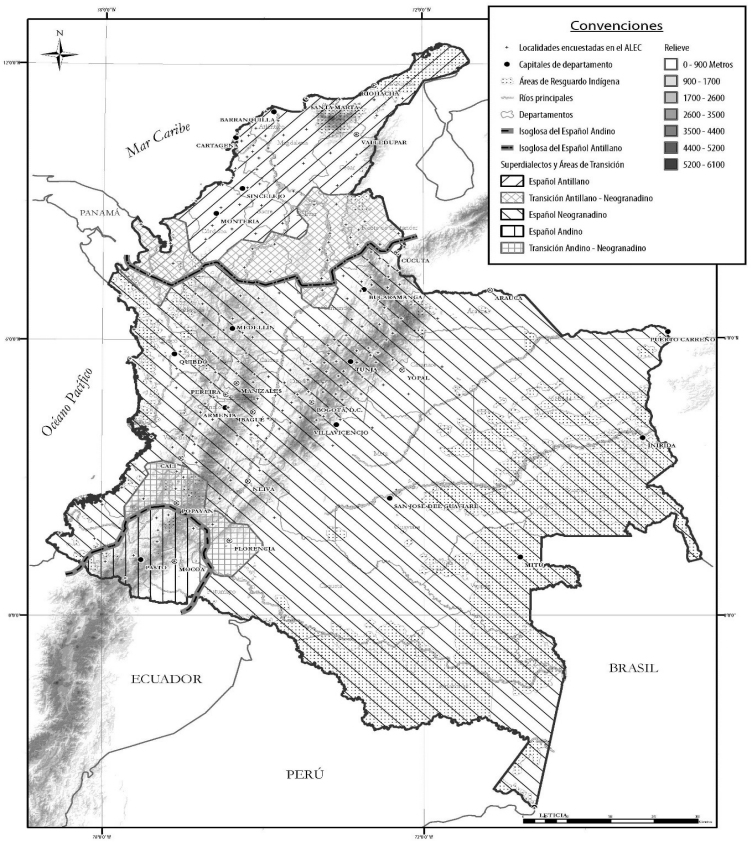


Figura 7: Superdialectos del español en Colombia propuestos por Ruíz (Ruíz, 2020)

La clasificación de Ruíz (2020) en forma detallada es de la siguiente manera:

Superdialectos	Dialectos	Subdialectos	Hablas regionales
Español Antillano	Caribeño colombiano	Cartagenero	Nte. de Bolívar / Sucre
		Samario	Atlántico / Magdalena
		Fonsequero	Guajira / Cesar / Nte. Norte de Santander
		Sabanero	Córdoba / Urabá antioqueño / Sur de Bolívar
Español Neogranadino	Occidental	Antioqueño-caldense	Antioquia / Caldas / Quindío / Risaralda / Nte. del Valle del Cauca
		Caucano-valluno	Andes del Cauca / Valle geográfico del Cauca
		Pacífico Septentrional	Chocó
		Pacífico Meridional	Costas de Valle del Cauca, Cauca y Nariño
	Oriental	Santandereano	Santander / Sur de Norte de Santander
		Cundiboyacense	Boyacá / Cundinamarca
		Tolimense-huilense	Tolima / Huila / Occ. del Caquetá
		Llanero Norte	Arauca / Casanare
		Llanero Sur	Meta
		Llanero Este	Vichada
	Amazónico	Amazónico Norte	Guaviare / Guainía / Vaupés
		Amazónico Sur	Ori. de Caquetá / Amazonas
Español Andino	Andino colombiano	De tierras altas	Andes de Nariño
		De tierras bajas	Putumayo

Cuadro 2: Clasificación dialectal propuesta por Ruíz (2020)

La definición dialectal por parte de Ruíz (2020) la expresa a partir de diferencias morfosintácticas y léxicas registradas especialmente en la región superdialectal del español neogranadino, en la zona sur resalta que el dialecto Amazónico carece de un conjunto de datos robusto para una definición de diferencias precisas, lo cual se puede ver de forma reiterada en la caracterización dialectal de otros autores; en el nivel subdialectal Ruíz (2020) sin embargo establece las diferencias a partir de los rasgos léxicos de las propuestas mencionadas en esta investigación que toman como materia prima los datos del ALEC.

6.4. Aplicaciones informáticas

Existen múltiples aplicaciones informáticas con herramientas estadísticas integradas que pueden permitir el desarrollo, diseño e implementación de un modelo espacial para la estimación de isoglosas y clasificación dialectal a partir de datos geolingüísticos del tomo III del Atlas Lingüístico Etnográfico de Colombia (ALEC), sin embargo dentro de esta investigación algunas de las de mayor incidencia en los resultados de este desarrollo son las siguientes.

6.4.1. Gabmap

Gabmap es una aplicación web de libre acceso y código abierto, según Leinonen et al. (2016) su principal enfoque es el análisis de información

lingüística y su variación en el espacio geográfico, sin embargo tiene algunas características diferentes, lo que le ha otorgado éxito en dialectometría, actualmente permite la automatización del análisis del lenguaje con el uso de los siguientes métodos:

Herramientas de análisis lingüistico Gabmap
Generación y visualización de mapas de distribución.
Medición de distancias lingüísticas.
Medición y visualización de dialectos continuos.
Identificación de grupos dialectales.

Cuadro 3: Métodos de análisis lingüistico Gabmap
(Leinonen et al., 2016)

Dentro de esta investigación el uso del aplicativo informático Gabmap se restringe a la implementación y cálculo de la medida de distancia lingüística propuesta Goebl (1987), el índice relativo de identidad (IRI), índice base para el diseño, desarrollo e implementación del modelo espacial a generar.

6.4.2. GeoDa

GeoDa es una herramienta de software de código abierto que sirve para el procesamiento y modelamiento de datos geográficos (Geoda, 2020), fue desarrollado por el Doctor Luc Anselin de la Universidad de Chicago hacia finales del año 2003, proporciona métodos de análisis exploratorio de datos espaciales, estadísticas de autocorrelación espacial, análisis de regresión espacial básico y algunos algoritmos de minería de datos espaciales para el agrupamiento de grandes conjuntos de datos con un componente geográfico.

Dentro de esta investigación el uso de Geoda abarca distintos procesos que se llevarán a cabo dentro de su desarrollo, en primer lugar el análisis exploratorio de autocorrelación espacial de las variables a incluir dentro del modelo, en segundo lugar la implementación de distintos modelos espaciales de regresión y por último, el agrupamiento y/o clasificación de la información empleando el algoritmo de minería de datos espaciales denominado regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (REDCAP) (Guo, 2008).

6.4.3. SpaceStat

El software SpaceStat (2020), lanzado en 1991, es un software desarrollado en inicio para el análisis de información geográfica dentro de econometría espacial, con el tiempo integró distintas herramientas espaciotemporales y estadísticas que permiten el análisis exploratorio de datos espaciales, el análisis econométrico espacial con ponderaciones espaciales, el modelado de datos a partir de semivariogramas así como la construcción e implementación de distintos métodos de interpolación espacial.

Dentro de esta investigación SpaceStat se utilizará para la implementación y validación del método de interpolación llamado: Kriging Ordinario, método a utilizar dentro del proceso de construcción de la superficie dialectal continua, superficie que a su vez será la base para el análisis de la distribución lingüística a lo largo del territorio así como para la propuesta de clasificación dialectal.

6.4.4. Python



Figura 8: Python para el análisis de información lingüística
(Python, 2020)

Según la pagina oficial (Python, 2020), Python es un lenguaje de programación multiparadigma, dinámico y multiplataforma, de fácil aprendizaje y eficiente, actualmente soporta licencia de código abierto y con los años se ha convertido en uno de los mayores referentes en el mundo computacional, soporta una extenso catálogo de librerías cuya documentación es robusta y dinámica, de tal forma que será utilizado para desarrollos tipo scripting y la implementación del algoritmo final del modelo propuesto.

Python dentro de esta investigación se utilizará para el desarrollo de múltiples scripts que automatizarán la selección de información correspondiente a los mapas léxicos del ALEC, el análisis dialectométrico de la información, así como su visualización, y finalmente la implementación y validación de la medida de similitud lingüística (calculada dentro del software en línea Gabmap), el índice relativo de identidad (IRI) (Goebl, 1987).

6.4.5. R

R es un lenguaje y a la vez un entorno de computación estadística desarrollado por Bell Laboratories (R, 2020), provee una amplia variedad de técnicas de análisis estadístico, entre ellas, modelos lineales, modelos no lineales, pruebas estadísticas clásicas, análisis de series de tiempo, agrupamiento jerárquico, etc. Quizás lo más llamativo de *R* es la facilidad que tiene para para la creación de librerías ya que dispone de rutas de código abierto para tal fin.

Para el tratamiento de información espacial *R* proporciona paquetes como *RGdal* y *SpatialReg*, paquetes utilizado dentro de esta investigación para la gestión de información geográfica así como para la creación e implementación de modelos espaciales, de tal manera que el modelo a desarrollar se puede crear e implementar en más de una aplicación informática.

6.4.6. ArcGIS

ArcGIS es un sistema informático que permite la recopilación, organización, administración, análisis y divulgación de información espacial (ArcGIS, 2020); dentro de esta investigación se utilizará el módulo ArcMap, una aplicación de escritorio para la creación de mapas, la realización de análisis espacial y la administración de datos geográficos, esta aplicación permitirá la creación de las principales visualizaciones así como la validación de resultados estadísticos de otras aplicaciones, pretendiendo así que los resultados sean veraces.

ArcMap cuenta con una extensión de procesamiento para el análisis geoestadístico (Geostatistical Analyst), con el uso de este modulo se realizará la exploración de los resultados así como la validación de las variables incluidas

dentro del modelo, verificando y ejecutando el método de interpolación Kriging Ordinario, necesario para la generación de la superficie dialectal continua y el mapa de isoglosas.

Capítulo II
Desarrollo de la investigación

7. Metodología



Figura 9: Metodología de la investigación

Ésta investigación utiliza una metodología cuantitativa de desarrollo, los resultados mostrarán cómo se relacionan los diferentes componentes de las mediciones geográficas y léxicas (ver capitulo dos - análisis de datos), la implementación de los resultados surge del análisis de iterativas observaciones y mediciones de algoritmos de agrupación y análisis multidimensional

(jerárquico) implementados en datos geolingüisticos de Colombia, los datos a utilizar provienen del tomo III del ALEC, estos fueron digitalizados recientemente mediante el diseño e implementación de una base de datos relacional en software libre (PostgreSQL), lo que garantiza disponibilidad, coherencia y menor redundancia en la información objeto de estudio.

El proceso de desarrollo metodológico se centra en seis fases (Figura 9), una fase exploratoria, cuyo propósito fue la búsqueda bibliográfica especializada referente a los temas a tratar, una fase de análisis de la información cuyo objetivo fue la organización de datos y la generación de un análisis estadístico del comportamiento de las variables geográficas con respecto a la variable lingüística, en tercer lugar el diseño de un modelo para la obtención de isoglosas y clasificación dialectal con información del tomo III del ALEC, cuarto la construcción de un modelo espacial a partir del uso de herramientas informáticas, quinto la implementación del modelo dentro de un entorno computacional y por último el diseño y desarrollo de un protocolo de pruebas para su evaluación, teniendo presente los resultados obtenidos por otros autores.

7.1. Exploración y recopilación de información

La fase exploratoria buscó la recopilación y lectura de información teórica y/o búsqueda bibliográfica acerca de información existente, cuya aplicación específica fuese la clasificación dialectal tanto a nivel local como internacional, esta búsqueda permitió tener una perspectiva clara de los resultados esperados a partir de procesos análogos desarrollados en contextos diferentes.

Entre la información encontrada los artículos más relevantes son:

- La dialectométrie dans l'Atlas linguistique de la Gascogne de Séguy (1973), quien propuso un método de análisis dialectométrico cuantitativo basado en las distancias lingüísticas y diferencias léxicas.
- Points chauds de l'analyse dialectométrique : pondération et visualisation de Goebl (1987), artículo que describe de manera formal el cálculo de las distancias lingüísticas por el método IRI e IPI, así como visualización de resultados preliminares .

- Dialectometry: A Short Overview Of The Principles And Practice Of Quantitative Classification Of Linguistic Atlas Data de Goebl (1991), este artículo presenta una aplicación de clasificación dialectal con los métodos propuestos por Goebl en dialectometría cuantitativa.

- Geolingüística: modelos de interpretación geográfica para lingüistas, un libro escrito en el Hernández Campoy (1999), en el se encuentra una de las premisas en las que se desarrolla esta investigación, los factores geográficos tales como la altura y la distancia lingüística son fuentes extralingüisticas de la variación dialectal.

- Toward a Dialectological Yardstick de Nerbonne y Kleiweg (2007), este artículo habla de cómo involucrar distintos parámetros para llegar a medidas de diferencia lingüistica obteniendo una coherencia en las mediciones y en los resultados obtenidos.

- On Quantitative Geolinguistics: An Illustration From Galician Dialectology de Dubert y Sousa (2016), uno de los más importantes. En el se busca proporcionar una visión general de las principales técnicas analíticas y de representación de información geolingüística. Además de ello, tiene la intención de demostrar, sobre la base de algunos ejemplos, nuevas metodologías y su utilidad para un mejor comprensión de la variedad espacial del idioma.

- Dialectones: Finding Statistically Significant Dialectal Boundaries Using Twitter Data de Rodríguez et al. (2018), en el se recoge información lingüística de twitter para encontrar las principales características lingüísticas de un idioma en una región particular. Para tal fin se utilizan isoglosas y métodos de la ecología tradicional.

- Dialectology for computational linguists de Nerbonne et al. (2019), los autores describen en este artículo una visión general del trabajo computacional en dialectología, centrándose en el procesamiento computacional, los supuestos de fondo y los problemas que pueden derivar de un estudio dialectométrico profundo.

Una vez contextualizados con la información relacionada al problema de investigación, se desarrollo un análisis exploratorio de la información léxica contenida en el tomo III del ALEC, donde se encontró la estructuración de ésta en una base de datos relacional desarrollada en PostgreSQL.

Haciendo uso del convenio existente entre la Universidad Distrital Francisco José de Caldas y el Instituto Caro y Cuervo, se logró acceder a esta base de datos cuya información contiene más de 300 mapas con información lingüística y más de 25.000 registros de datos relacionados a las variantes de cada uno de los mapas, estos datos corresponden a información obtenida entre los años 1956-1978 (Florez, 1983) (fecha en que se recolectó la información del ALEC), además de ello existen metadatos correspondientes a cada localidad, entre los cuales están:

Metadatos
Altura sobre el nivel medio del mar
Temperatura
Vías de Acceso
Actividad Económica Predominante

Cuadro 4: Metadatos de cada localidad ALEC

De estos datos se extrajo una muestra inicial referente a cada una de las localidades. Con especial relevancia fueron tenidos en cuenta los datos de altura sobre el nivel medio del mar y temperatura, variables geográficas que pueden ser tenidas en cuenta dentro del modelo, esto ya que según Hernández Campoy (1999, pg. 81) son factores físicos que intervienen en la difusion de los cambios lingüísticos constituyendo en la dialectología tradicional núcleos que determinan modos o rasgos separados por líneas imaginarias denominadas isoglosas, de hecho en Inglaterra, una de las fronteras Dialectales más notable atraviesa una zona pantanosa aislada apodada 'the Fens', lo que expresa de forma clara como las barreras físicas pueden ser fronteras lngüisticas medibles.

Finalmente, esta fase comprendió la búsqueda de software y/o aplicaciones relacionadas con clasificación dialectal y dialectometría, esto con el fín de desarrollar, hacer pruebas e implementar el modelo propuesto, entre los software a utilizar se encuentra:

7.2. Análisis de datos

Según Chambers y Trudgill (1998) la lengua y su variación se pueden estudiar desde la dialectología haciendo uso de métodos proporcionados por la geografía lingüística, métodos donde la dimensión espacial cobra un sentido relevante; es por esto que el enfoque del siguiente desarrollo consiste en la realización de análisis cuantitativo y estadístico de información lingüística y geográfica, para ello en esta etapa se consideran algunas variables cuantitativas a utilizar en el modelo, las variables geográficas son continuas y son utilizadas como índices climáticos y geológicos (riesgo sísmico, altura sobre el nivel del mar, entre otros), cabe resaltar que todos los datos fueron suministrados por entidades de orden estatal cuyo enfoque es netamente geográfico lo cual supone una veracidad anticipada de la información.

La teoría lingúistica expresada por Hernández Campoy (1999) manifiesta que los cambios lingüísticos son producidos por condiciones cuyas causas son intralingüísticas o sistemáticas y/o causas extralingüísticas o extrasistemáticas, las primeras son aquellas relacionadas con mecanismos internos propios de los sistemas lingüístico, mientras que las segundas tienen un valor significativo ya que son aquellas referidas a la influencia de posibles factores externos.

Los factores extralingüísticos en esencia no guardan ninguna relación con los fenómenos lingüísticos propios del sistema, pero según Hernández Campoy (1999) al igual que la clase social, la edad, la raza o la religión tienen una importancia primordial en la diferenciación de la lengua, su difusión y su cambio, desde esta perspectiva los factores físicos de tipo geográfico, como los ríos, las montañas, la distancia euclidiana, y grandes fenómenos espaciales intervienen de manera fundamental en la propagación lingüística.

En palabras de Labov (1995), la separación geográfica conduce de forma natural e inevitable a la separación lingüística, demostrando que generalmente la configuración de determinadas isoglosas en mapas lingüísticos guardan una alta relación con determinadas barreras físicas, de tal manera que sitios aislados separados por valles o por colinas, generalmente mantienen una diferencia lingüística considerable, difiriendo con aquellos sitios cuya barrera geográfica es una vía principal de comunicación de núcleos urbanos. De he-

cho, como ya se mencionó, una barrera dialectal en Inglaterra, atraviesa una zona pantanosa de difícil acceso denominada 'the Fens' entre los condados de Cambridge y Lincoln.

Como barreras geográficas dentro del desarrollo de esta investigación se tomarán aquellos límites naturales de distribución espacial referentes a un fenómeno geográfico (Cuadro 5), entre los cuales están, el índice de disponibilidad hídrica, la aptitud agroclimática media, la precipitación promedio , la temperatura promedio multianual, entre otros, a continuación se ahondará en cada uno de ellos.

Datos	**Tipo**	**Entidad**
Actividades Económicas	Categórico	Metadatos ALEC - Instituto Caro y Cuervo
Altura sobre el nivel medio del mar	Numérico	Instituto GeográficoAgustín 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
Similitud Lingüística	Numérico	Resultado IRI Instituto Caro y Cuervo
Distancia Geográfica	Numérico	Resultado formula de Vincenty Instituto Caro y Cuervo

Cuadro 5: Datos a utilizar dentro del desarrollo

7.2.1. Localidades seleccionadas

Las localidades seleccionadas fueron aquellas en las que el número de respuestas era considerable, es decir, aquellas localidades cartografiadas inicialmente en los tomos impresos del ALEC (238 localidades); la distribución de estas a lo largo del territorio tiene una desventaja, ya que abarca la parte central y norte del territorio colombiano principalmente, es decir, su densidad espacial se concentra en la zona centro-norte, por lo que para la región sur sólo se puede encontrar una localidad, aquella que se encuentra en Leticia, departamento del Amazonas, esto se debe a que en el resto del territorio para la época no había suficientes hablantes de español, predominando por entonces las lenguas indígenas, así mismo, eran regiones de difícil acceso. La distribución geográfica de estos lugares se puede observar en la figura 10

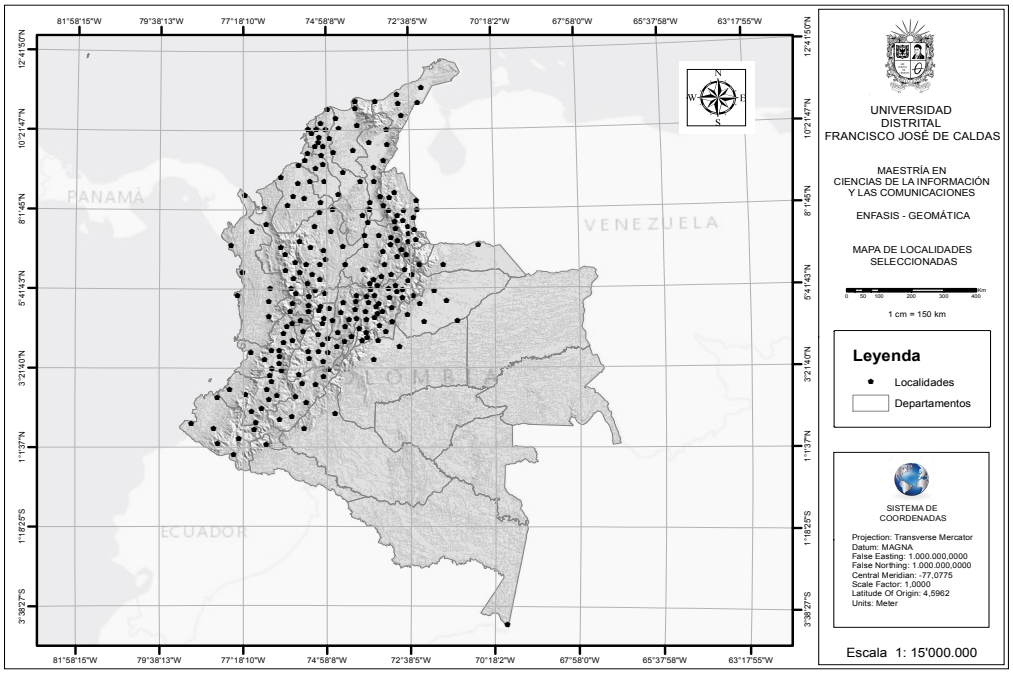


Figura 10: Localidades seleccionadas

Es importante tener presente que en el momento en que se seleccionaron las localidades a visitar se tuvieron en cuenta criterios geográficos, semejantes a aquellos con los que se va a trabajar en el desarrollo de este modelo, criterios tales como la temperatura, la altura sobre el nivel del mar, entre otros, la información adicional como vías de acceso, actividad económica predominante se utilizaran en un sentido cuantitativo, utilizando variables de tipo ficticio o dicotómico, lo cual mejorará la precisión del modelo, a continuación se hablará acerca de cada una de ellas y su relación con las localidades.

7.2.2. Actividades económicas predominantes

Entre los metadatos de cada localidad se recogieron aquellos relacionados con las actividades económicas presentes en cada una de ellas; para mejorar la calidad del modelo se utilizará aquella actividad principal que genera valor, por ejemplo para la localidad $Ch03$ (Coredó - Departamento del Chocó) un corregimiento para entonces del municipio de Juradó, al sureste de esta cabecera municipal, en la costa de la bahía de Humboltd en donde las actividades económicas son la pesca y la agricultura, sólo se tomará la actividad

principal según el orden de aparición, es decir, la pesca; entre las principales actividades económicas tenemos las siguientes:

Número	Actividad Económica
1	Agricultura
2	Comercio
3	Ganadería
4	Industria
5	Minería
6	Pesca

Cuadro 6: Actividades Económicas principales

La actividad predominante es la agricultura, seguida por la ganadería y la pesca, finalmente se encuentra la minería, el comercio y la industria, tal como se muestra en la figura 11.

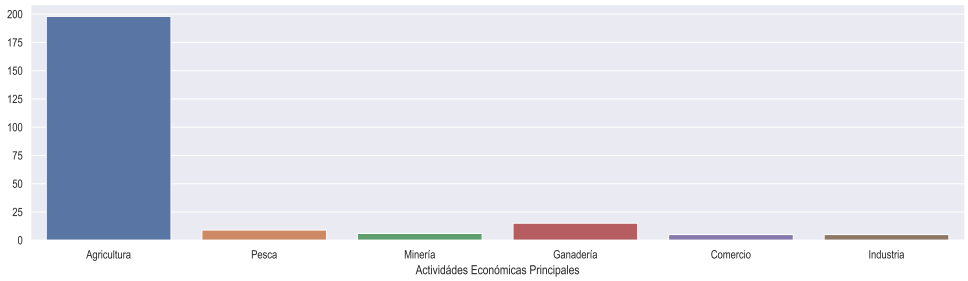


Figura 11: Actividades Económicas predominantes

Dado que la información de las actividades económicas se encuentra en forma cualitativa, su transformación a variables ficticias (variables reclasificadas: 0 - 1) es necesaria para su interpretación dentro del modelo cuantitativo, una breve visualización sería de la siguiente manera:

Localidad	Agricultura	Minería	Pesca	Actividad
Ch01	1	0	0	Agricultura
Ch10	1	0	0	Agricultura
Ch30	1	0	0	Agricultura
Ch03	0	0	1	Pesca

Cuadro 7: Actividades Económicas - Variables reclasificadas

Es de resaltar que dentro de la matriz se toman los valores cero (0) para aquellos elementos muéstrales en los que no se presenta una característica y

uno (1) cuando se presenta; en el cuadro 7 las actividades económicas para las primeras cuatro localidades son la Agricultura y la Pesca, la Minería sin embargo no presenta ningún elemento asociado.

7.2.3. Atura media sobre el nivel medio del mar

La altura sobre el nivel medio del mar fue tomada de los metadatos de la información recopilada en la base de datos espacial del Instituto Caro y Cuervo, sin embargo dado que la información tenia valores nulos para algunas de las 238 localidades fue necesario recurrir al Instituto Geográfico Agustin Codazzi, ente rector de la información cartográfica de nuestro país, para ello, se tomaron las curvas de nivel del territorio nacional (Figura 12) a una escala 1:50.000, estas se interpolaron por el método IDW (Distancia Inversa Ponderada), método que permite generar una superficie continua con la cual se pueden tener valores aproximados de la altura sobre el nivel medio del mar con una precisión de 12.5 m.

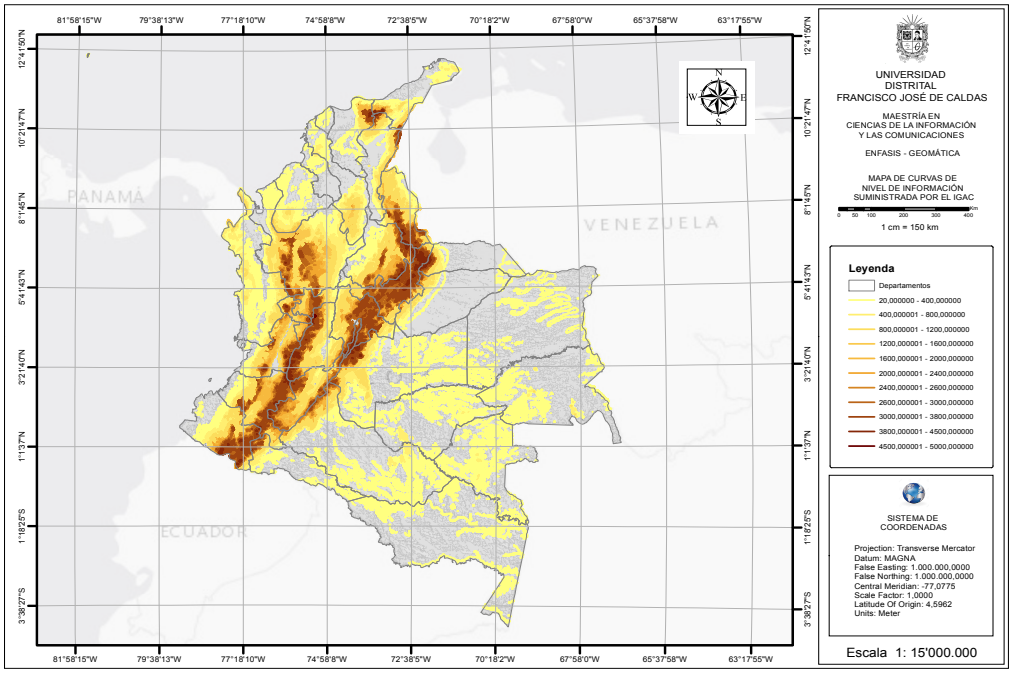


Figura 12: Información básica IGAC, altura sobre el nivel medio del mar

Una vez hecho esto se actualizó la información correspondiente a cada

localidad y su altura media sobre el nivel medio del mar, dado que es un valor flotante no fue necesaria ningún tipo de transformación, los valores más relevantes son:

Mínimo:	80 m
Máximo	3833,12793 m
Suma	301365,295982 m
Medía	1266,240739 m
Desviación Estándar	1064,5801 m

Cuadro 8: Valores referentes para la altura sobre el nivel medio del mar

Su distribución dentro del territorio colombiano sería de la siguiente manera (Figura 13):

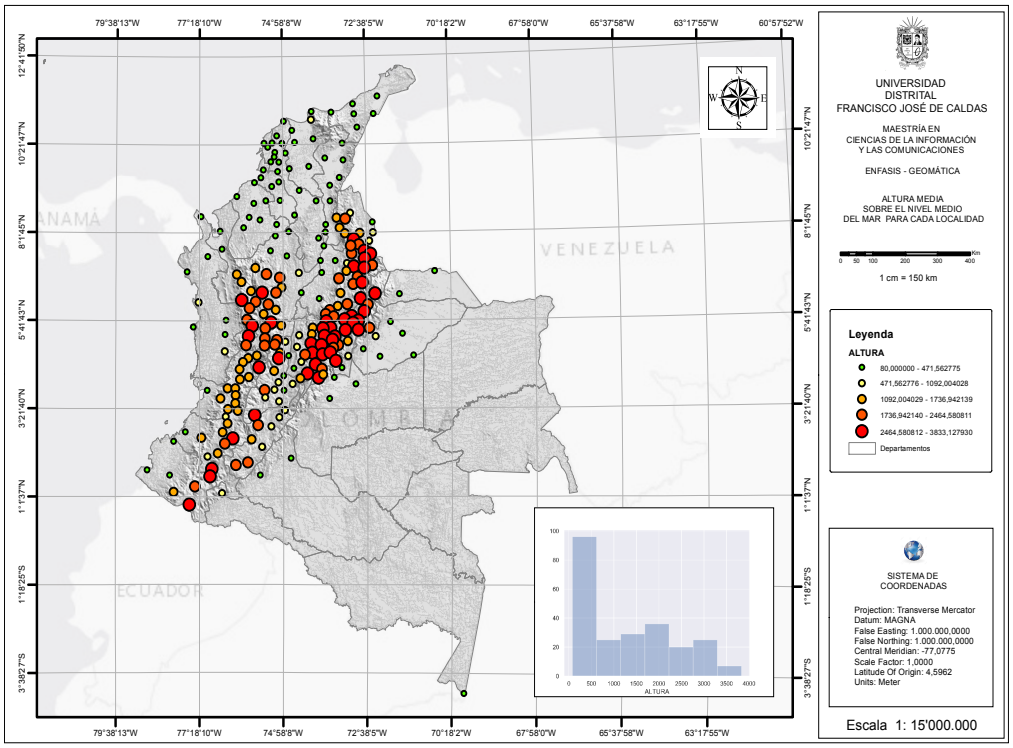


Figura 13: Distribución espacial, altura sobre el nivel medio del mar

Se puede concluir que la mayor parte de localidades se encuentra a una altura entre ochenta y quinientos metros sobre el nivel medio del mar, adi-

cionalmente que estas localidades se encuentran en la parte norte del país, distribuidas entre la región Atlántica y la región del Pacífico.

7.2.4. Vías de acceso

Las vías de acceso correspondientes a la época en la que se realizó la encuesta para cada localidad fueron obtenidas de los metadatos recopilados en la base de datos espacial del ALEC, estas vías de acceso se encuentran catalogadas de la siguiente manera:

Número	**Vía de Acceso**
1	Caminos carreteables
2	Aérea
3	Caminos de herradura
4	Carretera
5	Carretera y férrea
6	Fluvial
7	Férrea

Cuadro 9: Vías de acceso a localidades

Dado que la información de estas vías de acceso se encuentra en formato cualitativo, fue necesaria su transformación a variables ficticias de tal manera que se pudiesen interpretar dentro del modelo cuantitativo, de la siguiente manera:

Localidad	**Aérea**	**Carretera**	**Fluvial**	**Férrea**	**Vía de Acceso**
Ch01	0	0	1	0	fluvial
Ch10	0	0	1	0	fluvial
Ch30	0	0	1	0	fluvial
Ch03	0	1	0	0	carretera
Ch04	0	1	0	0	carretera

Cuadro 10: Vías de acceso a localidades - Variables ficticias o reclasificadas

Finalmente los datos de frecuencia muestran como la vía de acceso predominante es la vía tipo Carretera, seguida de la vía fluvial, tal como se muestra en el siguiente gráfico:

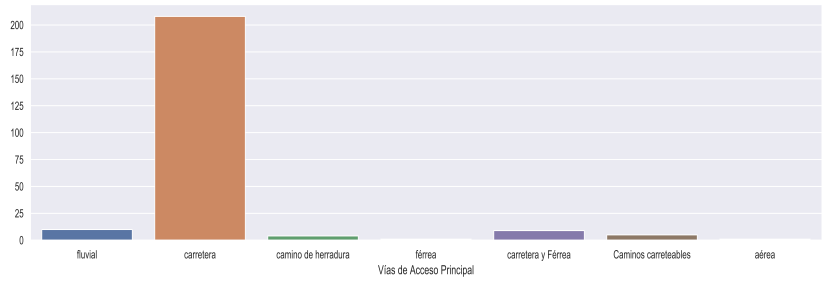


Figura 14: Histograma de las vías de acceso de localidades

7.2.5. Temperatura media

Los valores de temperatura fueron tenidos en cuenta para la selección inicial de las localidades, valores que como variable geográfica permiten conocer la magnitud de energía presente en un espacio geográfico; al igual que con otros metadatos al revisar la base de datos espacial para el ALEC existían valores nulos por lo tanto se decidió tomar aquellos valores que residen en la base de datos abierta del IDEAM, entidad que proporcionó el promedio de temperatura multianual del intervalo de tiempo 1981-2010; estos datos se pueden visualizar de la siguiente manera:

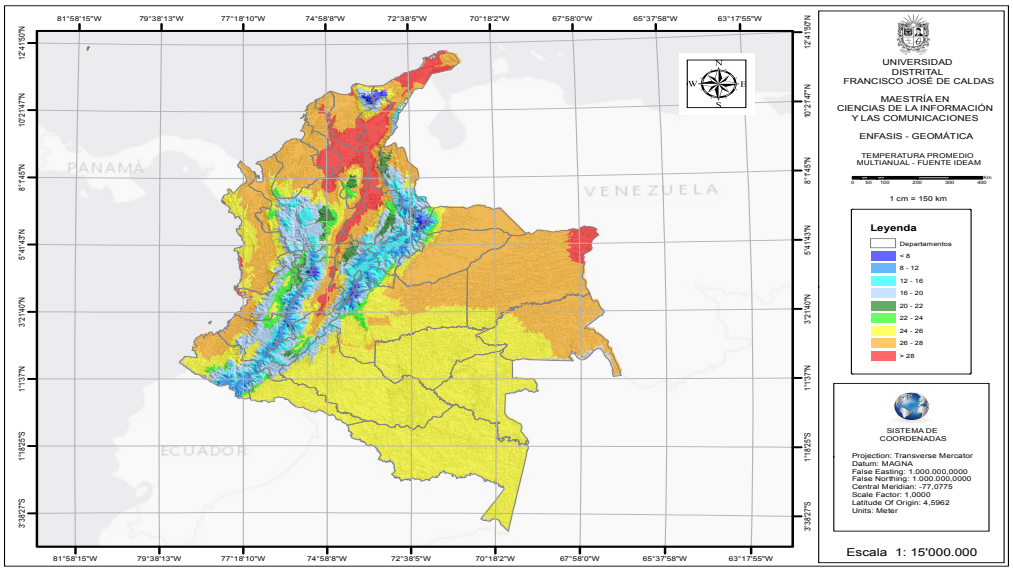


Figura 15: Temperatura promedio multianual - fuente IDEAM

Dado que los valores se encuentran en intervalos, se tomaron como variables cualitativas inicialmente, una vez realizada la intersección espacial con la ubicación de cada localidad se pudo relacionar esta temperatura con cada sitio visitado, se realizó su análisis y respectivo diagrama (Figura 16), que permitió concluir que los valores de temperatura para todas las localidades se concentran principalmente en el rango que va entre 26 y 28 grados Celcius, seguido del que se encuentra entre 16 y 20 grados Celcius.

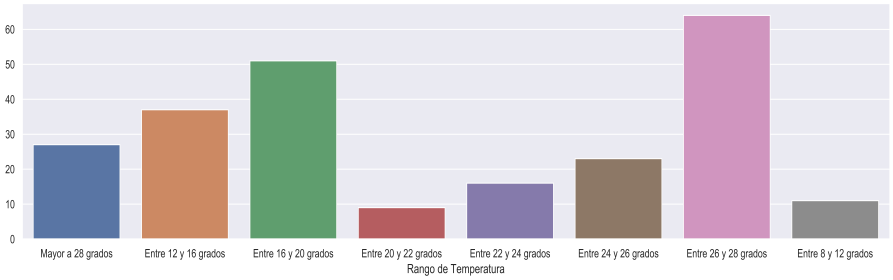


Figura 16: Temperatura promedio multianual - Histograma

Finalmente los valores cualitativos fueron transformados a valores cuantitativos con el uso de variables dicotómicas, quedando del siguiente tipo:

Localidad	Entre 24 y 26 grados	Entre 26 y 28 grados	Mayor a 28 grados	Rango Temperatura
Ch01	0	1	0	Entre 26 y 28 grados
Ch10	0	1	0	Entre 26 y 28 grados
Ch30	1	0	0	Entre 24 y 26 grados
Ch03	1	0	0	Entre 24 y 26 grados

Cuadro 11: Temperatura en variables ficticias o reclasificadas

Como se mencionó este tipo de matriz toma los valores cero (0) para aquellos elementos muestrales en los que no se presenta una característica y uno (1) cuando se presenta.

7.2.6. Aptitud agroclimática

Salvatore (2009) define la aptitud agroclimática como la capacidad de un determinado lugar de producir un cultivo especifico con base en las condiciones agroclimáticas y del suelo, la aptitud agroclimática expresa en niveles cuantitativos y cualitativos el nivel de satisfacción de una especie agrícola en una región según sus exigencias bioclimáticas requeridas (Pascale, 2010),

es decir, fija, la zonificación de cultivos a partir de la identificación de áreas potenciales, las cuales surgen como una integración de variables tales como el suelo, el clima, la precipitación, entre otras (Jiménez, 2004).

En esta investigación los datos de aptitud agroclimática media son fijados en el intervalo temporal 1981 - 2010, datos próximos a la realización de la encuesta bajo la cual surgió el ALEC, para su determinación en cada localidad se realizo la intersección de cada polígono de información con la posición de cada localidad, la figura 17 representa los diferentes niveles que rigen la clasificación del IDEAM.

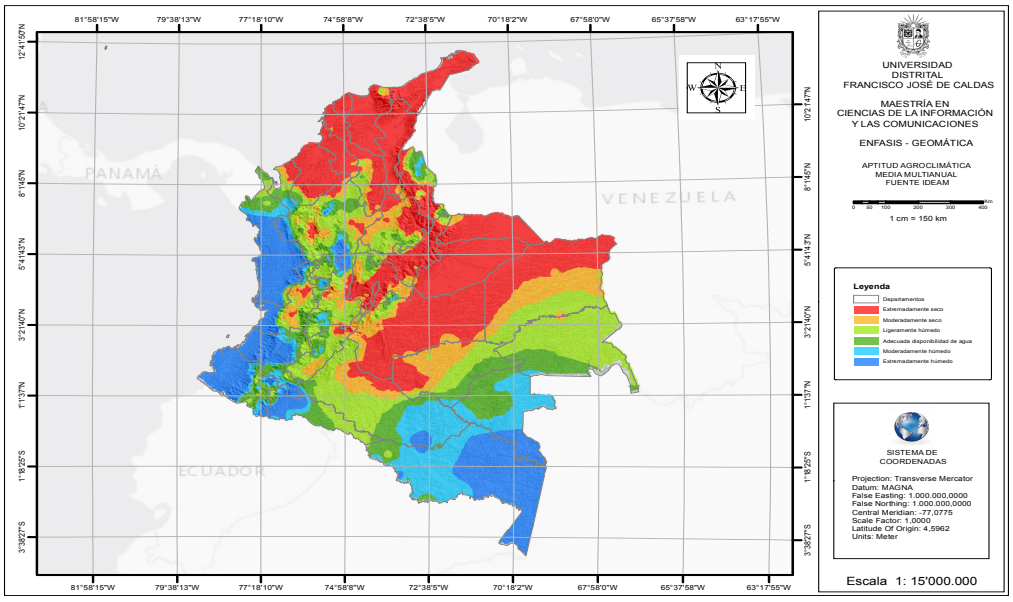


Figura 17: Aptitud agroclimática promedio multianual - fuente IDEAM

Niveles de clasificación agroclimática	
1	Extremadamente seco
2	Moderadamente seco
3	Ligeramente Húmedo
4	Adecuada disponibilidad de agua
5	Moderadamente húmedo
6	Extremadamente húmedo

Cuadro 12: Niveles agroclimáticos según el IDEAM

Dada la normalización de esta información en un tipo de variable cualita-

tiva, para su manejo dentro del modelo y desarrollos posteriores se realizó su transformación en variables ficticias, variables que se incluyen con el propósito de tener en cuenta variaciones de tipo cualitativo en un modelo cuantitativo.

Estas variables también llamadas dummys, binarias o dicotómicas toman los valores cero (0) para aquellos elementos muestrales en los que no se presenta una característica y uno (1) cuando se presenta, generalmente este tipo de variables son explicativas en modelos de regresión, donde el regresando y las demás variables explicativas son variables cuantitativas (Barreiro et al., 2004), como en este caso, donde la variable a explicar es la diferencia lingüística, a continuación se muestra la representación de la aptitud agroclimática en variables de este tipo:

Localidad	Adec	Ex.Húm	Ex.Seco	Lig.Húmedo	Mod. Húmedo	RANGO
Ch01_	1	0	0	0	0	Adecuada disponibilidad
Ch10_	0	0	0	1	0	Ligeramente húmedo
Ch30_	0	1	0	0	0	Extremadamente húmedo
Ch03_	0	0	0	0	1	Moderadamente húmedo
Ch04_	0	1	0	0	0	Extremadamente húmedo

Cuadro 13: Diferencia en valores de aptitud agroclimática

Como se observa, la información de la tabla 13 hace referencia a una matriz cuyos valores representan para cada localidad el valor de referencia de la aptitud agroclimática, en este caso relacionadas por el código que se utilizo en la cartografía digital del ALEC.

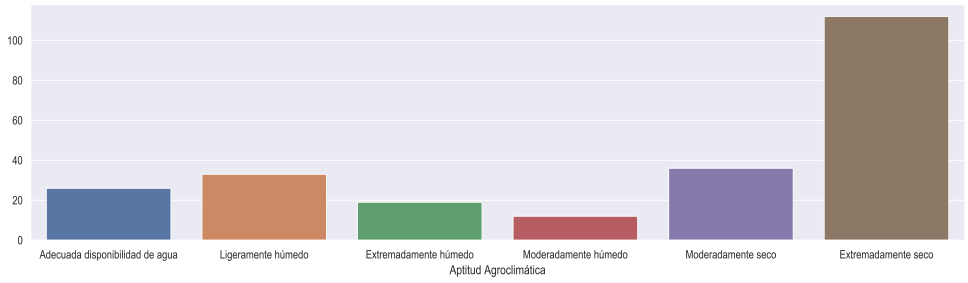


Figura 18: Aptitud agroclimática promedio multianual - Histograma

Finalmente se muestra como la concentración de valores de aptitud agroclimática se concentra principalmente en las zonas extremadamente secas,

que tienen una relación relativa a los valores de temperatura mencionados atrás.

7.2.7. Índice de disponibilidad hídrica (IDH)

El índice de disponibilidad hídrica es, al igual que la aptitud agroclimática, un indicador, sin embargo el IDH se utiliza para la identificación de zonas con periodos de exceso o deficiencia de agua, su cálculo se hace a partir del valor de evotranspiración potencial donde se tiene en cuenta la oferta y demanda de agua según el nivel de clasificación del tipo de tierra, desde aquellas que son muy secas, hasta aquellas que son muy húmedas.

Para la correcta clasificación climática el IDEAM ha adoptado una metodología utilizada en el ámbito internacional, esta metodología garantiza que se puedan analizar tanto las deficiencias hídricas como los excesos hídricos, para ello se utiliza el factor de humedad Thornthwaite; sin embargo dado que no existe semejanza total con la diversidad Climática de Colombia se hicieron algunas modificiaciones, por ejemplo para el rango de -60 a cero la clasificación del factor Thornthwait incluye tipos climáticos árido y semiárido, en Colombia los valores encontrados en este nivel son aproximadamente 2/3 de los valores totales, es decir, se hizo indispensable ajustar los valores de humedad a las condiciones climáticas del país, para ello se diseño el Índice de Disponibilidad Hídrica, este índice facilita la relación de deficiencia con exceso hídrico, la ecuación que permite su cálculo es la siguiente:

$$IDH = \frac{ET + (0{,}25 * EX)}{ETP} * 100 \tag{2}$$

En donde:

- IDH = Índice de Disponibilidad Hídrica
- ET = Evotranspiración real
- ETP = Evotranspiración potencial
- Ex = Excesos de agua

El índice de disponibilidad hídrica se clasifica entonces en las siguientes categorías según el IDEAM:

Número	Categoría
1	Adecuado
2	Húmedo
3	Seco
4	Semihúmedo
5	Semiseco

Cuadro 14: Categorías IDEAM - IDH

Categorías que visualmente se pueden representar en el siguiente gráfico:

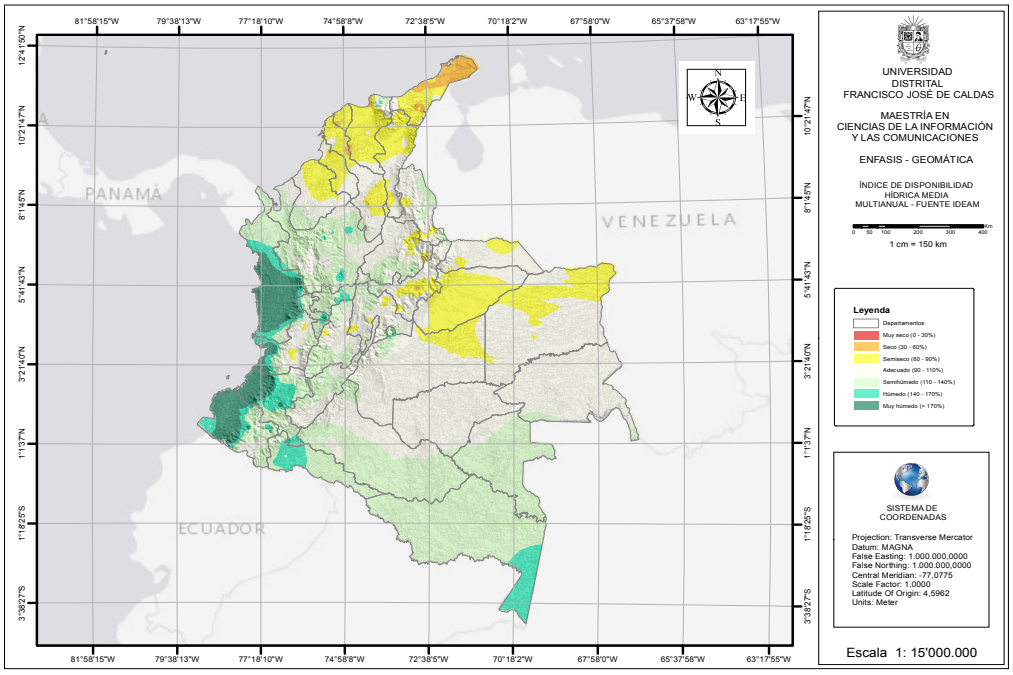


Figura 19: IDH promedio multianual

Recopilada esta información se realizó una intersección con la información espacial de cada una de las localidades, a partir de ello se logró obtener las variables ficticias correspondientes, de tal manera que dentro del modelo se pudiese tratar como información numérica.

Localidad	Adecuado	Húmedo	Seco	Semihúmedo	Semiseco	Condición IDH
Ch01	0	0	0	0	1	Semiseco
Ch10	1	0	0	0	0	Adecuado
Ch30	0	1	0	0	0	Húmedo
Ch03	0	0	0	1	0	Semihúmedo

Cuadro 15: IDH - Variables Ficticias o reclasificadas

Finalmente se obtuvo un histograma, que representa las distribuciones de frecuencias del IDH de las localidades, este permite concluir que las localidades se encuentran en una zona con IDH principal de tipo Semiseco, seguido a su vez del Adecuado, su gráfico se observa a continuación:

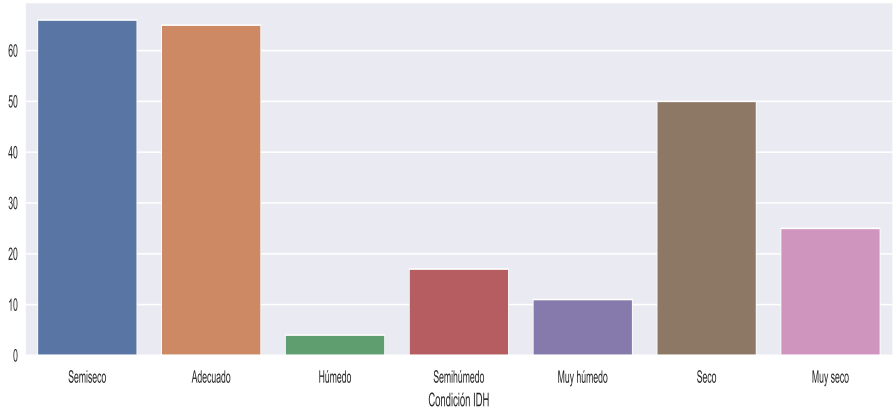


Figura 20: IDH promedio multianual por localidades - Histograma

7.2.8. Riesgo por amenaza sísmica

El riesgo de sismicidad representa para cada sitio o zona de estudio la intensidad sísmica estimada expresada en categorías diferentes de amenaza, la entidad encargada de proveer esta información es el Servicio Geológico Colombiano, quien definió una base de datos espacial con la información sísmica de todo el territorio nacional, este catálogo sísmico permitió construir un modelo de zonificación que permite delimitar por geometrías la ocurrencia de sismicidad equiprobable en el tiempo y espacio siguiendo una distribución de magnitudes de tipo Gutenberg-Richter (Servicio Geológico, 2020).

La visualización de las categorías de amenaza proporcionadas por el Servicio Geológico son las siguientes:

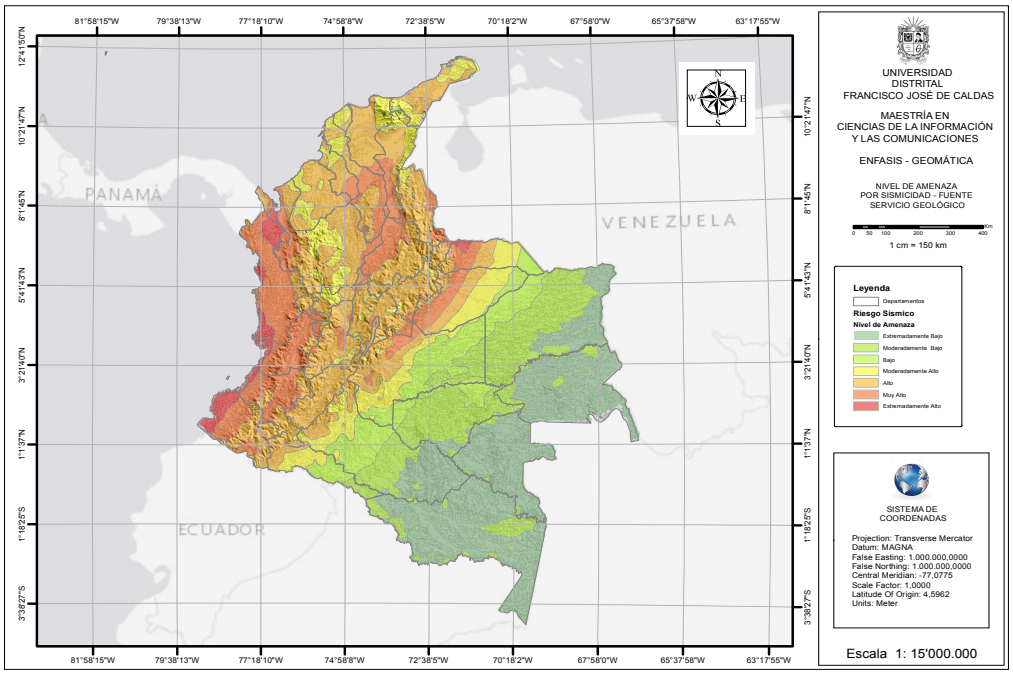


Figura 21: Riesgo por Amenaza Sísmica - Fuente: Servicio Geológico

Categorías que se dividen de la siguiente manera:

Número	Categoria
1	Extremadamente Bajo
2	Moderadamente Bajo
3	Bajo
4	Moderadamente Alto
5	Alto
6	Muy Alto
7	Extremadamente Alto

Cuadro 16: Amenaza Sísmica - Categorías

Al relacionar la información espacial de amenaza sísmica con la información espacial de las localidades se pudo constatar que los niveles bajo

los cuales se encuentra la mayor parte de localidades son solamente 5, Alto, Extremadamente Alto, Extremadamente Bajo, Moderadamente Alto y Muy Alto, tal como se muestra en el siguiente gráfico:

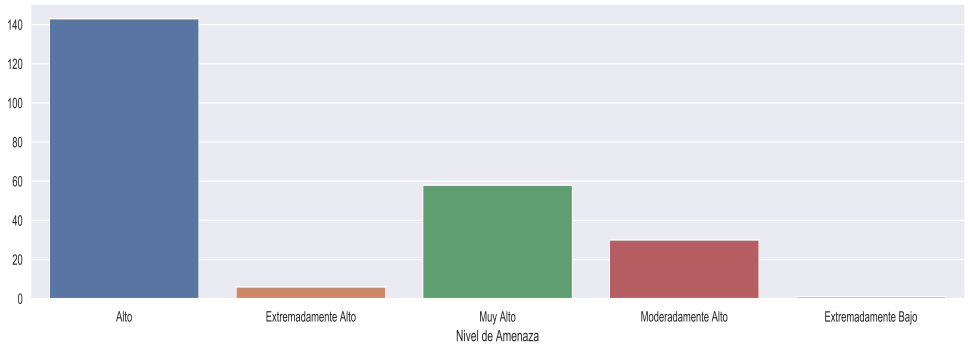


Figura 22: Amenaza Sísmica Localidades- Histograma

Finalmente al igual que con otros objetos geográficos mencionados atrás, la información cualitativa se transformo en variables ficticias para su análisis y manejo dentro del modelo cuantitativo.

Localidad	**Alto**	**Extre. Alto**	**Muy Alto**	**Nivel de Amenaza**
Ch01	1	0	0	Alto
Ch10	0	1	0	Extremadamente Alto
Ch30	0	1	0	Extremadamente Alto
Ch03	0	0	1	Muy Alto

Cuadro 17: Amenaza Sísmica - Variables reclasificadas ejemplo

7.2.9. Precipitación media multianual

Este objeto geográfico según el IDEAM, representa la distribución espacial de la precipitación media total anual medida en milímetros sobre el territorio colombiano, durante el periodo de tiempo comprendido entre el año 1981 y 2010, los datos fueron tomados de las estaciones meteorológicas en manos del IDEAM cuyos rangos más representativos se encuentran entre los 0 y los 11000 milímetros, los datos se pueden visualizar en la Figura 23, donde la mayor parte del territorio colombiano se encuentra ubicado en una

zona con valores de precipitación que se encuentran en el rango de 2000 a 3000, principalmente en la zona sur, departamentos del Amazonas, Vaupés, Vichada, Guainia, entre otros.

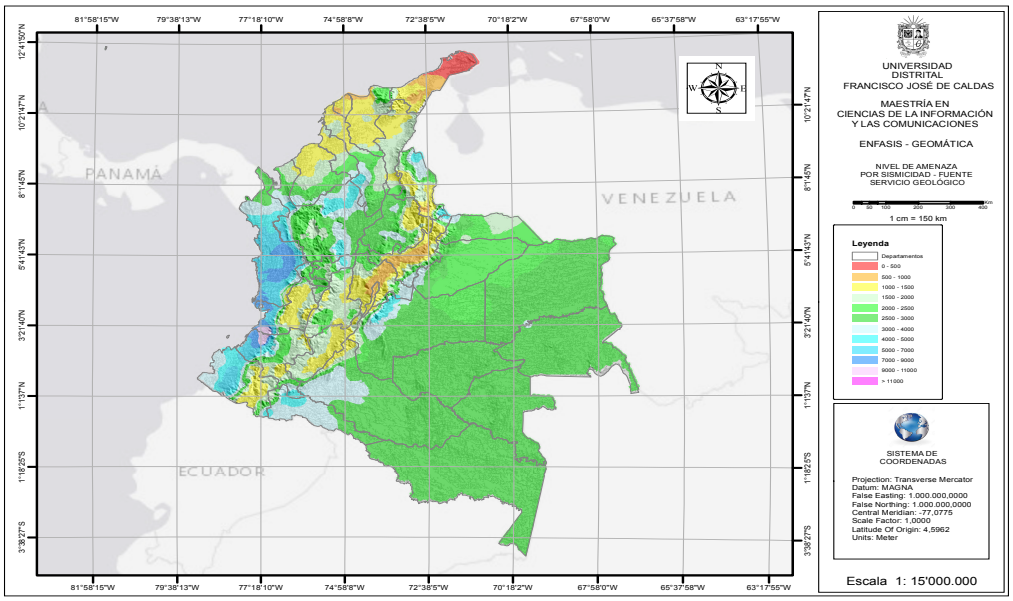


Figura 23: Precipitación media multianual - Fuente IDEAM

La relación espacial con las localidades se generó a partir de la intersección de coordenadas, los rangos fueron evaluados con el uso de diagramas de frecuencia, permitiendo así tener una perspectiva acerca del comportamiento de este objeto en las localidades visitadas.

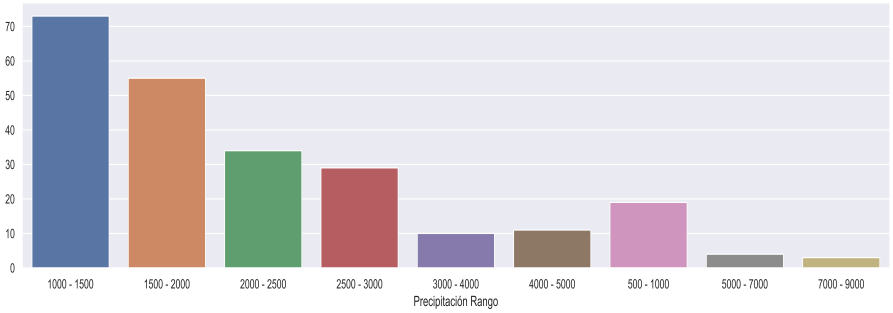


Figura 24: Precipitación Localidades - Histograma

Finalmente al igual que con otros objetos geográficos mencionados atrás, la información cualitativa correspondiente a la Precipitación multianual, se transformó en variables dicotómicas para su análisis y manejo dentro del modelo cuantitativo.

Localidad	Entre 2500 y 3000	Entre 4000 y 5000	Entre 5000 y 7000	Nivel
Ch01	1	0	0	Entre 2500 y 3000
Ch10	1	0	0	Entre 2500 y 3000
Ch30	0	0	1	Entre 5000 y 7000
Ch03	0	1	0	Entre 4000 y 5000

Cuadro 18: Precipitación - Variables reclasificadas

7.2.10. Similitud lingüistica

Como ya se mencionó según Goebl (1987) existen dos posibilidades de medir la proximidad o semejanza entre dos vectores de una matriz de datos lingüisticos:

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

En esta investigación se utilizará el Indice Relativo de Identidad (IRI_{jk}), una medida de similitud no ponderada entre vectores j , k, conocida en Alemán como "Relativer Identitätswert (RlW_{jk})", nombre adoptado por Goebl (1987); para ello inicialmente se requirió contar con una matriz de datos lingüisticos, datos que fueron tomados de los mapas léxicos del tomo III del ALEC.

Para ello inicialmente a partir de la base de datos que reside en PostgreSQL se realizó una consulta en lenguaje SQL que permitió obtener los mapas cartografiados del tomo III, tomando particularmente aquellos ya que tienen el mayor número de datos para cada una de las localidades, una vez consultados todos los mapas cartografiados se obtuvieron aquellos mapas Léxicos.

Los mapas léxicos contienen aquellas palabras que caracterizan un lugar, en este caso el espacio geográfico en que se realizaron las encuestas fue denominado Localidad, lugar donde se recogió la información referente al uso

de diferentes elementos lingüísticos para referirse a un mismo concepto, esta información se depuro de elementos etnográficos, fonéticos y suplementos adicionales; la selección de esta información se hizo por medio de Scripts de Python, donde se seleccionaron los mapas (100) con menor cantidad de valores nulos de la primera variante de cada entrada por localidad, tal como se observa en la siguiente tabla:

Id Mapa	**Valores Nulos**
3066	0
3243	0
3114	0
3067	0
3063	1
3064	1
3170	1

Cuadro 19: Valores nulos - Determinación de mapas léxicos

Seleccionada la información del archivo consultado en la base de datos espacial, se realizó la relación con las localidades y se obtuvo la información final (Cuadro 20), la cual fue procesada para generar la matriz de similitud; esta matriz fue generada siguiendo el algoritmo IRI de Goebl (1987) en Scripts de Python y verificada en el software en línea Gabmap, software especializado para el tratamiento de información lingüistica.

Id Localidad	**3001**	**3002**	**3019**	**3032**	**3033**	**3044**
1	enamorar	anillo	comadrona	tetero	arrullar	quitar el seno
2	enamorar	argolla	comadrona	tetero	arrullar	despechar
3	enamorar	argolla	comadrona	tetero	arrollar	destetar
4	enamorar	argolla	comadrona	tetero	cantar	destetar
5	enamorar	argolla	comadrona	tetero	arrullar	destetar

Cuadro 20: Datos finales - Ejemplo de información subida y procesada en Gabmap

La información lingüística obtenida, fue una matriz cuadrada que muestra la relación entre localidades con su respectiva **diferencia** lingüística:

	Ch01	Ch10	Ch30	Ch03
Ch01	0	0,445945946	0,581081081	0,527027027
Ch10	0,445945946	0	0,587837838	0,506756757
Ch30	0,581081081	0,587837838	0	0,358108108
Ch03	0,527027027	0,506756757	0,358108108	0

Cuadro 21: Matriz diferencia lingüistica - Relación por Localidad

Esta diferencia lingüística es igual a uno cuando no tiene ninguna coincidencia con ninguna de las localidades, es decir, para la localidad Ch01, la diferencia es cero cuando todas las respuestas en otra localidad son iguales, y distinta de cero cuando una o mas localidades tienen alguna respuesta equivalente.

Hecho esto se determinó la media para los valores de cada localidad, es decir, el valor medio de similitud para realizar los posteriores cálculos dentro del modelo (Cuadro 22), la distribución de los valores de similitud lingüística y la media de la distancia lingüística se muestran en la Figura 25 es de resaltar que el uso de la media es utilizado para la visualización y agrupamiento por Goebl (1987), sin embargo en este proyecto sólo se utilizo para reducir el número de valores al número de localidades.

Localidad	Similitud Lingüística
Ch01	0.5114845791875694
Ch10	0.5484902070044664
Ch30	0.5850693129212045
Ch03	0.5211848900010854

Cuadro 22: Similitud lingüística media por Localidad

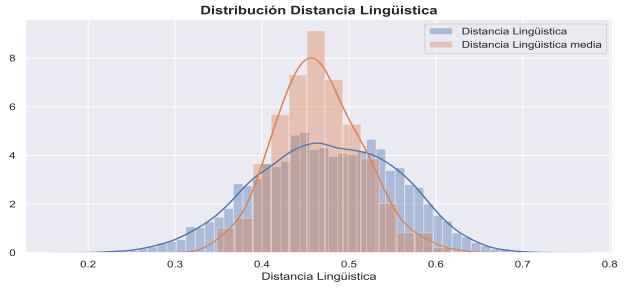


Figura 25: Similitud lingüística (Az)- Similitud lingüística media (Narj)

Como se puede observar la distribución de probabilidad de los valores de similitud lingüística son de tipo Normal, con una campana Gaussiana que resalta en el gráfico de comparación de similitud lingüística (Figura 25), para conocer si la distribución de probabilidad de los datos de similitud lingüística son equivalentes a los valores de la distribución Normal se realizó el gráfico $QQ - Plot$, de la siguiente manera:

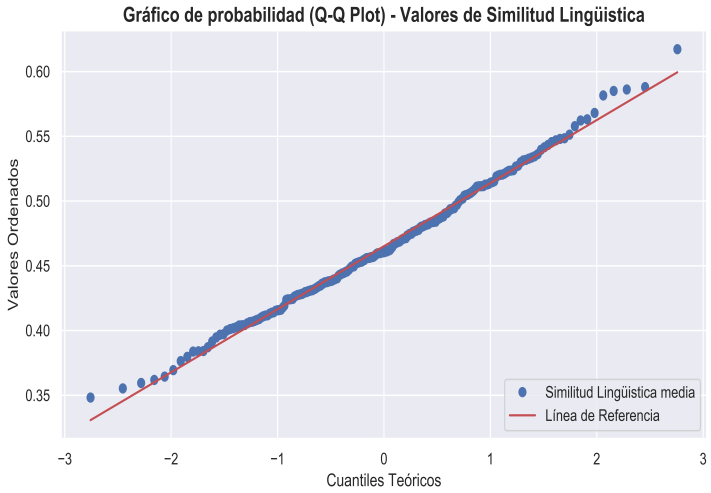


Figura 26: Gráfico de probabilidad QQ similitud lingüistica media

La figura 26 muestra como los valores de la similitud lingüística correspondientes a cada localidad se ajustan a una distribución normal ideal, se puede comprobar con otro tipo de elementos, por ejemplo el Test de Shapiro-Wilk, el cual plantea la hipótesis nula que la muestra proviene de una distribución normal, tomando generalmente como valor de significancia 0.05 ($\alpha = 0.05$), con la hipótesis alternativa de que la distribución proviene de una distribución no normal, o la prueba K-cuadrado prueba de D'Agostino, prueba que a partir de la curtosis y la asimetría permite saber si la prueba esta normalmente distribuida, la notación respectiva es de la siguiente manera:

Shapiro - Wilk y K2 de D'Agostino

$$\begin{aligned} H_0 &= La\ distribución\ es\ normal \\ H_1 &= La\ distribución\ no\ es\ normal \end{aligned} \tag{3}$$

Los resultados de las pruebas se muestran a continuación:

Test	**Valor Estadístico**	**p-valor**	**Significancia α**
Shapiro - Wilk	0.994	0.442	0.05
Prueba K2 de D'Agostino	2.807	0.246	0.05

Cuadro 23: Resultados de pruebas de Normalidad

Dado que el p-valor es mayor al nivel de significancia α, se acepta la hipótesis nula que propone la muestra como una fuente de datos proviene con distribución normal, es decir, tanto la prueba Shapiro - Wilk como K2 de D'Agostino nos dan resultados similares: una variable respuesta de tipo normal, variable que servirá para posteriores análisis. Finalmente la visualización final de la distancia lingüistica dividida en categorías para cada Localidad, es de la siguiente manera:

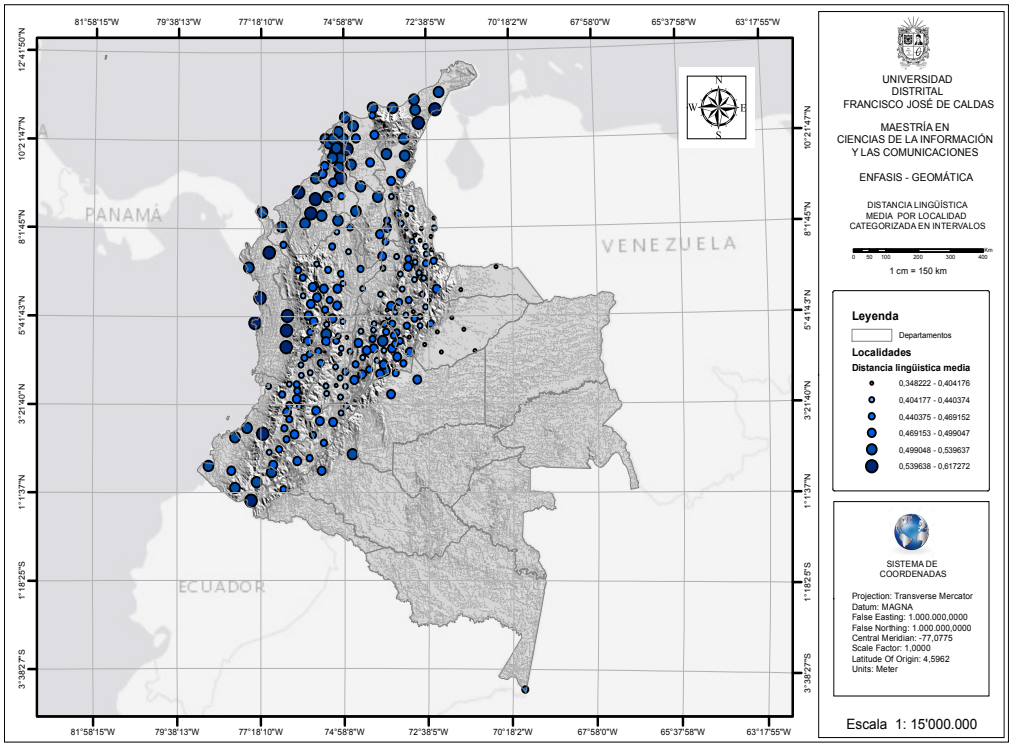


Figura 27: Distancia lingüística por localidad

7.2.11. Distancia geográfica

La distancia entre dos puntos es igual a la longitud de recta representada numéricamente, para este trabajo se utilizará la distancia geodésica, distancia medida entre dos puntos sobre una superficie de revolución denominada elipsoide.

Existen numerosas técnicas de medición de distancias sobre la superficie del elipsoide, sin embargo en este proyecto se utilizará la formula de Vincenty, formula que ha generado resultados precisos y con ventajas notables en el tiempo de cálculo (Esenbuğa y Çolak, 2016).

Con uso de la librería Geopy de Python se implementó la formula de Vincenty, referenciando en el procesamiento los valores asociados a las coordenadas (x,y) de cada localidad y asumiendo como elipsoide de referencia el WGS84, una vez hecho esto se obtuvo una matriz cuadrada, que permitió relacionar cada valor de distancia entre las distintas localidades, de la siguiente manera:

	Ch01	**Ch10**	**Ch30**	**Ch03**
Ch01	0	119,8071357	251,7752858	167,3827172
Ch10	119,8071357	0	136,7562652	77,26507002
Ch30	251,7752858	136,7562652	0	96,29309692
Ch03	167,3827172	77,26507002	96,29309692	0

Cuadro 24: Matriz distancia geográfica

La distancia geográfica está medida en kilómetros y cuenta con varias cifras decimales para mejorar la precisión del cálculo, su verificación se hizo por medio del software en línea Gabmap, este proporciona la distancia geográfica utilizando la formula de Vincenty, para terminar, los datos de distancia fueron calculados para cada localidad, a partir de la media aritmética, de tal manera que todos los datos fueron normalizados para iniciar el desarrollo del modelo.

Por último, los valores se categorizaron y visualizaron, mostrando que la distancia geográfica promedio para cada localidad tiene una distribución espacial acorde a la densidad de puntos por área, es decir, donde hay mayor cantidad de puntos, su distancia media es menor que aquellas áreas donde hay una menor densidad.

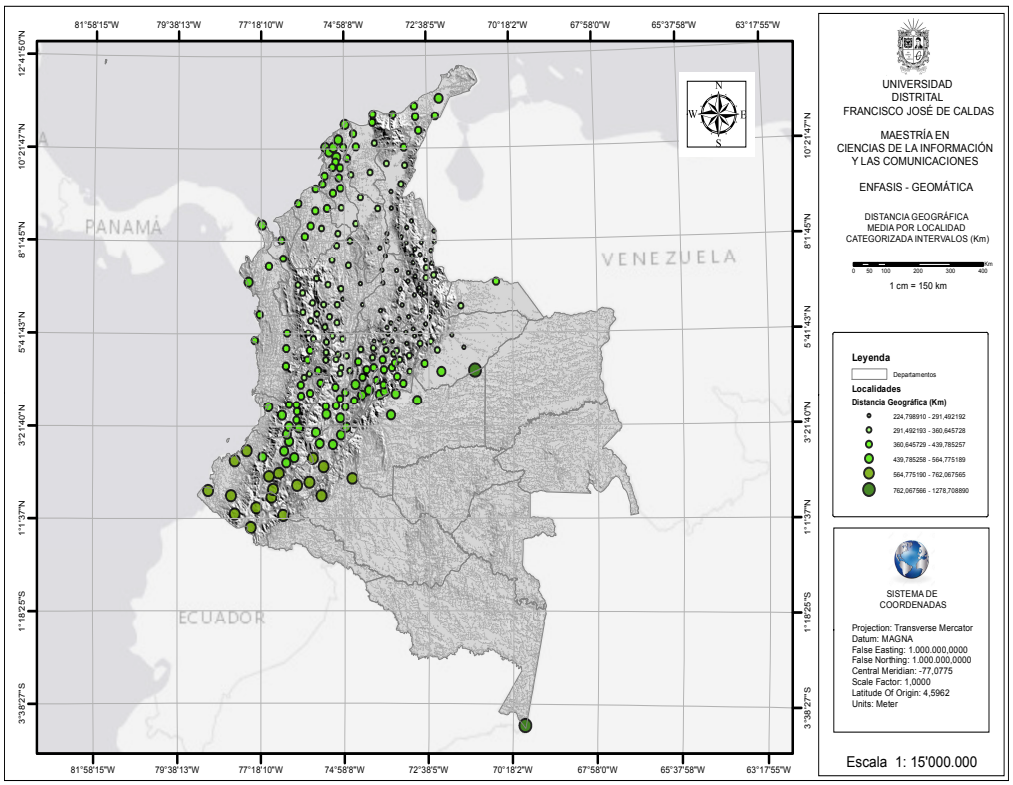


Figura 28: Distancia Geográfica por Localidad

7.3. Diseño del Modelo

7.3.1. Pesos Espaciales

Los pesos espaciales son un elemento clave para el modelamiento de datos geográficos, permiten probar y estimar modelos de autocorrelación espacial basados en la representación espacial de los elementos de un conjunto de datos; se definen como una estructura basada en vecindades, donde formalmente, los pesos expresan la estructura contigua entre las observaciones (en este caso el número de localidades) como una matriz W de nxn dimensiones en la que los elementos W_{ij} son los pesos espaciales (Anselin y Smirnov, 2006).

Según Anselin y Rey (2014) los pesos espaciales W_{ij} no son cero cuando i y j son vecinos, y cero en caso contrario. Convencionalmente se excluye la

relación consigo mismo, de tal manera que los elementos diagonales de W sean cero, $W_{ii} = 0$. La forma más simple y general de la matriz de pesos espaciales expresa la existencia de una relación vecina de forma binaria, con pesos 0 y 1. Por notación cada unidad espacial es representada como una fila i y los vecinos potenciales por las columnas j, con $j \neq i$. la existencia de la relación fila i y la columna correspondiente j sería representada como $W_{ij} = 1$.

$$\mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & ...w_{1n} \\ w_{21} & w_{22} & ...w_{2n} \\ ... & ... & ... \\ w_{n1} & w_{n2} & ...w_{nn} \end{bmatrix} \tag{4}$$

Hay muchos criterios sobre los cuales la construcción de la matriz de pesos puede basarse, sin embargo los dos enfoques operativos más comunes son la relación de vecindad basada en la noción de contigüidad y a partir de la medida de distancia.

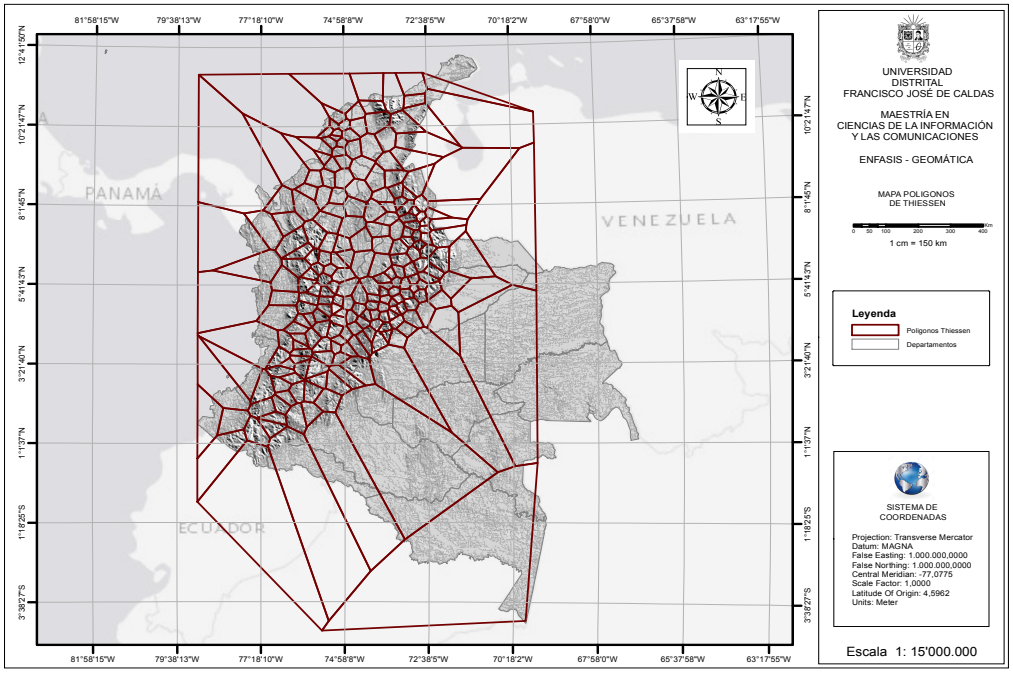


Figura 29: Polígonos de thiessen de las localidades

Según Anselin y Rey (2014) la contigüidad intrínseca es más apropiada

para los datos geográficos expresados como polígonos, mientras que la distancia es adecuada para los datos con geometría tipo punto, dentro del modelo a desarrollar sin embargo se utilizará una teselación basada en los polígonos de Thiessen (Figura 29), esta permitirá la determinación de relaciones de contigüidad entre los polígonos de la teselación.

La contigüidad significa que dos objetos espaciales comparten un lindero en común de longitud distinta de cero, al generar los valores de la matriz de pesos espaciales se pueden distinguir dos criterios de contigüidad, de torre y de reina (Figura 30), criterios que hacen referencia a los movimientos permitidos a las piezas así llamadas en un tablero de ajedrez.

En primer lugar el criterio por contigüidad de torre define vecinos por la existencia de un arco (lindero o linea, revisar teoría de grafos) común entre dos objetos espaciales, el criterio de la reina sin embargo no solo define los objetos espaciales contiguos por el arco compartido sino por el vértice, lo cual hace que sea utilizada de forma regular (Anselin y Rey, 2014).

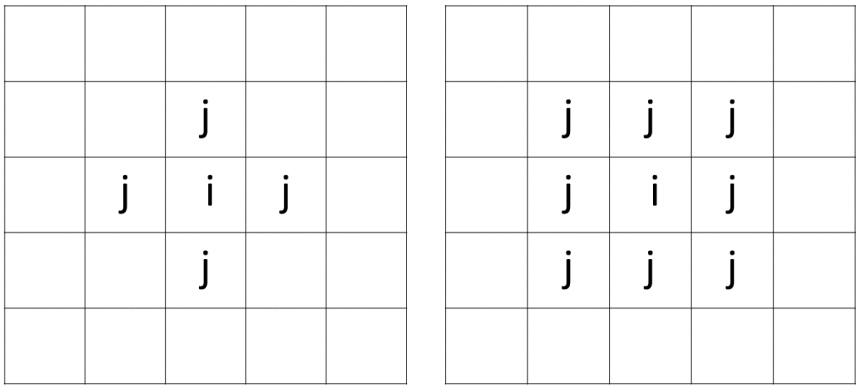


Figura 30: Contigüidad de Torre (Iz) vs Reina (Der)

En esta investigación se utilizará el criterio de pesos de contigüidad basado en los vértices y arcos compartidos, es decir, teniendo presente la hipótesis

que presenta la variación lingüística como el resultado del contacto social y la distancia geográfica (Nerbonne, 2010), por lo tanto lugares colindantes tendrán un mayor peso sobre las observaciones debido a su cercanía con el lugar encuestado (Anselin y Rey, 2014), en este caso para cada localidad.

La implementación de pesos espaciales para las localidades se realizó por medio del software GeoDa (sección 6.4.2), la creación de estos pesos se hace con el módulo de Gestor de Pesos que permite no solo su creación sino su caracterización por medio de su histograma y mapa de conectividad.

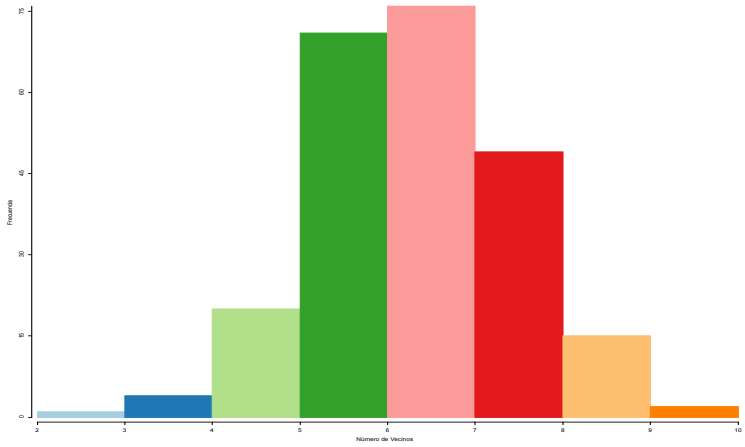


Figura 31: Histograma de conectividad - pesos espaciales

La figura 31) muestra el número de observaciones para cada valor de cardinalidad, es decir, el número de observaciones dado el número de vecinos, es evidente como el patrón general es simétrico, con centro en 7, que representa el 31,9 % del total, el valor mínimo de observaciones vecinas en un punto son dos, y un máximo de nueve en las 238 localidades seleccionadas.

Finalmente, estos datos se exportaron al dataset de localidades, y se pueden visualizar de la siguiente manera:

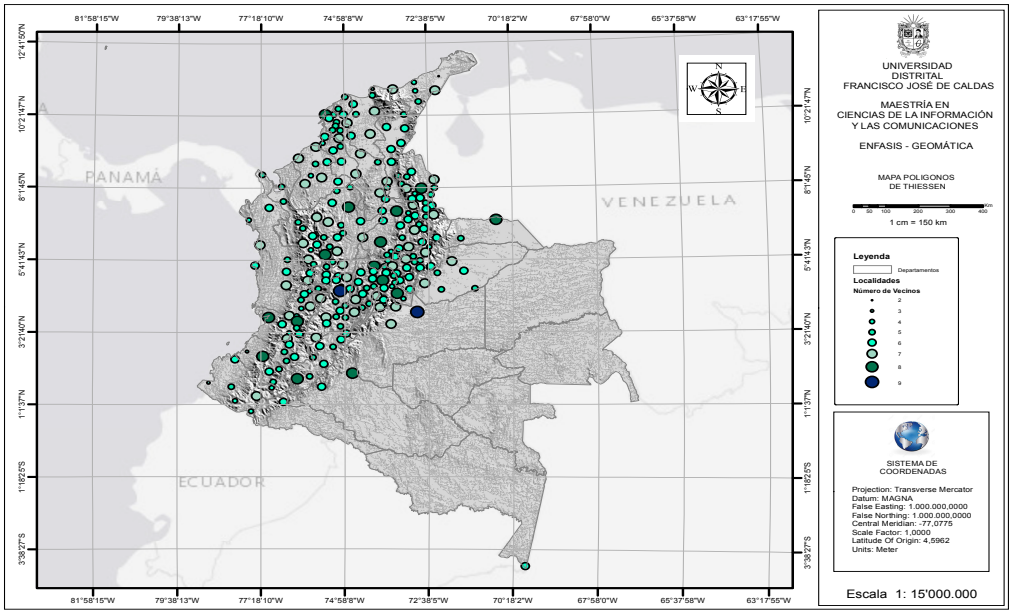


Figura 32: Valores de contigüidad visualización

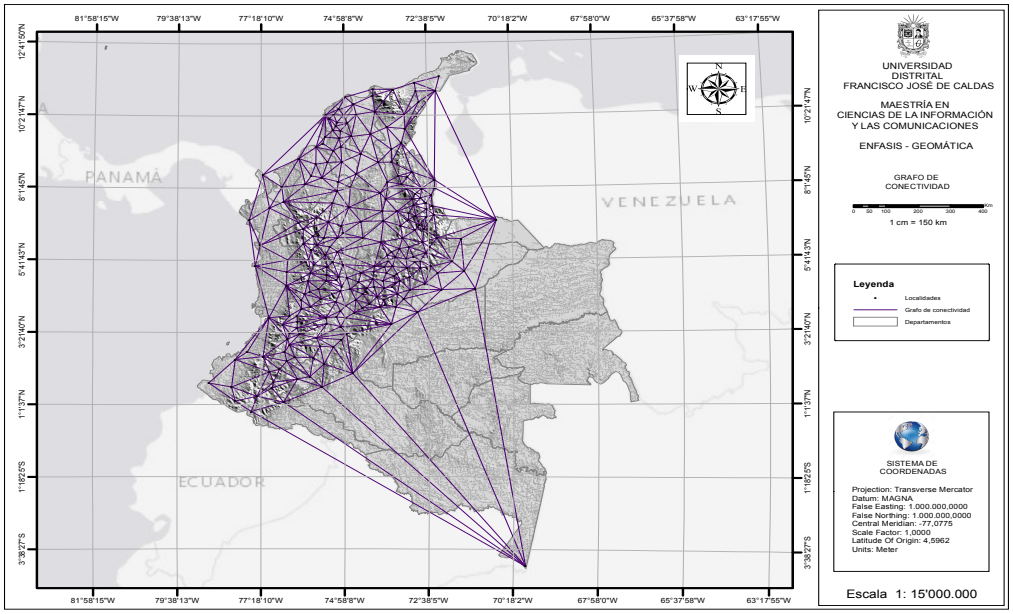


Figura 33: Grafo de conectividad

7.3.2. Correlación Espacial

Una vez definidos los pesos espaciales para cada localidad se buscará determinar la concentración o dispersión de los valores de una variable dentro del territorio, para ello inicialmente se utilizara la autocorrelación espacial, reflejando el grado en que objetos o actividades en una unidad geográfica son similares a otras actividades u objetos en unidades geográficas cercanas (Goodchild, 1987).

Cuando se habla de dependencia espacial, el valor de la variable en un espacio geográfico es parcialmente función del valor de la misma variable en espacios vecinos (Flint et al., 2000), en este trabajo se verá como las variables pueden estar o no autocorrelacionadas geográficamente, con el uso de técnicas estadísticas, específicamente con el uso del índice I de Moran una medida para el diagnostico de correlación espacial (Bohorquez y Ceballos, 2008).

El índice I de Moran es una de las herramientas más antiguas para la medición de autocorrelación espacial (Bohorquez y Ceballos, 2008), este método es similar a la correlación de Pearson, donde los valores varían en el intervalo $(1, -1)$, si el valor $I = 1$, existe una relación positiva, los valores están concentrados en un espacio geográfico y hay correspondencia, si el valor $I = -1$ existe una autocorrelación negativa y los valores están perfectamente dispersos, si el valor $I = 0$, los valores son espacialmente aleatorios (Vilalta, 2005); para el cálculo del índice de Moran se utiliza la asociación de los valores determinada por una matriz de distancias o de contigüidad (matriz de pesos) que predefine los valores vecinos , su formula matemática es de la siguiente manera:

$$I = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \cdot \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i=1}^{i=n}(x_i - \overline{x})^2} \tag{5}$$

Bohorquez y Ceballos (2008)

Donde (ecuación 5), n es el número de localidades o unidades geográficas, W_{ij} es la matriz que define las distancias, o la matriz de contigüidad (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 esta dada por los valores de Z, con el supuesto de una distribución normal (Goodchild, 1987).

La implementación de este índice se hizo por medio del software Geoda, así como su visualización, las variables utilizadas son aquellas mencionadas en items anteriores.

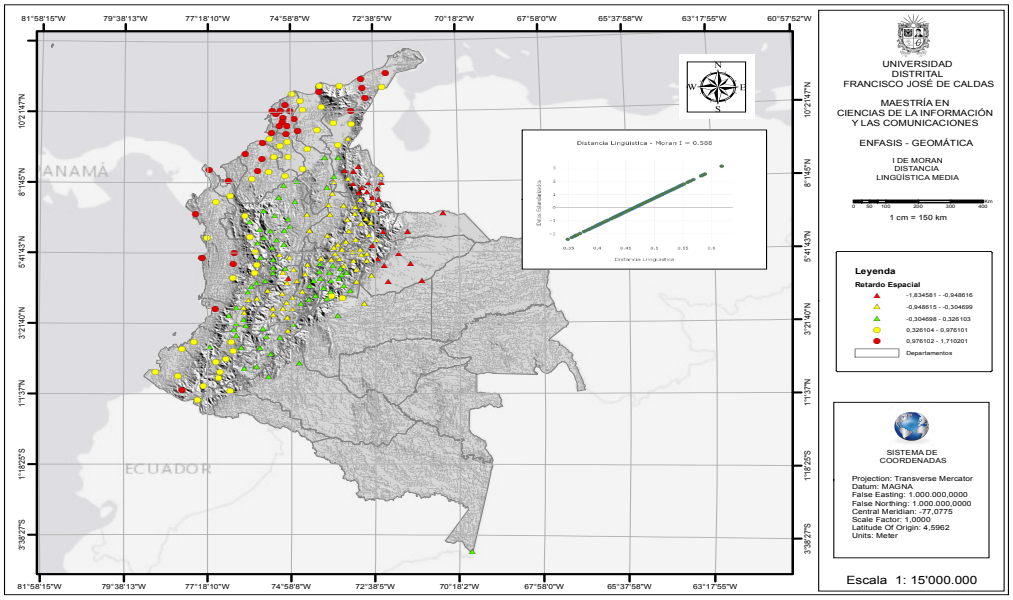


Figura 34: Índice de Moran para la Distancia Lingüistica

La Figura 34 muestra el diagrama de dispersión para los valores estandarizados del retardo espacial obtenido por el índice I de Moran para la variable distancia lingüística, valor que para este caso muestra una autocorrelación positiva alta de $I = 0{,}588$, adicionalmente los valores de retardo espacial se representan espacialmente divididos en cuantiles, el retardo espacial es el promedio de los valores que adopta una variable en el conjunto de observaciones vecinas (por eso la matriz de pesos espaciales) (Coro, 2005), se puede ver como estos valores se encuentran relacionados en tonalidad siendo más oscuros cuanto más es el valor del retardo espacial, especialmente en el área andina departamento de Cundinamarca en tonos claros y más oscuros en la región Pacífica y Caribe, es de resaltar que el diagrama de dispersión el valor

de la pendiente de la recta que forman los valores estandarizados es el denominado estadístico I de Moran de autocorrelación espacial de tipo global, la correlación positiva como se puede observar está relacionada en los cuadrantes uno y tres.

La autocorrelación por tipo de variable es de la siguiente manera:

Actividad Económica	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Comercio	**0.138**	**0.001**	**-0.1284**	**0.001**
Ganadería	0.030	0.200	0.031	0.1270
Industria	-0.035	0.236	0.034	0.1090
Minería	0.057	0.054	0.003	0.4460
Agricultura	0.136	0.004	0.078	0.004
Pesca	**0.235**	**0.001**	**0.123**	**0.001**

Cuadro 25: I de Moran para actividades económicas global y bivariante

El cuadro 25 muestra los valores para la medición de la autocorrelación utilizando el Índice de Moran, el mayor valor es para la actividad económica Pesca, con una autocorrelación positiva de **0.235** y con un p-valor menor al valor $\alpha = 0{,}05$ de significancia, validando la hipótesis inicial que indica como los valores están concentrados en un espacio geográfico con una dispersión baja y por lo tanto rechazando la hipótesis nula de no autocorrelación; para su cálculo se hicieron en total 999 permutaciones (al igual que para los otros valores), de tal manera que los valores tuviesen una mayor confiabilidad técnica, adicionalmente se hizo la relación entre cada una de las actividades económicas y la diferencia lingüistica con el indice **I de moran bivariante, el cual muestra la relación existente entre la distancia lingüistica media (DLM) de cada localidad y los respectiva datos de actividades económicas**.

Aptitud Agroclimática	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Ligeramente húmedo	0.105	0.006	-0.113	0.001
Moderadamente seco	**0.217**	**0.001**	**-0.121**	**0.001**
Moderadamente húmedo	0.260	0.001	0.018	0.245
Adecuada disponibilidad de agua	0.263	0.001	-0.015	0.2610
Extremadamente húmedo	**0.483**	**0.001**	**0.1895**	**0.001**
Extremadamente seco	**0.659**	**0.001**	**0.063**	**0.015**

Cuadro 26: I de Moran para aptitud agroclimática global y bivariante

En el cuadro 26 se observan los valores respectivos del Índice I de Mo-

ran para la aptitud agroclimática, estos indican como la autocorrelación es mucho más alta con respecto a las actividades económicas, los valores altos corresponde a las zonas ***extremadamente secas*** cuyo nivel de significancia permite rechazar la hipótesis nula de no correlación, adicionalmente se muestran los valores del I de Moran Bivariante, que relacionan la distancia lingüística media y la aptitud agroclimática, siendo bajos en su mayoría, a diferencia de las categoría mencionada ***Extremadamente seco*** y ***Extremadamente Húmedo***.

Valores de Altura	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Altura	**0.655**	**0.001**	**-0.310**	**0.001**

Cuadro 27: I de Moran global y bivariante para la altura

Para la altura media sobre el nivel medio del mar (Cuadro 27), los valores están autocorrelacionados de una forma alta, con un nivel de significancia similar, su relación con los valores de distancia lingüistica es negativa, es decir, en cada localidad los valores de la variable distancia lingüistica tienen un disminución del **0.31 (31 %)** con respecto al promedio de los valores vecinos de la variable altura.

Precipitación Rango	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Entre 1000 y 1500	0.307	-0.002	0.078	0.006
Entre 1500 y 2000	0.083	-0.004	-0.040	0.066
Entre 2000 y 2500	0.118	-0.013	-0.091	0.002
Entre 2500 y 3000	**0.165**	**-0.068**	**-0.095**	**0.001**
Entre 3000 y 4000	0.063	0.103	0.037	0.093
Entre 4000 y 5000	0.060	0.116	0.067	0.007
Entre 500 y 1000	0.492	-0.128	-0.038	0.078
Entre 5000 y 7000	**0.062**	**0.108**	**0.137**	**0.001**
Entre 7000 y 9000	**0.111**	**0.045**	**0.134**	**0.001**

Cuadro 28: I de Moran global y bivariante para la precipitación

El cuadro 28 muestra como los valores de precipitación tienen una autocorrelación alta para cada rango, especialmente para el que va ***entre 500 y 1000*** con un I de Moran de **0.492**, sin embargo con un p-valor más alto que el nivel de significancia; el I de moran Bivariante por lo contrario es más alto para los rangos que van ***entre 5000 y 7000***, y ***7000 y 9000***, con un nivel de significancia menor que alfa, lo que permite corroborar la hipótesis de correlación espacial entre la precipitación y la diferencia lingüística.

Temperatura Rango	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Entre 12 y 16 grados	**0.244**	**0.001**	**-0.143**	**0.001**
Entre 16 y 20 grados	0.263	0.001	-0.107	0.002
Entre 20 y 22 grados	0.090	0.018	-0.058	0.025
Entre 22 y 24 grados	0.303	0.001	-0.029	0.143
Entre 24 y 26 grados	0.192	0.001	0.035	0.094
Entre 26 y 28 grados	**0.510**	**0.001**	**0.242**	**0.001**
Entre 8 y 12 grados	0.139	0.004	-0.113	0.001
Mayor a 28 grados	0.435	0.001	0.063	0.006

Cuadro 29: I de Moran global y bivariante para la temperatura

Para la variable temperatura el Cuadro 29 muestra como el I de Moran de autocorrelación tiene un mayor valor en temperaturas altas, especialmente en el intervalo que va ***entre 26 y 28 grados***, adicionalmente indica como la relación entre la diferencia lingüística y temperatura tiene una mayor correlación para el mismo intervalo, es importante resaltar que el nivel de significancia fue menor a alfa ($\alpha = 0{,}5$) rechazando la hipótesis nula de no correlación.

IDH	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Adecuado	**0.383**	**0.001**	**-0.080**	**0.002**
Húmedo	**0.067**	**0.084**	**0.0808**	**0.004**
Muy húmedo	0.612	0.001	0.196	0.001
Muy seco	**0.645**	**0.001**	**0.375**	**0.001**
Seco	0.453	0.001	-0.036	0.093
Semihúmedo	0.262	0.001	0.0033	0.454
Semiseco	0.444	0.001	-0.261	0.001

Cuadro 30: I de Moran global y bivariante para el IDH

En el caso del índice de disponibilidad hídrica (Cuadro 30) las localidades con mayor autocorrelación espacial son aquellas que se encuentran en ***zonas muy secas***, valores que en relación con los datos de diferencia lingüística indican una correlación espacial media, con un nivel de significancia alto, es importante ver como estas zonas son las mismas en la que la temperatura está ***entre 26 y 28 grados***, mostrando no solo la relación entre las variables señaladas atrás, sino esencialmente con la diferencia lingüística media del espacio geográfico habitado por las mismas.

Vías de Acceso	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Via Aérea	-0.004	0.450	0.010	0.328
Via Férrea	-0.006	0.064	-0.031	0.111
Via Camino de herradura	**0.065**	**0.090**	**0.074**	**0.006**
Via Carretera y Férrea	0.084	0.024	0.023	0.1820
Via Carretera	0.147	0.003	-0.041	0.070
Via Fluvial	**0.179**	**0.002**	**0.151**	**0.001**
Caminos carreteables	**0.420**	**0.001**	**-0.206**	**0.001**

Cuadro 31: I de Moran global y bivariante para las vías de acceso

Por último se realizó el análisis para las vías de acceso y la distancia geográfica, en el caso de las vías de acceso la mayor autocorrelación espacial es representada por los ***caminos carreteables*** con un valor de ***0.420***, así mismo esta variable tiene la **mayor correlación bivariante inversamente proporcional a la distancia lingüística**, de la cual se puede inferir que **a mayor cantidad de caminos carreteables menor diferencia lingüística o que puede existir dispersión de los valores de distancia lingüística con respecto a los caminos carreteables**, es importante resaltar existe un nivel de significancia alto para esta categoría, por lo que se rechaza la hipótesis nula de no correlación. En el caso de la distancia geográfica (Cuadro 32) la autocorrelación es alta, y la relación con la distancia geográfica en unidades vecinas es positiva, lo que indica **que a mayor distancia geográfica mayor diferencia lingüística**, relación ya probada por otros autores (Goebl (1987), Séguy (1973)).

Distancia Geográfica	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Distancia Geográfica	**0.798**	**0.001**	**0.269**	**0.001**

Cuadro 32: I de Moran global y bivariante para la distancia geográfica

Con el resultado de la relación de las variables obtenido, se puede iniciar el proceso de construcción del modelo, este proceso se mostrará en la siguiente sección.

7.4. Construcción del Modelo

Según Anselin (1988) las causas de la autocorrelación espacial se pueden identificar a partir de dos hechos, el primero tiene que ver con la existencia de errores de medición en las observaciones de unidades geográficas vecinas y la segunda, aquella que comprende la existencia de varios fenómenos de interacción espacial, en esta investigación se desarrollará esta interacción, teniendo como premisa que todo aquello que ocurre en un territorio afecta a

otros territorios, es decir lo que pasa en una localidad está determinado por aquello que pasa en otras localidades (a mayor distancia entre una localidad y otra, existe mayor diferencia lingüística (Goebl, 2006)), existe entonces una interdependecia espacial en unidades geográficas, que permite efectos de difusión, procesos de dispersion, interacciones, externalidades, jerarquias, entre otros (Pérez, 2006).

La existencia de este tipo de fenómenos relacionados con la dependencia espacial (en ocasiones causa de la misma) permite considerar herramientas para estimarlos y agregarlos dentro de modelos geoestadísticos y/o modelos espaciales econométricos; tal es el caso de la matriz de pesos espaciales y/o matriz de retardos espaciales, instrumentos que permiten fusionar hechos como la interdependencia y multidireccionalidad en datos geográficos (Pérez, 2006), como se evidenció en enunciados anteriores.

Analizada la dependencia espacial para las variables geográficas que hacen parte del ALEC, el siguiente paso consiste en integrarlas y analizarlas dentro de un modelo espacial; existen distintos y variados modelos (generalmente de regresión) que permiten incorporar la dependencia espacial de manera formal, a continuación se presenta una breve introducción de la estructura básica del modelo a utilizar, junto con las consecuencias e implicaciones de su uso.

Modelo autorregresivo de regresión espacial, SAR(1)

Anselin (2001) define la econometría espacial como una parte de la econometría que se ocupa del tratamiento de la interacción espacial y la estructura de datos espaciales, para su análisis y observación dentro de modelos geográficos, hace la inclusión de efectos de dependencia espacial y heterogeneidad y/o heteroscedasticidad dentro del análisis de regresión; es importante destacar que los modelos dentro de esta rama del conocimiento también denominados modelos espaciales dinámicos, son aquellos modelos de regresión espacial que consideran de manera explicita la existencia del efecto espacial de dependencia o autocorrelación.

Entre los modelos más destacados se encuentran el modelo de regresión

espacial autorregresivo, este es la especificación básica de un modelo espacial, permitiendo identificar sistemáticamente niveles de dependencia para una variable a partir de los distintos valores que toma en distintas unidades geográficas, su expresión es de la siguiente manera:

$$y = \rho W_y + \mu \tag{6}$$

$$\mu \approx N(0, \sigma^2 I) \tag{7}$$

Yrigoyen (2003, pg. 93)

Donde y es un vector $(N, 1)$ de observaciones de la variable explicada, W la matriz de pesos espaciales de la misma variable siendo W_y el retardo espacial, ρ el coeficiente autorregresivo espacial (un valor escalar), **valor que toma la intensidad de las interdependencias entre las observaciones** y finalmente μ la perturbación aleatoria.

El modelo SAR(1) describe cómo los valores de una variable, (en este caso y) en un determinado punto geográfico (i) están condicionados por los valores que adopta esta variable en otros puntos geográficos (j), dando lugar a fenómenos de agrupamiento; en términos estadísticos es definido como la $Cov(y_i, y_j) \neq 0$, una de las características fundamentales es que este modelo de autocorrelación tiene un carácter secuencial, no simultáneo, sin cabida en el ámbito temporal como se observa en la Figura 35.

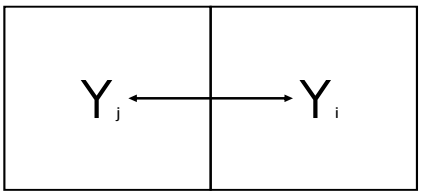


Figura 35: Ilustración Modelo SAR(1)
(Yrigoyen, 2003)

El modelo SAR(1) muestra la ausencia de aleatoriedad espacial en la variable (y) sin considerar variables adicionales que expliquen su distribución,

algo que no ocurre con el modelo mixto regresivo espacial autorregresivo (Modelo Retardo Espacial y Modelo de Error espacial respectivamente), como se muestra a continuación.

Modelo básico de regresión

Cuando existe la posibilidad de existencia de **NO dependencia espacial** en una variable, esta se puede recoger en un grupo de regresores a partir de la siguiente estructura:

$$y = X\beta + \mu \tag{8}$$

$$\mu \approx N(0, \sigma^2 I) \tag{9}$$

Yrigoyen (2003, pg. 94)

Donde X es una matriz de tamaño (K, N) de K variables y N observaciones, finalmente el vector B representa los parámetros de estas variables, con un tamaño $(K, 1)$; como se puede observar la especificación de un modelo básico solo sería correcto en terminos espaciales cuando el efecto espacial (en este caso la diferencia lingüistica) esté totalmente explicado por valores de una o más variables en dicho lugar (i).

Generalmente, el uso de este modelo a partir de la estimación por el método de mínimos cuadrados produce un efecto de dependencia espacial significativo debido a la mala estimación del mismo, ya que no logra explicar la estructura de la variable respuesta, en este tipo de casos hay problemas de correlación espacial, por lo tanto la forma más veraz de abordar este tipo de cuestiones es con el uso de modelos de dependencia espacial, como el modelo residual (modelos de error espacial) o el modelo de dependencia espacial sustantiva (modelos de retardo espacial) (Yrigoyen, 2003).

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

El modelo de dependencia espacial en la perturbación aleatoria según Yrigoyen (2003) es una estructura que permite definir la existencia de factores o variables no considerados en el modelo y enunciarlos en términos del error,

de esta manera la relación de dependencia entre la variable respuesta (y) se explica no solo por las variables independientes sino por aquellas que se encuentran ausentes (dependencia espacial residual), su representación en forma general es de la siguiente manera:

$$y = X\beta + \mu \tag{10}$$

$$\mu = \lambda W_{\mu} + \epsilon \tag{11}$$

$$\epsilon \approx N(0, \sigma^2 I) \tag{12}$$

Yrigoyen (2003, pg. 95)

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

Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial

El modelo del retardo espacial a diferencia de los casos mencionados 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 la variable en las localizaciones vecinas, de la siguiente forma:

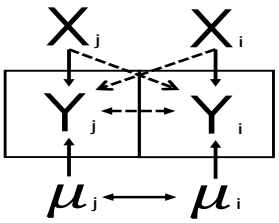


Figura 36: Ilustración Modelo Retardo Espacial (Yrigoyen, 2003)

El modelo de retardo espacial es adecuado cuando se está poniendo en manifiesto un proceso de difusión espacial de la diferencia lingüística, de

tal forma que los valores de dicha variable (y) en una localidad (i) estaría incrementando la probabilidad de ocurrencia de valores en lugares vecinos, es decir cuanto más aproximas las localidades geográficamente menor diferencia lingüística.

$$y = \rho W_y + X\beta + \mu \tag{13}$$

$$\mu \approx N(0, \sigma^2 I) \tag{14}$$

Yrigoyen (2003, pg. 96)

Como en el modelo del error espacial, la expresión puede adoptar una forma que la variable respuesta (y) en un punto (i) está afectado por el valor no solo de las variables independientes en ese punto (i) sino también por las demás localizaciones utilizando $(1 - \lambda W)^{-1}$.

$$y = (1 - \rho W_y)^{-1} X\beta + (1 - \rho W)^{-1} \mu \tag{15}$$

$$\mu \approx N(0, \sigma^2 I) \tag{16}$$

Yrigoyen (2003, pg. 96)

Modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR)

Una vez definidos los modelos principales de retardo espacial y del error espacial es posible precisar modelos adicionales que incluyan combinaciones de los anteriores, el modelo general es el denominado modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) cuya conceptualización es de la siguiente manera :

$$y = \rho W_1 y + X\beta + \mu \tag{17}$$

$$\mu = \lambda W_2 \mu + \epsilon \tag{18}$$

$$\epsilon = N(0, \sigma^2 I) \tag{19}$$

Herrera Gomez et al. (2012, pg. 14), Anselin y Florax (1995)

El modelo surge de la combinación de los modelos tanto de retardo espacial como de error espacial; al tener en cuenta dos matrices espaciales como W_1 y W_2 estas pueden ser iguales, dependiendo de la estructura espacial de las unidades geográficas.

Estimación del modelo de regresión por mínimos cuadrados ordinarios (MCO):

Las variables geográficas, variables respuesta de la diferencia lingüística (según la hipótesis de este proyecto) se incluirán dentro de un modelo de regresión lineal, esta definición tiene como propósito detectar la relación de cada una de ellas con la variable explicada, adicionalmente verificar la relación NO espacial que existe entre ellas, teniendo como criterio fundamental la correspondencia entre variables y diferencia lingüística; el modelo:

Variable	Estimate	Std.Error	t value	Pr(>—t—)	Significancia
(Intercept)	0.2780479	0.0898982	3.093	0.00227	**
Distancia Geográfica					
I(log(DistanciaG))	0.0507104	0.0113067	4.485	1.24e-05	***
Altura media sobre el nivel medio del mar					
I(log(ALTURA))	-0.0084437	0.0060186	-1.403	0.16223	
Actividades Económicas					
Comercio	-0.0234454	0.0182801	-1.283	0.20117	
Ganadería	0.0067890	0.0104548	0.649	0.51686	
Industria	0.0116415	0.0186368	0.625	0.53293	
Minería	0.0133656	0.0192970	0.693	0.48937	
Pesca	0.0070947	0.0141850	0.500	0.61753	
Aptitud Agroclimática					
Extremadamente húmedo	-0.0042131	0.0281010	-0.150	0.88098	
Extremadamente seco	0.0016166	0.0129710	0.125	0.90094	
Ligeramente húmedo	-0.0098911	0.0106781	-0.926	0.35543	
Moderadamente húmedo	0.0273977	0.0192674	1.422	0.15663	
Moderadamente seco	-0.0029856	0.0112700	-0.265	0.79135	
Precipitación					
Entre 1500 y 2000	-0.0041059	0.0076222	-0.539	0.59073	
Entre 2000 y 2500	-0.0044775	0.0103434	-0.433	0.66558	
Entre 2500 y 3000	-0.0104097	0.0099207	-1.049	0.29534	
Entre 3000 y 4000	0.0177010	0.0158478	1.117	0.26540	
Entre 4000 y 5000	0.0151328	0.0167037	0.906	0.36608	
Entre 500 y 1000	0.0001693	0.0107708	0.016	0.98748	
Entre 5000 y 7000	0.0163368	0.0301831	0.541	0.58895	
Entre 7000 y 9000	0.0442917	0.0325188	1.362	0.17476	

Amenaza Sísmica					
Extremadamente Alto	-0.0204395	0.0219393	-0.932	0.35267	
Extremadamente Bajo	-0.1096399	0.0513014	-2.137	0.03383	*
Moderadamente Alto	0.0158319	0.0079435	1.993	0.04765	*
Muy Alto	0.0053712	0.0076531	0.702	0.48362	
Temperatura					
Entre 16 y 20 grados	-0.0148255	0.0096233	-1.541	0.12504	
Entre 20 y 22 grados	-0.0151730	0.0155148	-0.978	0.32930	
Entre 22 y 24 grados	-0.0240873	0.0144654	-1.665	0.09749	.
Entre 24 y 26 grados	-0.0366812	0.0146446	-2.505	0.01307	*
Entre 26 y 28 grados	-0.0242457	0.0163935	-1.479	0.14076	
Entre 8 y 12 grados	-0.0129337	0.0133125	-0.972	0.33248	
Mayor a 28 grados	-0.0302542	0.0185255	-1.633	0.10406	
IDH					
Húmedo	0.0147276	0.0358030	0.411	0.68127	
Muy húmedo	0.0211108	0.0335360	0.629	0.52976	
Muy seco	0.0423385	0.0150355	2.816	0.00536	**
Seco	0.0097586	0.0119370	0.818	0.41464	
Semihúmedo	-0.0192633	0.0191843	-1.004	0.31657	
Semiseco	-0.0077817	0.0095763	-0.813	0.41744	
Vias					
Camino de herradura	-0.0222642	0.0263463	-0.845	0.39911	
Caminos carreteables	-0.1450533	0.0247028	-5.872	1.83e-08	***
Carretera	-0.0445338	0.0163446	-2.725	0.00702	**
Carretera y Férrea	-0.0439274	0.0208021	-2.112	0.03599	*
Férrea	-0.0819993	0.0432531	-1.896	0.05947	.
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**					
Residual standard error: 0.03683 on 195 degrees of freedom **Multiple R-squared: 0.525,** **Adjusted R-squared: 0.4227** **F-statistic: 5.132 on 42 and 195 DF, p-value: 1.726e-15**					

Cuadro 33: Modelo de regresión no espacial

La implementación del modelo lineal se realizó en el software estadístico R, los valores del Cuadro 33 resaltan la influencia que tienen algunas variables con respecto a la distancia lingüística media, en este caso muestra que la mayor significancia del modelo es el logaritmo de la distancia geográfica, hipótesis que ha sido ya verificada y mencionada por otros autores (Nerbonne et al. (2019), Goebl (2006), Dubert y Sousa (2016), Séguy (1973)); se puede observar como el logaritmo de la altura tiene menor significancia, siendo el valor del parámetro β igual a **-0.0084437**, las variables clasifica-

das como las actividades económicas principales, la aptitud agroclimática, la precipitación tienen una relevancia casi nula, sin embargo valores como los de Amenaza Sísmica, en las localidades donde hay un riesgo moderadamente alto y extremadamente bajo se obtuvo un parámetro β con un valor de **0.0158319** y **-0.1096399** respectivamente, adicionalmente las localidades con rango de temperatura 24 y 26 grados (Everett et al., 2015), tienen una significancia más alta siendo el valor del parámetro β igual a **-0.0366812**, la aptitud agroclimática para localidades ubicadas en zonas muy secas representa uno de los valores más significativos, después de la distancia geográfica y las vías de acceso, con un parámetro β para el modelo de regresión equivalente a **0.0423385**, finalmente las vías de acceso representan uno de los valores más llamativos de este modelo no espacial, ya que son significativos aquellos lugares en los que existe caminos carreteables, carreteras y carreteras junto con vías férreas, siendo los parámetros β iguales a **-0.1450533**, **-0.0445338**, **-0.0439274** respectivamente, en general el modelo tiene un error residual estándar de **0.03683** y un coeficiente de determinación (R-squared) de **0.525**, ajustado a **0.4227** es decir el modelo explica el **42 %** del valor de variabilidad de la diferencia lingüística, el estadístico F finalmente en forma general fue de **5.132** con un p-valor menor a **1.726e-15** permitiendo rechazar la hipótesis nula de un mejor ajuste del modelo sólo con el intercepto y concluyendo que el modelo mejora el ajuste de datos de información lingüística con información geográfica y algunos metadatos de las localidades del ALEC.

Contraste de dependencia espacial: Hecho esto se puede contrastar la estimación de este modelo con diferentes test para detectar y analizar algún tipo de dependencia espacial, como ya se mencionó la hipótesis nula indica que el tipo de dependencia es igual a cero, es decir, no existe algún tipo de correlación espacial y la hipótesis alternativa en contraste muestra que existe algún tipo de dependencia espacial, para ello Anselin (2005) plantea una prueba de multiplicadores de Lagrange, este test incluye cinco alternativas para hallar el mejor modelo para la información de entrada, los dos primeros test 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 se tiene el modelo LM-SARMA, una última alternativa de orden superior que combina el retardo espacial con el error espacial, estas pruebas según Anselin (2005), son de alta potencia y permiten

elegir la alternativa adecuada que esté diseñada para representar nuestros datos.

Según Anselin y Rey (2014) antes de considerar las estadísticas robustas deben ser consideradas las versiones estándar (LM-Lag o LM-Error) y verificar que las pruebas sean significativas, Anselin y Rey (2014) manifiestan que todo el conjunto de estadísticas en principio pueden parecer desconcertantes, sin embargo, indica que existe una forma bastante intuitiva para poder avanzar a través de los diferentes resultados, hacia una especificación de regresión espacial, este proceso se puede resumir en la Figura 36 donde, se inicia en la parte superior considerando las estadísticas estándar, LM-Error y LM-Lag, si ninguno de estos test es significativo y no se puede rechazar la hipótesis nula se puede dejar el modelo estimado por mínimos cuadrados ordinarios (modelo ya estimado), en el caso de que entre los estadísticos de prueba uno rechace la hipótesis nula y el otro no, la decisión es sencilla, si al final los dos rechazan la hipótesis nula se puede proceder con la parte inferior del gráfico y considerar las formas robustas de los estadísticos de prueba o el modelo LM-SARMA, un modelo que puede combinar estas dos opciones.

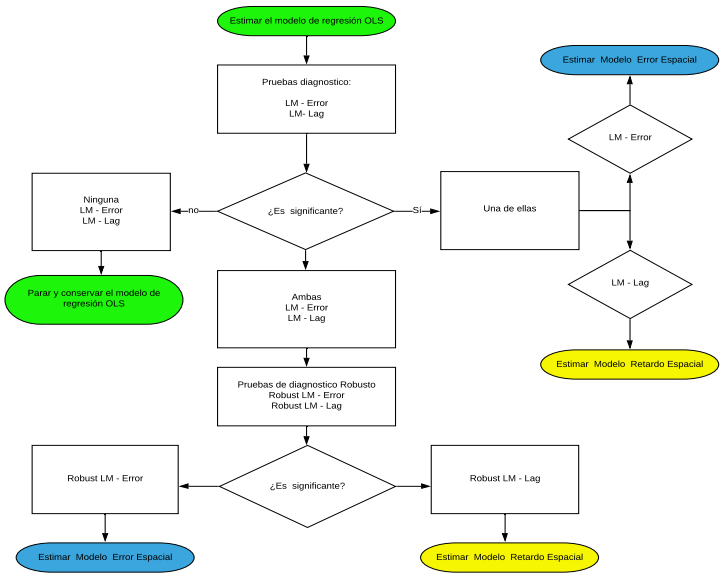


Figura 37: Ilustración generación del modelo espacial (Anselin y Rey, 2014)

Prueba LM-ERR: como se mencionó está basado en los multiplicadores de Lagrange (Yrigoyen, 2003), fue propuesto inicialmente por Burridge (1980) y su expresión es de la siguiente manera:

$$LM - ERR = \frac{\left[\frac{e'We}{s^2}\right]^2}{tr[W'W + W^2]} \tag{20}$$

Donde $s^2 = \frac{e'e}{N}$ es el estimador de máxima verosimilitud de la varianza de la perturbación aleatoria y e es el vector de residuos de la estimación por mínimos cuadrados ordinarios (MCO) (Yrigoyen, 2003).

Este tipo de contraste espacial es un test asintótico, distribuido como una χ^2 con un grado de libertad.

Prueba LM-EL: la prueba LM-EL es un tipo de contraste de dependencia espacial de la perturbación aleatoria, al igual que el test LM-ERR, sin embargo es robusto ya que permite la especificación de la existencia de esta dependencia en la variable respuesta espacialmente retardada (Anselin et al., 1996).

$$LM - EL = \frac{\left[\frac{e'We}{s^2} - T_1(R\tilde{J}_{\rho-\beta})^{-1}\frac{e'Wy}{s^2}\right]^2}{T_1 - T_1^2(R\tilde{J}_{\rho-\beta})} \tag{21}$$

con:

$$(R\tilde{J}_{\rho-\beta})^{-1} = \left[T_1 + \frac{(WX\beta)'M(WX\beta)}{s^2}\right]^{-1} \tag{22}$$

Donde:

$$T_1 = tr(W'W + W^2) \tag{23}$$

W es la matriz de pesos espaciales y $WX\beta$ el retardo espacial de los valores procedentes de la estimación por mínimos cuadrados ordinarios.

El estadístico de prueba se distribuye como una χ^2 con un grado de libertad, este fue calculado teniendo en cuenta que la matriz de pesos espaciales presentes en el termino del error es equivalente a aquella que se utiliza para la estimación del retardo espacial de la variable respuesta.

Prueba LM-LAG: el contraste de retardos espaciales LM-LAG de la variable respuesta es válido para la hipótesis de normalidad, hipótesis ya descrita en el análisis de los datos de distancia lingüística; ademas de ello su naturaleza es asintótica; esta prueba se distribuye al igual que los otros test como una variable de tipo χ^2 con un grado de libertad.

$$LM - LAG = \frac{\left[\frac{e'Wy}{s^2}\right]^2}{R\tilde{J}_{\rho-\beta}} \tag{24}$$

Donde y es el vector de tamaño $(N, 1)$ de las observaciones de la variable dependiente. Las pruebas LM-ERR y LM-LAG según Yrigoyen (2003) proporcionan la mejor alternativa entre la dependencia estructural y la aleatoria, como ya se mencionó, la selección del mejor test va a depender de los indicativos de significancia, el que tenga un mayor valor (confianza máxima) será aquel que muestre la dependencia espacial correcta y aquella dependencia a seleccionar.

Prueba LM-LE: el contraste de retardo espacial robusto se orienta a la medición de la presencia espacial local, su expresión es de la siguiente manera:

$$LM - LE = \frac{\left[\frac{e'W_1y}{s^2} - \frac{e'W_1e}{s^2}\right]^2}{R\tilde{J}_{\rho-\beta} - T_1} \tag{25}$$

Esta prueba sigue al igual que en casos anteriores la distribución de una χ^2 con un grado de libertad, siguiendo el supuesto que indica que las matrices de pesos espaciales son similares para la variable respuesta y la perturbación aleatoria de la misma.

Prueba SARMA: finalmente se puede contrastar la presencia tanto de dependencia espacial sustantiva junto con la residual, se trata de la prueba SARMA, este estadístico es planteado de la siguiente manera (Yrigoyen, 2003) :

$$SARMA = \frac{\left[\frac{e'Wy}{s^2} - \frac{e'We}{s^2}\right]^2}{R\tilde{J}_{\rho-\beta} - T_1} + \frac{\left[\frac{e'We}{s^2}\right]^2}{T_1} \tag{26}$$

Este estadístico se distribuye como una χ^2 con dos grados de libertad.

Finalmente, los estadísticos de prueba para el modelo generado fueron calculados e implementados en el software estadístico R, es de resaltar que la matriz de pesos fue la misma descrita anteriormente, matriz de contigüidad basada en el movimiento de la Reina dentro de un tablero de ajedrez.

Prueba	**Valor**	**Grados de Libertad**	**p-valor**
LM - ERR	17.496	1	2.879e-05
LM - EL	10.98	1	0.0009209
LM - LAG	49.895	1	1.622e-12
LM - LE	43.379	1	4.509e-11
SARMA	60.876	2	6.04e-14

Cuadro 34: Valores de pruebas de dependencia espacial

Como se mencionó, la selección del modelo depende del nivel de significancia; se puede observar en el Cuadro 34 que los valores son significantes y robustos para todas las pruebas siendo el p-valor menor a α ($\alpha = 0{,}05$) rechazando la hipotesis nula de no dependencia espacial, según Anselin (2005) antes de considerar las pruebas robustas es necesario considerar primero las estándar, en este caso la prueba LM-LAG es más significativa que LM-ERR así como la prueba robusta más significativa es para el rezago espacial, la prueba SARMA finalmente constituye el test de mayor significancia, lo que permite considerar tanto el modelo de rezago espacial como el modelo SARMA dentro de los modelos a implementar.

Estimación del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial

La construcción e implementación del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial se realizó en el software estadístico *R*, al igual que para el contraste de dependencia y con los Test mencionados en la sección anterior, se utilizó la matriz de pesos espaciales simétrica construida en el software GeoDa, los parámetros generados son los siguientes:

Variable	Estimate	Std.Error	z value	Pr(>—z—)
(Intercept)	0,103180	0,073724	1,399600	0,161640
Distancia Geográfica				
I(log(DistanciaG))	0,023852	0,009509	2,508400	0,012130
Altura media sobre el nivel medio del mar				
I(log(Altura))	0,000497	0,004832	0,102900	0,918060
Actividades Económicas				
Comercio	-0,014525	0,014536	-0,999300	0,317650
Ganadería	0,003928	0,008309	0,472800	0,636380
Industria	0,007504	0,014804	0,506900	0,612210
Minería	0,013829	0,015329	0,902200	0,366980
Pesca	0,005970	0,011274	0,529600	0,596410
Aptitud Agroclimática				
Extremadamente húmedo	-0,000041	0,022325	-0,001800	0,998530
Extremadamente seco	0,005630	0,010308	0,546200	0,584930
Ligeramente húmedo	-0,004228	0,008485	-0,498200	0,618310
Moderadamente húmedo	0,030111	0,015311	1,966700	0,049220
Moderadamente seco	0,000851	0,008955	0,095000	0,924300
Precipitación				
Entre 1500 y 2000	-0,002872	0,006058	-0,474	0,63547
Entre 2000 y 2500	-0,008194	0,008216	-0,9973	0,31862
Entre 2500 y 3000	-0,006818	0,007922	-0,8606	0,38947
Entre 3000 y 4000	0,013427	0,012588	1,0666	0,28615
Entre 4000 y 5000	0,010094	0,013271	0,7606	0,44687
Entre 500 y 1000	-0,000688	0,008557	-0,0804	0,93588
Entre 5000 y 7000	0,005806	0,023975	0,2422	0,80864
Entre 7000 y 9000	0,017498	0,025839	0,6772	0,49828
Amenaza Sísmica				
Extremadamente Alto	-0,015980	0,017448	-0,9159	0,35974
Extremadamente Bajo	-0,057156	0,040971	-1,395	0,16301
Moderadamente Alto	0,009563	0,006313	1,5149	0,1298
Muy Alto	0,003624	0,006084	0,5957	0,55138
Temperatura				
Entre 16 y 20 grados	-0,011731	0,007650	-1,533400	0,125170
Entre 20 y 22 grados	-0,009399	0,012337	-0,761900	0,446150
Entre 22 y 24 grados	-0,016288	0,011536	-1,412000	0,157960
Entre 24 y 26 grados	-0,021231	0,011700	-1,814600	0,069590
Entre 26 y 28 grados	-0,009330	0,013078	-0,713400	0,475590
Entre 8 y 12 grados	-0,011708	0,010575	-1,107200	0,268210
Mayor a 28 grados	-0,011457	0,014776	-0,775400	0,438100
IDH				
Húmedo	0,009976	0,028440	0,350800	0,725760
Muy húmedo	0,023243	0,026674	0,871400	0,383560
Muy seco	0,019521	0,012212	1,598400	0,109940
Seco	0,006916	0,009547	0,724500	0,468790
Semihúmedo	-0,019580	0,015239	-1,284800	0,198850
Semiseco	-0,005806	0,007607	-0,763300	0,445300
Vías				
Camino de herradura	-0,010851	0,020927	-0,518500	0,604100
Caminos carreteables	-0,084884	0,020114	-4,220200	0,000024
Carretera	-0,032409	0,012989	-2,495100	0,012590
Carretera y Férrea	-0,032430	0,016527	-1,962300	0,049730
Férrea	-0,065812	0,034376	-1,914500	0,055560
Rho: 0.54529, LR test value: 47.475, p-value: 5.5721e-12 **Asymptotic standard error: 0.067348** **z-value: 8.0966, p-value: 6.6613e-16** **Wald statistic: 65.555, p-value: 5.5511e-16**				

Cuadro 35: Estimación del modelo de rezago espacial

El modelo de rezago espacial muestra como el coeficiente ρ refleja la dependencia espacial inherente a los datos de la muestra (**0.54**), midiendo la influencia promedio en las observaciones de sus observaciones vecinas, en este caso tiene un efecto positivo y es muy significativo, en el caso de las variables independientes los efectos siguen siendo muy parecidos a los mencionados en el modelo no espacial, siendo la distancia geográfica un elemento de nuevo

relevante dentro de la estimación, las variable referente a la precipitación, así como la altura y las actividades económicas no tienen una influencia significativa dentro del modelo por lo que se puede pensar el prescindir de ellas para una mejor estimación, en cuanto a la aptitud agroclimática se entiende que aquellas zonas moderadamente húmedas presentan una alta significancia con respecto a el cambio en la distancia lingüistica.

Entre otras variables las más sugerentes hacen referencia a la temperatura, el índice de disponibilidad hídrica y la amenaza sísmica, variables que en alguno de sus rangos tienen un nivel de significancia alto (señalado en color verde el más significante y color amarillo aquel que tiene un poco menos de significancia), sin embargo difieren un poco en comparación con el modelo lineal no espacial estimado anteriormente; finalmente, las vías de acceso tienen una alta significancia dentro del modelo, con valores β negativos donde la vía de acceso con mayor influencia son los caminos carreteables, con un p-valor de **0.000024**; las variables explicativas en el modelo son significantes según el test de Wald, permitiendo contrastar la hipótesis acerca de la coherencia del modelo y el valor que las variables agregan al mismo, en este caso un test bastante significativo, descartando la hipótesis nula de no coherencia; dado que se utilizaron bastantes variables se intentará realizar una mejor estimación sustrayendo algunas variables, en este caso aquellas menos representativas e influyentes dentro del fenómeno lingüístico objeto de estudio y se culminará contrastando los modelos para escoger el de mayor veracidad.

Estimación del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial (Variables con mayor significancia (VMS))

La construcción e implementación del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial para las variables con mayor significancia se realizó en el software estadístico R, la matriz de pesos utilizada es equivalente a la ya generada y mencionada en la sección 7.3.1, las variables fueron seleccionadas a partir de múltiples iteraciones de integración y descarte a partir de las cuales se estimo el modelo con parámetros:

Variable	Estimate	Std.Error	z value	Pr(>—z—)
(Intercept)	-0,0041025	0,0490852	-0,0836	0,9333909
Distancia Geográfica				
I(log(DistanciaG))	0,0236311	0,0082084	2,8789	0,0039908
Actividades Económicas				
Minería	0,025458	0,012917	1,9709	0,0487367
Aptitud Agroclimática				
Extremadamente seco	0,0091526	0,0052458	1,7447	0,0810289
Ligeramente húmedo	-0,0068969	0,0063624	-1,084	0,2783598
Moderadamente húmedo	0,0259769	0,0125158	2,0755	0,0379371
Precipitación				
Entre 2000 y 2500	-0,0110223	0,0062273	-1,77	0,0767298
Entre 500 y 1000	0,0064258	0,0077946	0,8244	0,4097157
Amenaza Sísmica				
Moderadamente Alto	0,0068347	0,0061211	1,1166	0,2641779
Temperatura				
Entre 20 y 22 grados	-0,0058762	0,0098401	-0,5972	0,5503967
Entre 26 y 28 grados	0,0021984	0,0050846	0,4324	0,6654811
IDH				
Semihúmedo	-0,0203181	0,0108448	-1,8735	0,0609942
Semiseco	-0,0086323	0,0050384	-1,7133	0,0866584
Vías				
Camino de herradura	0,0711757	0,0202489	3,5150000	0,0004397
Caminos carreteables	0,0413851	0,0130986	3,1595000	0,0015804
Carretera	0,0443168	0,0165841	2,6722000	0,0075345
Carretera y Férrea	0,0683297	0,0162778	4,1977000	0,0000270
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**				

Cuadro 36: Estimación del modelo de retardo espacial para variables significativas

El Cuadro 36 muestra como los valores más significativos son utilizados para cada variable, dado que las variables utilizadas son diferentes al modelo de rezago espacial generado los $p - valores$ son en esencia distintos, los parámetros β también varían siendo en este caso una medida representativa del cambio, contrastando en forma positiva con los β calculados anteriormente, es destacable que el valor del intercepto es poco significativo, lo que hace que el estadístico de Wald tenga un valor alto y significativo y mostrando así la coherencia final del modelo, finalmente, el parámetro ρ que refleja la dependencia espacial inherente a los datos fue de **0.61495** es un valor alto permitiendo descartar la hipótesis nula de no correlación.

Estimación del modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR): La estimación del modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) se realizó en el software estadístico R, la matriz de pesos utilizada es la mencionada en la sección 7.3.1, para esta estimación en primer lugar se utilizaron todas las variables luego de ello las más significativas, cabe mencionar que este modelo SARAR incluye tanto las perturbaciones aleatorias como las variables sustantivas expresadas en el modelo de retardo espacial

por lo que es una combinación de los dos modelos, los parámetros del modelo SARAR y sus características estimadas para todas las variables son las siguientes:

Variable	Estimate	Std.Error	z value	Pr(>—z—)
(Intercept)	0,063060	0,056161	1,122800	0,261510
Distancia Geográfica				
I(log(DistanciaG))	0,012205	0,006984	1,747600	0,080540
Altura media sobre el nivel medio del mar				
I(log(Altura))	0,002029	0,004082	0,497200	0,619083
Actividades Económicas				
Comercio	-0,009307	0,012376	-0,752100	0,452020
Ganadería	0,012230	0,007845	1,559000	0,119007
Industria	0,013596	0,012982	1,047300	0,294946
Minería	0,007620	0,014276	0,533800	0,593513
Pesca	0,002870	0,008820	0,325400	0,744909
Aptitud Agroclimática				
Extremadamente húmedo	0,003796	0,019066	0,199100	0,842189
Extremadamente seco	0,004590	0,008360	0,549000	0,583017
Ligeramente húmedo	-0,006209	0,007198	-0,862600	0,388379
Moderadamente húmedo	0,022272	0,013320	1,672100	0,094498
Moderadamente seco	-0,000885	0,007021	-0,126000	0,899698
Precipitación				
Entre 1500 y 2000	-0,000430	0,005275	-0,081400	0,935093
Entre 2000 y 2500	-0,005873	0,007213	-0,814200	0,415514
Entre 2500 y 3000	-0,001966	0,006302	-0,312000	0,755007
Entre 3000 y 4000	0,016503	0,011007	1,499300	0,133789
Entre 4000 y 5000	0,006940	0,010732	0,646600	0,517864
Entre 500 y 1000	-0,000087	0,006039	-0,014400	0,988497
Entre 5000 y 7000	0,009193	0,022930	0,400900	0,688475
Entre 7000 y 9000	0,034297	0,020840	1,645700	0,099815
Amenaza Sísmica				
Extremadamente Alto	-0,031827	0,015849	-2,008100	0,044629
Extremadamente Bajo	-0,058235	0,039182	-1,486300	0,137209
Moderadamente Alto	0,011864	0,005441	2,180400	0,029225
Muy Alto	0,002571	0,005099	0,504300	0,614051
Temperatura				
Entre 16 y 20 grados	-0,012633	0,006775	-1,864500	0,062251
Entre 20 y 22 grados	-0,007614	0,010631	-0,716200	0,473889
Entre 22 y 24 grados	-0,010392	0,009820	-1,058200	0,289954
Entre 24 y 26 grados	-0,015554	0,010050	-1,547800	0,121674
Entre 26 y 28 grados	-0,006189	0,010938	-0,565900	0,571472
Entre 8 y 12 grados	-0,011930	0,009757	-1,222700	0,221439
Mayor a 28 grados	-0,009767	0,011777	-0,829300	0,406917
IDH				
Húmedo	0,011908	0,026193	0,454600	0,649383
Muy húmedo	0,008987	0,024453	0,367500	0,713244
Muy seco	0,011198	0,010063	1,112800	0,265786
Seco	0,000082	0,007975	0,010200	0,991829
Semihúmedo	-0,020475	0,014230	-1,438800	0,150196
Semiseco	-0,005696	0,006403	-0,889700	0,373647
Vías				
Camino de herradura	-0,025950	0,018754	-1,383700	0,166446
Caminos carreteables	-0,082517	0,017846	-4,623800	0,000004
Carretera	-0,039486	0,012213	-3,233200	0,001224
Carretera y Férrea	-0,039930	0,014902	-2,679500	0,007373
Férrea	-0,055689	0,031022	-1,795200	0,072628
Rho: 0.77318 **Asymptotic standard error: 0.061512** **z-value: 12.569, p-value: <2.22e-16** **Lambda: -0.74407** **Asymptotic standard error: 0.15913** **z-value: -4.6758, p-value: 2.9281e-06**				

Cuadro 37: Estimación del modelo SARAR

El modelo SARAR proporciona un ρ de dependencia espacial inherente en

los datos de **0.77318**, un valor mucho más alto que los modelos anteriores de rezago espacial, este modelo incorpora un valor adicional λ con un coeficiente de **-0.74407**, este valor como ya se mencionó es el coeficiente autorregresivo, en cuanto a las variables, no hay ningún cambio significativo más que el de nivel de significancia de la actividad económica, Ganadería con un p-valor de **0.119** que aunque es mayor que α sigue siendo importante, con respecto a la aptitud agroclimática, la precipitación, la amenaza sísmica, la temperatura y el IDH no hay mayores cambios, es un modelo que en definitiva reitera la incidencia de las variables dentro de la distancia lingüística, especialmente de las vías de acceso, de tal manera que se puede corroborar lo ya hecho en los modelos anteriores.

Estimación del modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) para variables significativas (VMS): La estimación del SARAR para las variables significativas igual que en ocasiones anteriores se realizó en el software estadístico R, los parámetros del modelo y sus características estimadas son las siguientes:

Variable	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-0,0195320	0,0348570	-0,5603000	0,5752460
Distancia Geográfica				
I(log(DistanciaG))	0,0147660	0,0062552	2,3605000	0,0182480
Actividades Económicas				
Minería	0,0182420	0,0115180	1,5838000	0,1132360
Aptitud Agroclimática				
Extremadamente seco	0,0058828	0,0043315	1,3582000	0,1744150
Ligeramente húmedo	-0,0081800	0,0058069	-1,4087000	0,1589370
Moderadamente húmedo	0,0190640	0,0114220	1,6691000	0,0951070
Precipitación				
Entre 2000 y 2500	-0,0101480	0,0057651	-1,7602000	0,0783760
Entre 500 y 1000	0,0055654	0,0058927	0,9445000	0,3449390
Amenaza Sísmica				
Moderadamente Alto	0,0088277	0,0052985	1,6661000	0,0957010
Temperatura				
Entre 20 y 22 grados	-0,0000774	0,0041051	-0,0189000	0,9849590
Entre 26 y 28 grados	-0,0054301	0,0089011	-0,6100000	0,5418350
IDH				
Semihúmedo	-0,0170580	0,0099993	-1,7059000	0,0880250
Semiseco	-0,0060900	0,0042182	-1,4437000	0,1488110
Vías				
Camino de herradura	0,0579870	0,0182080	3,1847000	0,0014490
Caminos carreteables	0,0343480	0,0115950	2,9622000	0,0030550
Carretera	0,0356180	0,0146940	2,4240000	0,0153520
Carretera y Férrea	0,0603920	0,0149170	4,0486000	0,0000515
Rho: 0.77931 **Asymptotic standard error: 0.061075** **z-value: 12.76, p-value: <2.22e-16** **Lambda: -0.50651** **Asymptotic standard error: 0.16726** **z-value: -3.0283, p-value: 0.0024591** **LR test value: 81.599, p-value: <2.22e-16**				

Cuadro 38: Estimación del modelo SARAR para variables significativas

El modelo SARAR generado para las variables significativas muestra como el valor de ρ aumenta en **0.003** con respecto al modelo anterior, sin embargo en este caso el número de reducido de variables hace que λ sea un valor menor, en este caso **-0.50651**; las vías de acceso siguen proporcionando la mayor fiabilidad y otorgan al modelo una bondad consistente verificada con el uso del test LR (prueba de razón de verosimilitud) con un valor alto y un p-valor de **2.22e-16** $< \alpha$, es decir, con un alto grado de significancia.

7.5. Selección del Modelo

Generados los modelos que incluyen información lingüística como respuesta de variables geográficas y metadatos del ALEC, se pudieron realizar pruebas de dependencia espacial sobre el modelo de regresión estimado por mínimos cuadrados ordinarios (MCO), modelo que sirvió como base para la estimación del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial, el modelo de retardo espacial para variables significativas, el modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) y el modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) para variables significativas, en este aparte se hará la selección del modelo óptimo con base en pruebas diagnostico.

7.5.1. Coeficiente de determinación

Según Nagelkerke (1991) el coeficiente de determinación denominado coeficiente de correlación múltiple es una medida de proporción de la varianza explicada por el modelo de regresión, refleja el éxito de la predicción de la variable dependiente a partir de las variables independientes; dado que para los modelos de regresión espacial se utilizó la estimación por máxima verosimilitud se utilizará el coeficiente de determinación pseudo R^2 de Nagelkerke, una medida equivalente definida de la siguiente manera:

$$R^2 = 1 - exp\left[-\frac{2}{n}\left\{l(\hat{\beta}) - l(0)\right\}\right] = 1 - \left\{\frac{L(0)}{L(\hat{\beta})}\right\}^{\frac{2}{n}} \tag{27}$$

Donde $l(\hat{\beta}) = log(L(\hat{\beta}))$ y $l(0) = log(L(0))$ denotando las probabilidades logarítmicas del modelo ajustado y "nulo", respectivamente.

Esta definición del pseudo coeficiente de determinación R^2 tiene las siguientes propiedades.

- Es consistente con el R^2 clásico.
- Es consistente con la máxima verosimilitud como método de estimación, es decir, las estimaciones de máxima verosimilitud de los parámetros del modelo maximizan a R^2.
- Es asintóticamente independiente del tamaño de la muestra n.
- Se puede interpretar como la proporción de variación explicada, o también, $1 - R^2$, que se puede decir que es la proporción de variación no explicada.
- Es adimensional.

Los valores de R^2 clásico y de Nagelkerke para los modelos estimados son los siguientes:

Tipo de Modelo	R^2	R^2 **Ajustado**	**Pseudo** R^2	ρ (dependencia espacial)
Modelo MCO	0.525	0.4227		
Modelo Retardo Espacial			0.6109	0.54529
Modelo Retardo Espacial VMS			0.58062	0.61495
Modelo SARAR			0.62926	0.77318
Modelo SARAR VMS			0.59323	0.77931

Cuadro 39: Estimación del coeficiente de determinación

El Cuadro 39 permite ver como los valores para los modelos espaciales son más altos en comparación con el de MCO, tanto el R^2 clásico como el R^2 ajustado, entre estos el modelo con mayor ajuste es el modelo SARAR para todas las variables, sin embargo se debe tener presente la dependencia espacial intrínseca entre los datos representada con el coeficiente ρ, entre mayor sea esta dependencia mucho mejor será la homogeneidad de los datos y mejor la estipulación del modelo, sin embargo para seleccionar el modelo correcto es necesario hacer algunas pruebas adicionales verificando tanto los valores predichos como sus residuos.

7.5.2. Análisis de términos del error

Los modelos espaciales al igual que los modelos de regresión lineal requieren que se cumplan con algunas hipótesis sobre los residuos, entre los más destacados se encuentran:

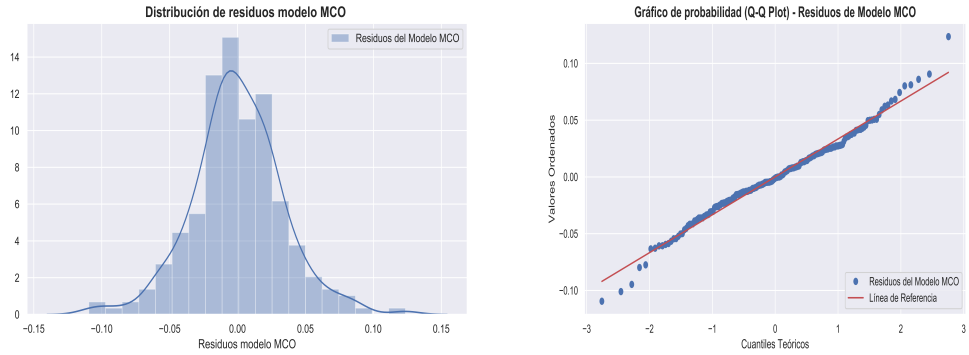
- Media cero: $E(e_i) = 0$ con $i = 1, 2, 3..n$
- Varianza constante: $Var(e_i) = \sigma^2 I$ con $i = 1, 2, 3..n$
- Residuos incorrelacionados: $Cov(e_i, e_j) = 0$

Cuando no se cumple alguna de las mencionadas hipótesis, se implica la no aleatoriedad de los residuos y por lo tanto la existencia de una estructura o relación no considerada en la estimación inicial del modelo, entre los dos efectos asociados a este comportamiento de ausencia de normalidad en los residuos se encuentra la heteroscedasticidad, y la autocorrelación o existencia de dependencia, efectos que pueden ser medidos con distintas pruebas estadísticas.

Heteroscedasticidad y normalidad: Se dice que en los residuos existe heteroscedasticidad cuando la varianza del mismo es diferente para las distintas observaciones que integran la muestra; su detección se puede realizar a través de distintos contrastes estadísticos, entre las pruebas más utilizados en modelos espaciales según Coro (2005) se encuentran el test del multiplicador de Lagrange, desarrollado por Breusch y Pagan, así como la versión "estudentizada"del mismo, esto va a depender de la distribución de la perturbación aleatoria (cuando esta no es normal, produce resultados sesgados).

Para el análisis de heteroscedasticidad de los residuos producidos por los modelos generados se tratará inicialmente de conocer su distribución así como de ejecutar en paralelo distintos test para el hallazgo de la misma, teniendo como medida base el contraste mencionado por Yrigoyen.

Modelo estimado por MCO: La estimación de los residuos para el modelo denominado MCO fue determinada en Python, su histograma es de la siguiente manera:



(a) Histograma (b) Grafico - QQ

Figura 38: Ilustración prueba normalidad modelo MCO

Al realizar los test de contraste de normalidad de los residuos del modelo MCO se obtiene en primer lugar que el estadístico de Shapiro-Wilk tiene un valor de **0.986** con un p-valor de **0.020**, en las pruebas de hipótesis el nivel de significancia utilizado α ha sido igual a 0,05, en este caso la hipótesis nula es que la distribución de los datos es normal, al ser el p-valor menor a alfa ($p < \alpha$) se rechaza la hipótesis nula de normalidad, por lo tanto es probable que los residuos no tengan esta distribución.

El test de de Anderson-Darling propuesto, se ejecutó de la misma manera que la prueba de Shapiro-Wilk, obteniendo un estadístico de **0.901844**, que para el nivel de significancia del 0,05 proporciona un valor crítico de **0.774**, se acepta la hipotesis nula de que los datos son normales cuando el estadístico de prueba calculado es menor que el valor crítico en un nivel de significancia elegido, dado que en este caso el estadístico es mayor al valor critico probablemente los datos no provengan de una distribución normal.

Para corroborar los resultados de los test de normalidad se utilizará finalmente la prueba D'Agostino's K^2 cuyo valor estadístico es de **8.049** con un p-valor de **0.018**, al igual que el test de Shapiro-Wilk la hipótesis nula expresa que los datos provienen de una distribución normal, como el nivel de significancia es menor a alfa se verifica la información resultante de las pruebas anteriores.

Teniendo como resultado que los residuos no provienen de una distribución normal se puede por último utilizar el test de Breusch-Pagan para la detección de heteroscedasticidad, este estadístico tiene el valor de **65.187** para el modelo MCO con un p-valor de **0.01243**, en este caso rechazando la hipótesis nula que que plantea que los datos son homocedásticos, es decir, existe heteroscedasticidad en los residuos del modelo.

La visualización de los residuos del modelo generado se realizó con el uso de polígonos de Thiessen, donde no solo se muestran los residuos sino también los valores predichos, es de aclarar que se debe verificar que estos residuos no sobrepasen el nivel de confianza $\pm 2\sigma = 0{,}052$, donde $\sigma =$**0.02625917**.

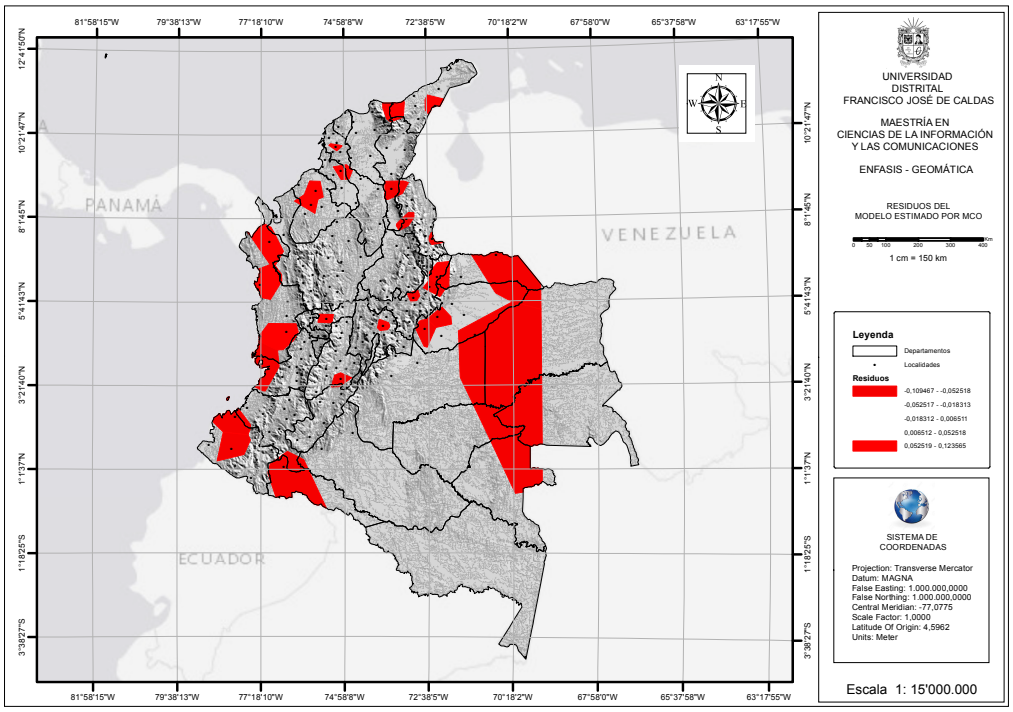


Figura 39: Residuos del modelo estimado por MCO

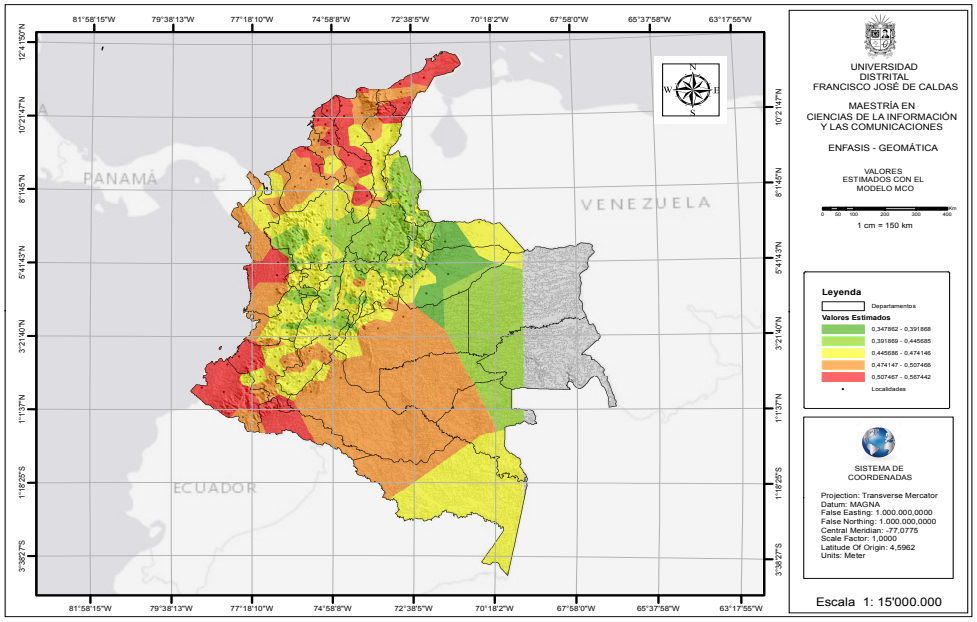


Figura 40: Valores estimados MCO

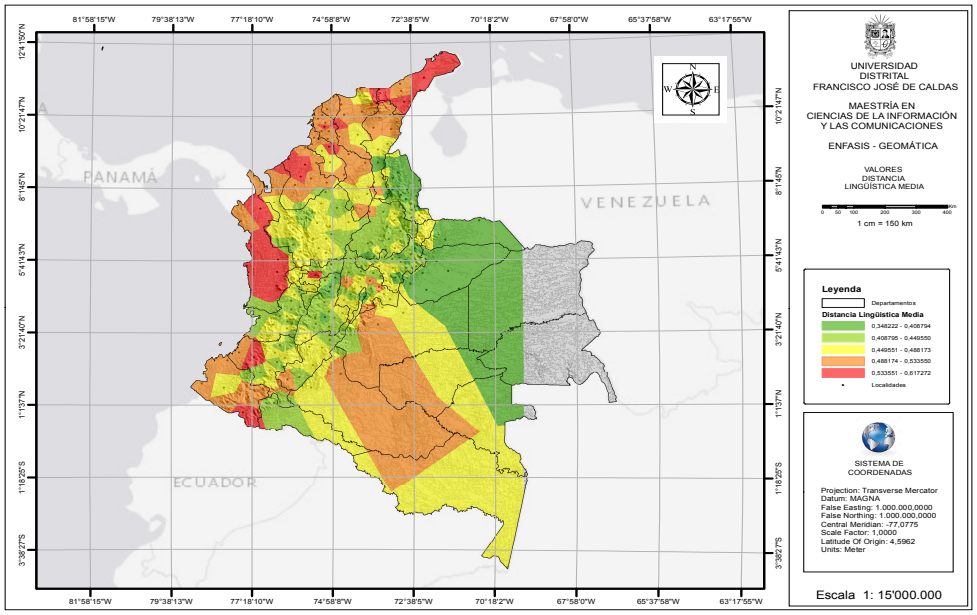
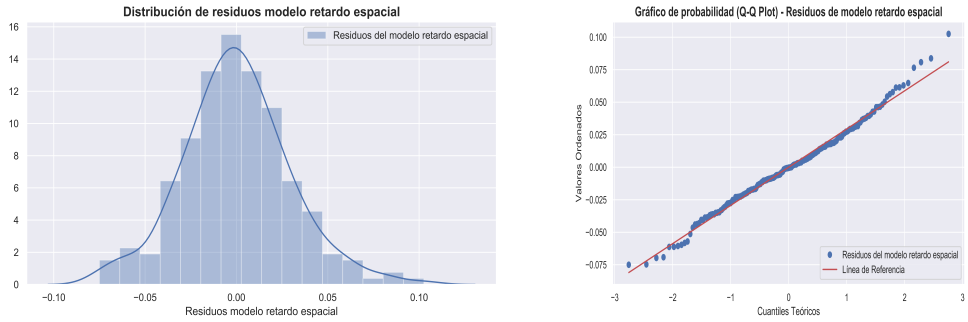


Figura 41: Valores distancia lingüistica media

En la Figura 39 se pueden visualizar los residuos del modelo estimado por MCO, en color rojo se señalan aquellos lugares en los que se supera el nivel de significancia señalado atrás, estos valores se encuentran principalmente en las fronteras especialmente en la zona Este del país, dado que los valores de los residuos están relacionados con los valores estimados, en las Figuras 40 y 41 se pueden observar que las zonas coinciden y que se encuentran en distinto color, además que el rango estimado es mucho más amplio por lo que el **estadístico de determinación 0.525 concuerda con la ilustración**.

Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial: Los residuos del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial fueron contrastados con la hipótesis de normalidad con base en las gráficas de probabilidad QQ y el histograma, la ilustración generada en Python es la siguiente:



(a) Histograma (b) Grafico - QQ

Figura 42: Ilustración prueba normalidad modelo retardo espacial

La Figura 42 muestra como los valores de los residuos se concentran en el valor cero (0) con una media de **-2.9854e-17** cumpliendo uno de los criterios generales para los residuos del modelo, el gráfico de probabilidad QQ muestra que las observaciones se encuentran alineadas con la distribución normal, por último, el contraste de Shapiro-Wilk tiene un valor de **0.990** con un p-valor de **0.090**,al igual que en el caso anterior las pruebas de hipótesis el nivel de significancia utilizado α ha sido 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 lo tanto es pro-

bable que los residuos tengan una distribución normal.

La segunda prueba de normalidad utilizada es aquella propuesta por Anderson-Darling, este test ofreció un resultado para el estadístico de **0.6298**, que para el nivel de significancia del 0,05 proporciona un valor crítico de **0.774**; se acepta la hipótesis nula de que los datos son normales cuando el estadístico de prueba calculado es menor que el valor crítico en un nivel de significancia elegido, en este caso el estadístico es menor al valor critico, por lo tanto probablemente los residuos tengan una distribución normal.

Finalmente se utilizará el test de contraste de D'Agostino's K^2 cuyo valor estadístico es de **5.919** con un p-valor de **0.052**, al igual que el test de Shapiro-Wilk la hipótesis nula expresa que los datos provienen de una distribución normal, como el nivel de significancia es mayor a alfa se puede afirmar que los residuos provienen de una distribución normal.

Se puede concluir que los residuos provienen de una distribución normal, para detectar la homogeneidad o heteroscedasticidad dentro de ellos se utilizará el test de Breusch-Pagan cuyo estadístico fue **69.634** con un p-valor de **0.004672**, la hipótesis nula plantea que existe homocedasticidad, al ser el valor de α menor a 0,05 ($\alpha < 0{,}05$) se rechaza esta hipótesis nula, por lo tanto existe heteroscedasticidad en los residuos del modelo y en el modelo mismo.

Para realizar la visualización de los residuos del modelo se utilizará como en el caso anterior los polígonos de Voronoi o de Thiessen, se mostrarán los residuos teniendo presente el nivel de significancia de los mismos, el nivel de confianza para este modelo será igual a $\pm 2\sigma$, donde sigma es $\sigma =$ **0.02931633**, es decir $\pm 2\sigma = \pm$**0.05863265**; para que los residuos cumplan criterio de estimación óptima el 0,95 del total de los datos debe estar en este intervalo.

La siguiente ilustración muestra como estos residuos tienen de manera intrínseca una relación estipulada en cuantiles, en color rojo se expresa el valor superior a los niveles de confianza dentro de cada uno de los polígonos

que contienen los centroides correspondientes a las localidades (Figura 44).

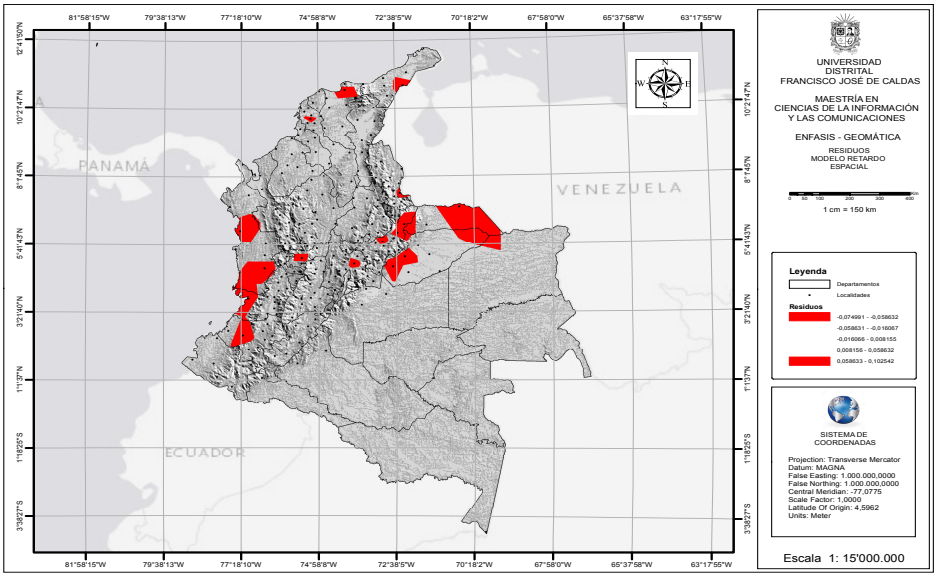


Figura 43: Residuos del modelo retardo espacial

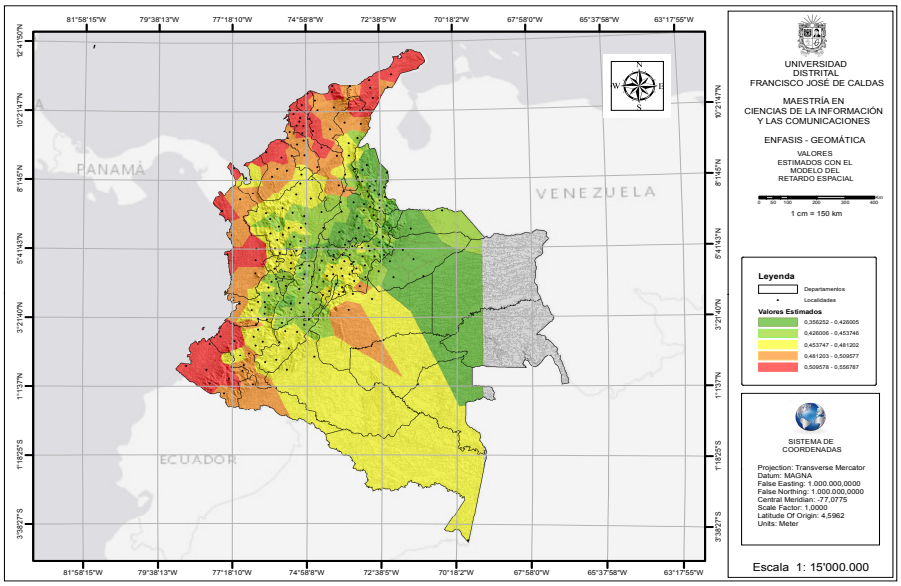
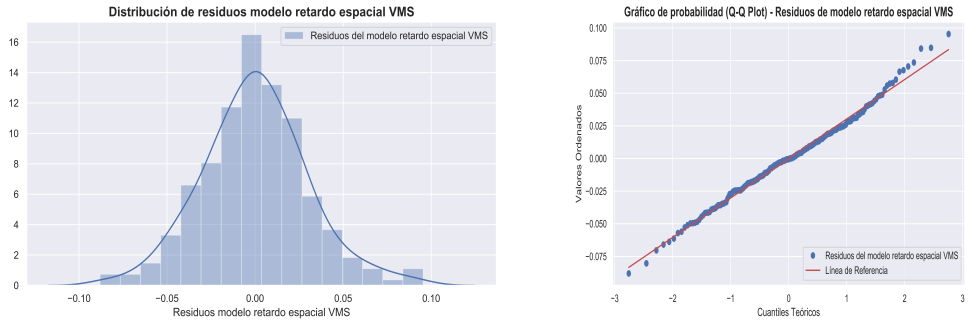


Figura 44: Estimados modelo retardo espacial

En la Figura 43 se pueden visualizar los residuos del modelo de retardo espacial, en color rojo se señalan aquellos lugares en los que se supera el nivel de significancia señalado atrás, estos valores en comparación con el modelo estimado por MCO son menores, y la ubicación geográfica se distribye con aleatoriedad, como se señaló los valores de los residuos están relacionados con los valores estimados, en la Figura 44 se pueden observar que las zonas coinciden en su mayor parte con unos tonos equivalentes a los de la distancia lingüistica (Figura 41), cabe destacar que el rango estimado es mucho más reducido que el de la distancia lingüistica media pero se ajusta más que los valores del modelo estimado por MCO, por lo que el estadístico de determinación 0,6109 se evidencia en la ilustración.

Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial (Variables con mayor significancia (VMS)): Los residuos del modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial para las variables con mayor significancia fueron contrastados con pruebas de normalidad teniendo como base las gráficas de probabilidad QQ y el histograma, las imágenes generadas en Python son la siguientes:



(a) Histograma (b) Grafico - QQ

Figura 45: Ilustración prueba normalidad modelo retardo espacial VMS

La Figura 45 muestra como los valores se concentran para el histograma del modelo retardo espacial VMS en cero ($\mu =$ **7.46368419e-18**), la forma de campana supone que la distribución de estos residuos es normal, el gráfico de probabilidad QQ normal muestra que los valores se ajustan en un **95 %**, esto a la vez supone que se está cumpliendo con uno de los criterios para los

residuos del modelo estimado.

La prueba de Shapiro-Wilk de contraste de normalidad tiene un estadístico calculado de **0.993** con un p-valor de **0.357**, como en casos anteriores las pruebas de hipótesis se han realizado con un 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 lo tanto es probable que los residuos tengan esta distribución.

En segundo lugar se utilizó la prueba de normalidad de Anderson Darling este test ofreció un resultado para el estadístico de **0.474872**, para el nivel de significancia del **0.05** proporciona un valor crítico de **0.774**; como ya se mencionó se acepta la hipótesis nula de que los datos son normales cuando el estadístico de prueba calculado es menor que el valor crítico en un nivel de significancia elegido, en este caso el estadístico es menor al valor critico, es probable entonces la hipótesis de normalidad, la última prueba a utilizar es el test de contraste de D'Agostino's K^2 cuyo valor estadístico es de **3.582** con un p-valor de **0.167**, la hipótesis nula expresa que los datos provienen de una distribución normal, como el nivel de significancia es mayor a alfa (α) se puede afirmar que según los test los residuos provienen de una distribución normal.

Para la detección de homogeneidad o heteroscedasticidad espacial dentro de los residuos se utilizará el test de Breusch-Pagan cuyo estadístico calculado fue **25.583** con un p-valor de **0.06018**, la hipótesis nula plantea que existe homocedasticidad, al ser el valor de α mayor a 0,05 ($\alpha > 0{,}05$) se acepta la hipótesis nula, por lo tanto existe homocedasticidad u homogeneidad en el modelo estimado.

La visualización de los residuos se hará con el uso de polígonos de Voronoi o de Thiessen, se mostrarán los residuos teniendo presente el nivel de confianza de los mismos, para este modelo será igual a $\pm 2\sigma$, donde sigma es $\sigma = \mathbf{0.03013126}$, es decir $\pm 2\sigma = \pm$ **0.06026252**.

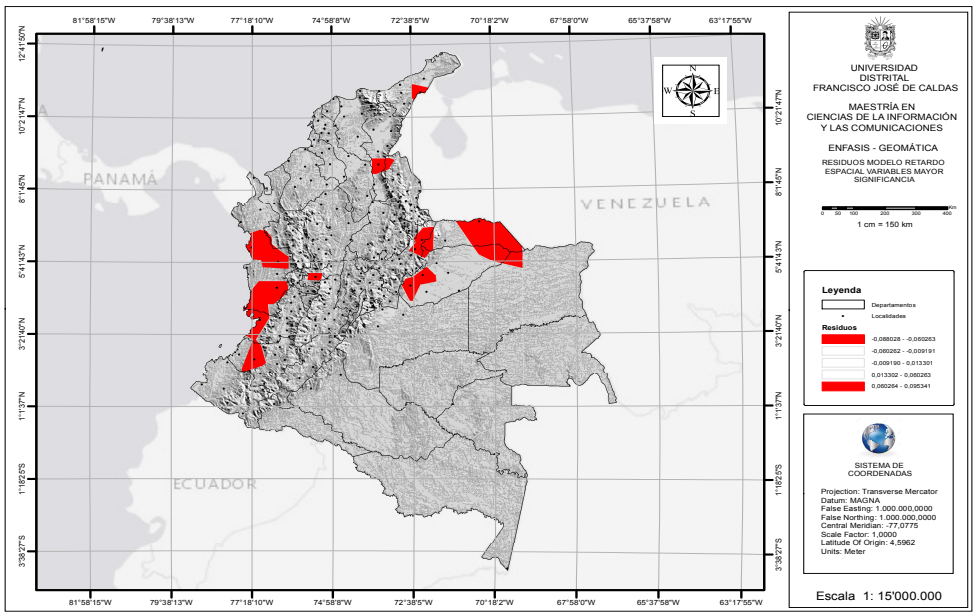


Figura 46: Residuos del modelo retardo espacial VMS

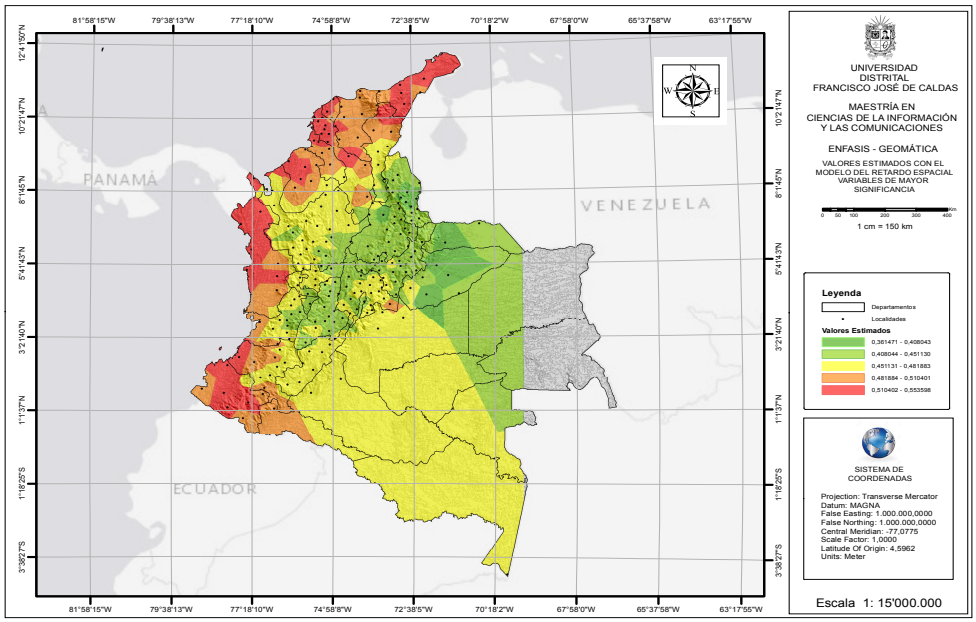
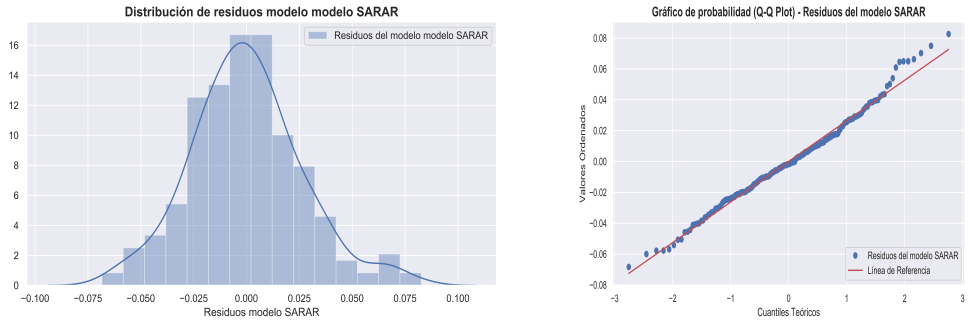


Figura 47: Valores estimados modelo retardo espacial VMS

Exceptuando algunas áreas espaciales observadas en las figuras 46 y 47 se revela como la distancia lingüística estimada con el modelo de retardo espacial se distribuye de una forma homogénea en todo el territorio, diferenciando aquellas zonas donde la distancia lingüística es mayor de aquellas donde no lo es, presentando esta distancia de una forma uniforme, es de resaltar que cerca del océano pacífico y en la costa atlántica es donde se presenta la mayor diferencia en comparación con el centro del país. El rango de valores en forma general se ajusta mucho mejor que el modelo anterior destacando que el I de Moran (indice de autocorrelación espacial) tiene un estadístico para los residuos de **-1.2771** con un p-valor de **0.2016**, es decir, se rechaza la hipótesis nula de autocorrelación y se puede decir que los residuos se distribuyen de forma aleatoria en el espacio.

Modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR): Los residuos del modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas se pueden representar como en otros casos con las gráficas de probabilidad QQ y el histograma, de la siguiente manera:



(a) Histograma (b) Grafico - QQ

Figura 48: Ilustración prueba normalidad modelo SARAR

La Figura 48 muestra como los valores se concentran para el histograma del modelo SARAR cero ($\mu = 0$), la forma de campana de residuos asemeja una distribución normal, la prueba de Shapiro-Wilk de contraste de normalidad tiene un estadístico calculado de **0.989** con un p-valor de **0.055**, al ser el p-valor mayor a alfa ($p > \alpha$) no se rechaza la hipótesis nula de normalidad,

por lo tanto es probable que los residuos tengan esta distribución.

La prueba de normalidad de Anderson Darling tiene un estadístico de **0.70140745**, para el nivel de significancia del **0.05** proporciona un valor crítico de **0.774**; en este caso el estadístico es menor al valor critico, por lo tanto es probable la hipótesis de normalidad, la última prueba, el test de contraste de D'Agostino's K^2 cuyo valor estadístico es de **6.362** con un p-valor de **0.042**, como el nivel de significancia es menor a alfa (α) los residuos no provienen de una distribución normal, en contraste esta distribución de los residuos puede que sea normal, aunque el nivel de significancia en los test es bajo.

En cuanto a la heteroscedasticidad espacial dentro de los residuos, el test de Breusch-Pagan cuyo estadístico calculado fue **62.611** tiene un p-valor de **0.02117**, al ser el valor de α menor a 0,05 ($\alpha > 0{,}05$) se rechaza la hipótesis nula, por lo tanto existe heteroscedasticidad, finalmente el nivel de confianza es igual a $\pm 2\sigma$, es decir $\pm 2\sigma = \pm$ **0,05251834**.

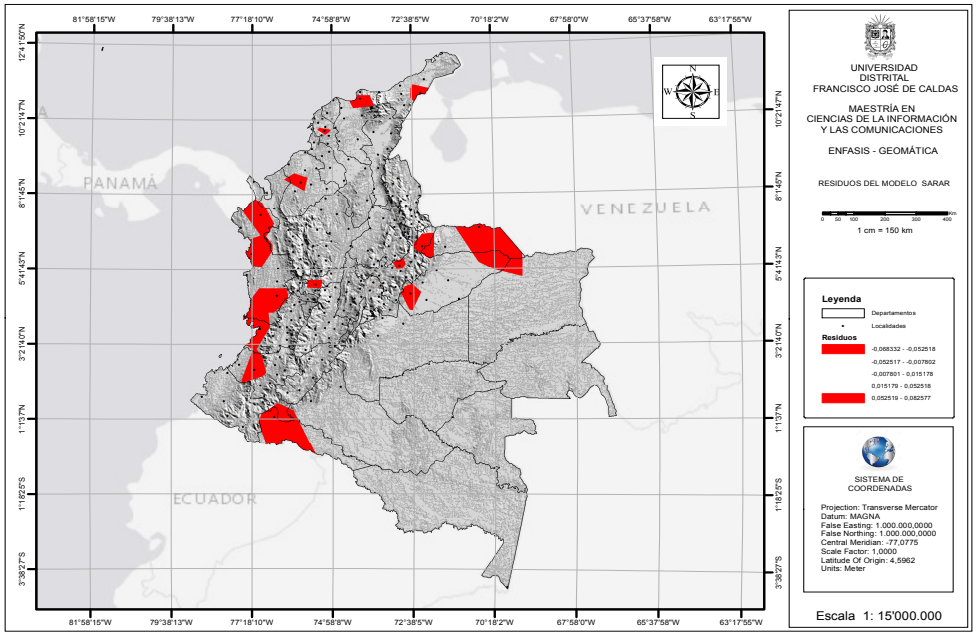


Figura 49: Residuos del modelo SARAR

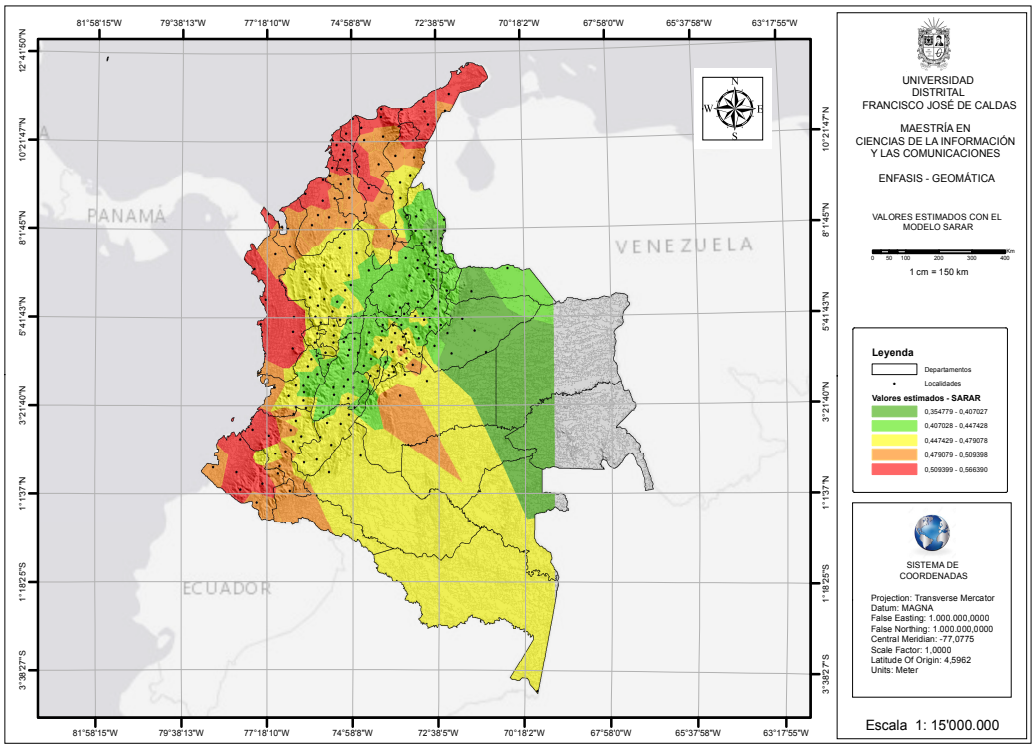
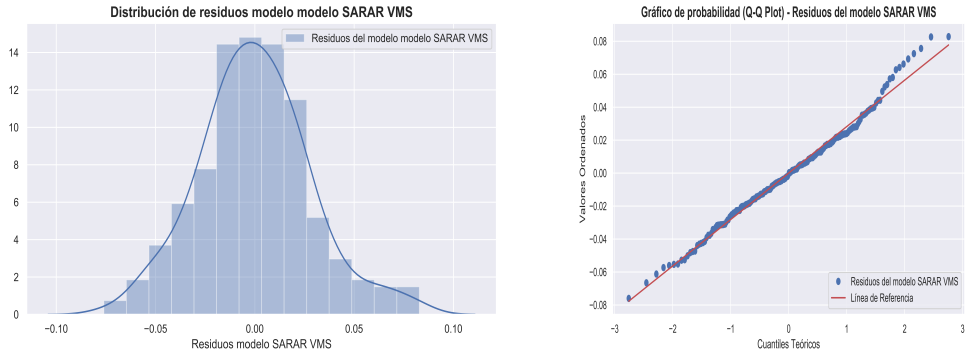


Figura 50: Valores estimados modelo SARAR

Este modelo a diferencia del modelo anterior muestra como los residuos superan en muchos casos el nivel de confianza estimado, aquellas zonas en color rojo (ver Figura 49) aparecen especialmente en los limites orientales y occidentales del país, el rango de valores en forma general se ajusta a la distancia lingüística y presenta una covarianza cercana a 0 (**0.00068954**), el índice I de Moran (indice de autocorrelación espacial) tiene un estadístico para los residuos de **-0.25413** con un p-valor de **0.7994**, por lo tanto se puede decir que los residuos se distribuyen en forma aleatoria.

Modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR) para variables significativas (VMS): Los residuos del modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas SARAR (VMS) se pueden representar de la siguiente manera:



(a) Histograma (b) Grafico - QQ

Figura 51: Ilustración prueba normalidad modelo SARAR VMS

La Figura 51 resalta la concentración de los valores en cero (μ =**3.73184e-18**), la prueba de Shapiro-Wilk de contraste de normalidad tiene un estadístico calculado de **0.990** con un p-valor de **0.112**, al ser el p-valor mayor a alfa ($p > \alpha$) es probable que los residuos tengan una distribución normal, el test de Anderson Darling tiene un estadístico de **0.542028765338**, para el nivel de significancia del 0,05 proporciona un valor crítico de **0.774**; en este caso el estadístico es menor al valor critico, por lo tanto es probable la hipótesis de normalidad, la última prueba, el test de contraste de D'Agostino's K^2 cuyo valor estadístico es de **4.679** con un p-valor de **0.096**, estos test prueban que los residuos probablemente se distribuyen normalmente.

En cuanto a la heteroscedasticidad espacial dentro de los residuos, el test de Breusch-Pagan con estadístico **27.053** y un p-valor de **0.04089**, muestran la existencia de este fenómeno, finalmente en este modelo se utilizará un nivel de confianza igual a $\pm 2\sigma$, es decir $\pm 2\sigma = \pm$ **0,05621706** para la visualización de los datos resultantes.

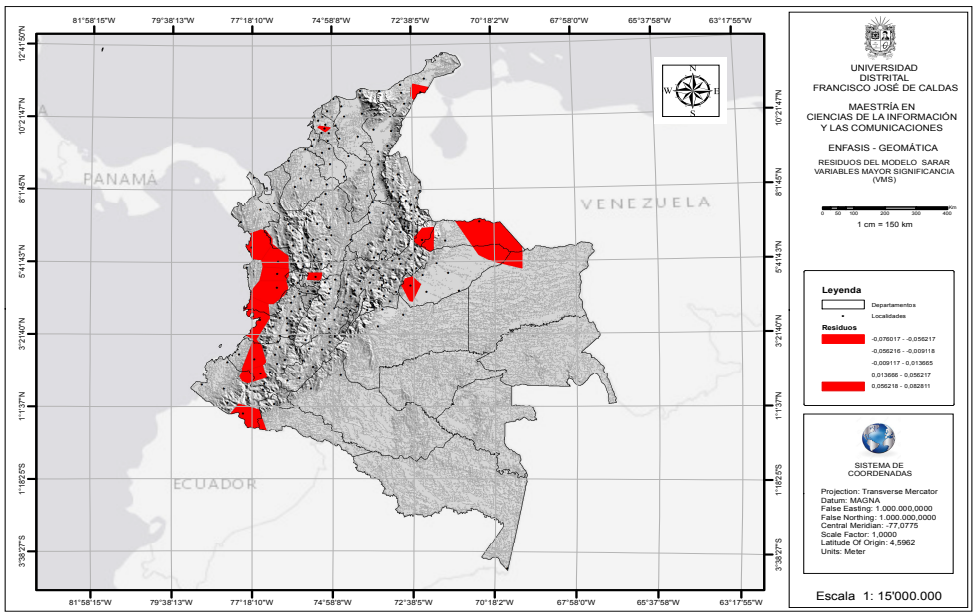


Figura 52: Residuos del modelo SARAR VMS

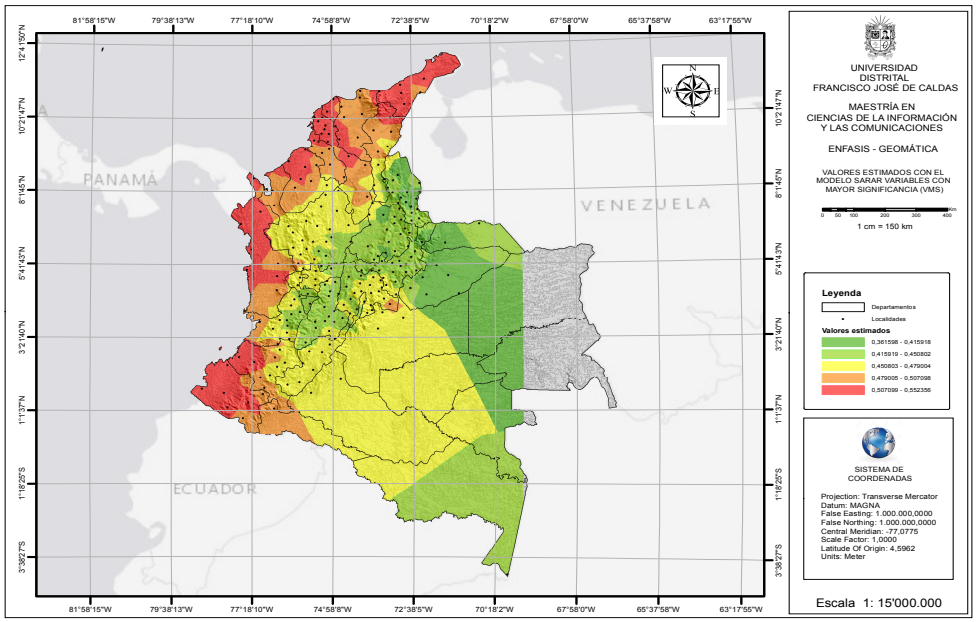


Figura 53: Valores estimados modelo SARAR VMS

Este modelo como el anterior superan en muchos casos el nivel de confianza estimado, sin embargo existe homogeneidad espacial con respecto a los valores de la distancia lingüística, aún cuando presenta una covarianza cercana a 0 (**0.00079009**) en los residuos, después de este análisis el modelo que se ajusta a todos los criterios de ajuste en los residuos es el modelo de retardo espacial para las variables significativas, los otros modelos aunque estimados faltan a uno o mas supuestos.

Autocorrelación espacial: Con el propósito de contrastar la hipótesis de que los residuos se encuentran localizados de forma aleatoria en el espacio existe el estadístico de I de moran, que calcula la asociación de valores similares o disimiles entre regiones vecinas (Vayá y Moreno, 2000), a continuación se presentan los valores para cada uno de los modelos:

Tipo de Modelo	**I de Moran**	**p - valor**
Modelo MCO	4.3583	1.311e-05
Modelo Retardo Espacial	-1.1631	0.2448
Modelo Retardo Espacial VMS	-1.2771	0.2016
Modelo SARAR	-0.25413	0.7994
Modelo SARAR VMS	-0.051805	0.9587

Cuadro 40: Autocorrelación residuos modelos

Dado que en los modelos espaciales generados el p-valor es mayor a alfa ($\alpha = 0{,}05$) 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, el patrón espacial es aleatorio en forma general; para el modelo generado por MCO existe una alta autocorrelación espacial positiva con significancia, esta estimación no sería la adecuada si se desea seleccionar un modelo eficiente.

7.5.3. Sumario modelo seleccionado:

El modelo final seleccionado es el modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial para las variables con mayor significancia (VMS); este modelo cumple con los supuestos de regresión estimados, tanto para la los valores ajustados como para los residuos, a continuación se

presentan las principales características del modelo y su representación gráfica dentro del territorio.

Modelo del retardo espacial VMS		
Coeficiente de determinación	0.58062	
Media Residuos	7.46368419e-18	
Pruebas de Normalidad		
Prueba	**Estadístico**	**p-valor**
Shapiro-Wilk	0.993	0.357
D'Agostino's K^2	3.582	0.167
Anderson Darling	**Estadístico**	**Valor Crítico**
	0.474872	0.774
Heteroscedasticidad		
Breusch Pagan	25.583	0.06018
Autocorrelación espacial en residuos	**Estadístico**	**p-valor**
	-1.2771	0.2016
Covarianza	0.00126386	

Cuadro 41: Sumario modelo seleccionado

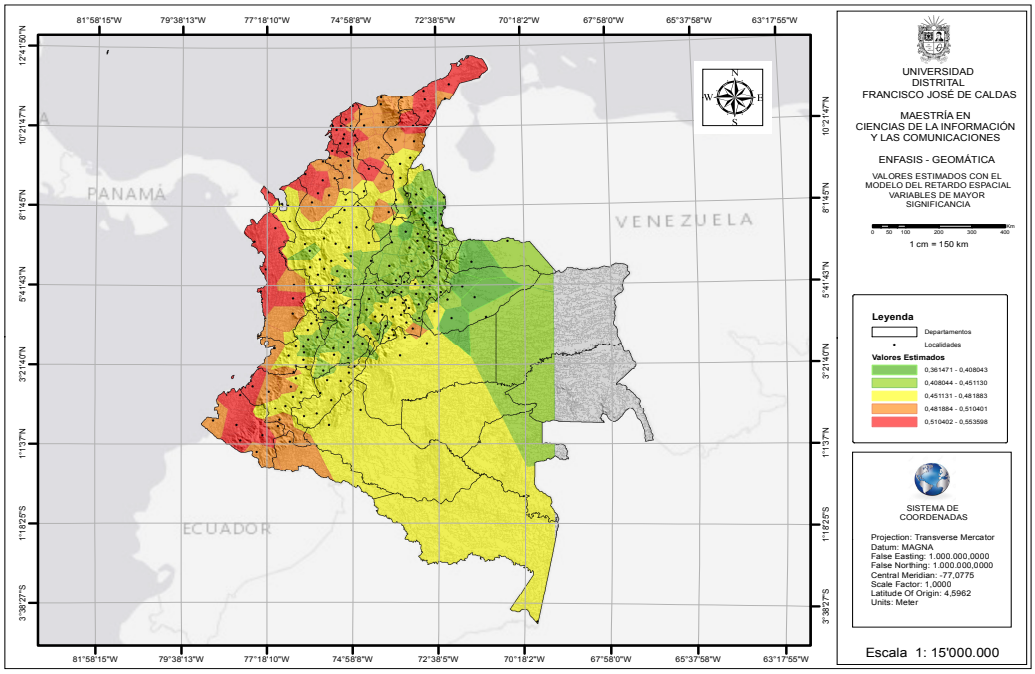


Figura 54: Modelo seleccionado: valores estimados

7.6. Generación de limites lingüísticos

Para una correcta clasificación dialectal se utilizarán limites geográficos, según Heeringa y Nerbonne (2001) los dos paradigmas clave de la dialectología son precisamente la forma como se conceptualizan estos límites, ya sea como isoglosas o como superficies dialectales continuas, las isoglosas como se mencionó atrás son lineas imaginarias que delimitan y separan las ocurrencias de diferentes variantes producto de fenómenos lingüísticos, mientras que la teoría de dialecto continuo plantea que la variación lingüística, ya sea individual o agregado, se distribuye espacialmente de forma gradual (Jeszenszky y Weibel, 2015), en esta investigación se seguirán estas dos teorías, teniendo como premisa la investigación de Grieve et al. (2011), donde se utilizaron diferentes herramientas para delimitar áreas dialectales, entre ellas la autocorrelación espacial, el análisis factorial y el análisis de conglomerados (clustering), vinculando la teoría de isoglosas con el dialecto continuo, es de resaltar que Jeszenszky y Weibel (2015) plantean que en realidad los fenómenos lingüísticos individuales analizados en pocas ocasiones muestran el tipo de patrones regionales bien definidos que a menudo se exhiben en los estudios tradicionales de clasificación de dialecto, sin embargo se intentará contrastar esta teoría con los resultados de la presente investigación.

7.6.1. Superficie dialectal continua

Para la generación de una superficie dialectal continua se utilizará el método geoestadístico de interpolación Kriging, este método según Hoch y Hayes (2010) utiliza el análisis de semivarianza, herramienta útil para comprender la variación lingüistica relacionada con la dependencia espacial, revelando información importante sobre las tasas de cambio en el espacio geográfico en función de las distancia entre las observaciones, o en dado caso la aleatoriedad de la distribución de la muestra objeto de estudio en el espacio.

El método de interpolación espacial Kriging basado en el análisis de semivarianza reconoce que las variables espaciales son bastante estocásticas para ser cartografiadas con métodos deterministas (Hoch y Hayes, 2010), por lo tanto utiliza variables regionalizadas con un componente sistemático (como la media), un componente estocástico (autocorrelación espacial) y un componente aleatorio (ruido o error) (Burrough y McDonnell, 1998), uno de las ventajas de utilizar este método es que proporciona estimaciones del error

para los valores interpolados en cualquier punto del espacio, determinando así la calidad de la estimación (Stein, 1999).

Variograma: El variograma es una herramienta para el análisis de datos espaciales que sirve como base para el Kriging (Webster y Oliver, 2007), según Lichtenstern (2013) el variograma tiene como conjetura que la relación espacial de dos observaciones no depende de su ubicación geográfica absoluta en sí, sino solo de su ubicación relativa; el variograma teórico definido por Matheron y Blondel (1962) y Cressie (2015) se expresa de la siguiente manera:

$$2\gamma(h) = Var(Z(x+h) - Z(x)) \tag{28}$$

Donde $Z(x)$ es el valor de la variable en un sitio x, $Z(x+h)$ es otra observación separada de la anterior por una distancia h y n es el número de parejas que se encuentran separadas por esa distancia.

La interpretación del semivariograma parte del criterio de que a menor distancia geográfica entre las localidades mayor similitud espacial entre las observaciones (Giraldo, 2015), premisa mencionada y verificada en la sección 7.4.

A continuación se presenta el variograma experimental teniendo presente que la variable regionalizada z es la distancia lingüística generada por el modelo estimado, cabe resaltar que su implementación se realizó en el software ArcGIS y fue verificado en el software SpaceStat:

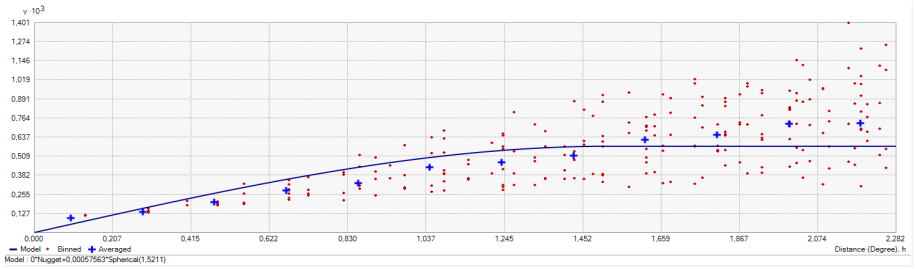


Figura 55: Variograma con valores estimados

El variograma en el origen tiene un comportamiento regular con un movimiento parabólico, lo que supone una distribución heterogénea en el espacio, el *spline* de ajuste al que se recurrió fue la función regular esférica, tratando de minimizar y mejorar el ajuste.

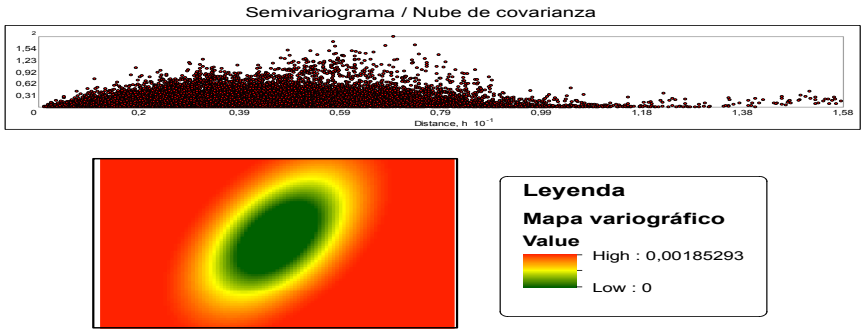


Figura 56: Mapa variográfico y nube variográfica

El mapa variográfico ilustrado en la Figura 56 muestra el variograma experimental en diferentes direcciones del espacio, en este caso en una escala de color, ayudando a distinguir la anisotropía, es decir, intentando conocer si el comportamiento de los valores de distancia lingüistica estimada pueden tener direcciones preferentes de continuidad en el espacio (Emery, 2013); el mapa generado para los valores de distancia lingüística muestra una mayor continuidad en la orientación norte con sesenta grados al este, por otro lado se tiene la nube variográfica, es decir, la nube de las diferencias cuadráticas, en este caso se muestra una concentración en el centro-izquierdo del mismo permitiendo poner en evidencia que existe en general una heterogeneidad en los datos y una diferencia mínima entre los puntos vecinos.

Construcción del Kriging: Para la construcción del Kriging se va a tener en cuenta que la variable regionalizada, en este caso la distancia lingüística estimada con el modelo de retardo espacial, es la realización de una función aleatoria Z estacionaria con media desconocida (Kriging Ordinario) tal que:

$$\begin{cases} \forall \mathbf{x} \in V, E[z(\mathbf{x})] = m \ desconocida) \\ \forall \mathbf{x}, \mathbf{x}+h \in V, \ cov[Z(\mathbf{x}+h), Z(x)] = C(h) \end{cases} \tag{29}$$

Donde V es la representación de la vecindad del Kriging (Emery, 2013).

Las propiedades principales del estimador Kriging son las siguientes:

- Linealidad: esta restricción se asegura cuando se toma como estimador en x_0.

$$Z^*(x_0) = a + \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha) \tag{30}$$

Donde los λ_α representa los pesos o ponderaciones de los valores originales

- Insesgo: en este caso la suma de los pesos debe ser igual a uno, de esta manera la esperanza del predictor es igual a la esperanza de la variable, la propiedad de insesgamiento se puede expresar de la siguiente manera.

$$E(Z^*(x_0)) = E(Z^(x_0)) \tag{31}$$

- Optimalidad: bajo la condición de insesgamiento se debe minimizar la varianza del error de predicción, es decir, reducir la expresión:

$$V(Z^*(x_0) - Z(x_0)) \tag{32}$$

El sistema de Kriging ordinario contiene una incógnita y una ecuación más que el sistema simple. Su notación matricial se puede escribir de la siguiente manera:

$$\begin{pmatrix} C(x_1 - x_1) & \cdots & C(x_1 - x_n) & 1 \\ \vdots & & \vdots & \vdots \\ C(x_n - x_1) & \cdots & C(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C(x_1 - x_0) \\ \vdots \\ C(x_n - x_0) \\ 1 \end{pmatrix} \tag{33}$$

El kriging ordinario tiene como principal característica el variograma (γ), una herramienta equivalente a la covarianza (C), cuya relación se denota $\gamma(h) = C(0) - C(h)$, su representación matricial final es la siguiente:

$$\begin{pmatrix} \gamma(x_1 - x_1) & \cdots & \gamma(x_1 - x_n) & 1 \\ \vdots & & \vdots & \vdots \\ \gamma(x_n - x_1) & \cdots & \gamma(x_n - x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ -\mu \end{pmatrix} = \begin{pmatrix} \gamma(x_1 - x_0) \\ \vdots \\ \gamma(x_n - x_0) \\ 1 \end{pmatrix} \tag{34}$$

La estimación final de la superficie continua se realizó en el software ArGIS, la visualización así como sus principales características se muestra a continuación:

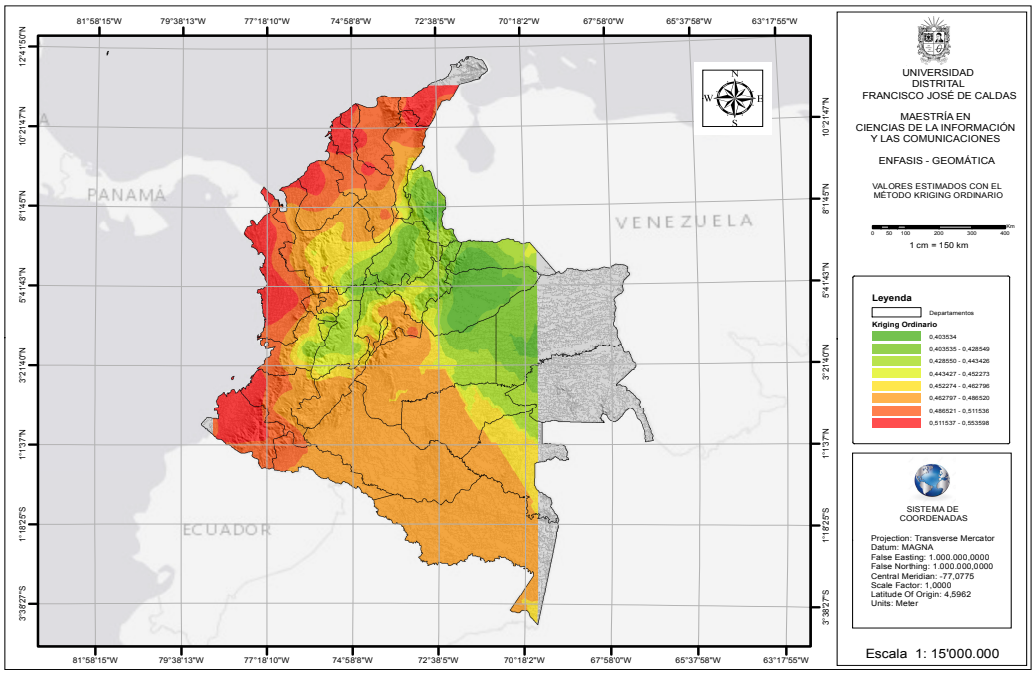


Figura 57: Kriging ordinario valores estimados

Entre las principales características de la estimación de Kriging Ordinario, se encuentra aquella que plantea que Kriging es un estimador exacto, según Emery (2013) la estimación para un sitio en particular vuelve con el mismo dato, mientras que la varianza en esa ubicación es nula, esto se verificó dentro del software ArcGis, con la herramienta *Geostatistical Analyst - Validation/Prediction*.

La estimación del mapa de los valores estimados por Kriging tiene la propiedad de suavizamiento, es decir, es mucho más suave de lo que sería el mapa de los valores reales (con menos variaciones) (Emery, 2013). Esto es lógico dado que dentro de la estimación no se pueden inventar detalles de la información que no aparece dentro de la observada.

La validación de la interpolación por Kriging se hace de forma cruzada verificando la forma como se adecuan los datos y los parámetros (modelo de variograma, vecindad de kriging), esto se hace estimando sucesivamente mediante Kriging cada uno de los datos, luego de ello se puede calcular el error de estimación, finalmente se puede realizar un análisis estadístico de los errores en todos los sitios; para esta validación se utilizó el software ArcGIS presentando la información de las pruebas de forma gráfica, tal como se muestra a continuación:

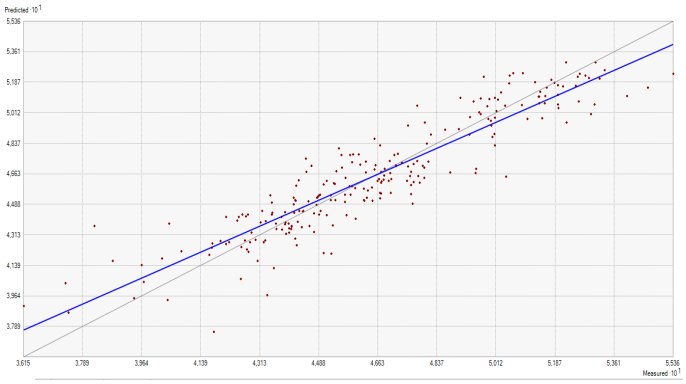


Figura 58: Kriging ordinario valores estimados - Valores medidos

La nube de correlación de los datos medidos y los datos estimados presenta una correlación lineal significativa, sin embargo los valores estimados tienen una pendiente positiva más alta que los observados.

Según Emery (2013) la media de los errores y de los errores estandarizados deben ser cercanos a cero, en este caso estos valores son **-0,0002383** y **-0,00780** respectivamente, por lo tanto se puede suponer que los valores obtenidos son insesgados, adicionalmente su distribución debería ser normal, por lo tanto se generó el gráfico QQ, que permite tener un acercamiento a

esta hipótesis.

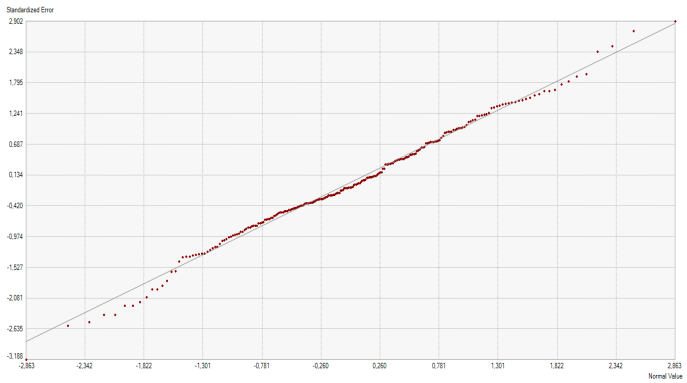


Figura 59: Kriging ordinario gráfico QQ para errores estandarizados

El gráfico 59 permite corroborar la hipótesis de normalidad de los errores, en este caso estandarizados, finalmente el nivel de ajuste de los datos predichos a los valores reales según el coeficiente de correlación de Pearson es **0.920048**, este valor corroborar el ajuste, en general óptimo para la información existente.

7.6.2. Construcción de Isoglosas

El objetivo de la dialectología, como se mencionó atrás, era marcar en la cartografía lingüística isoglosas que delimitaran áreas dialectales, estos limites fueron utilizados de manera regular en la geografía lingüística, convirtiéndose en conceptos recurrentes y genéricos (Chacón, 2016).

El uso de estos mapas lingüisticos pretende formalizar y constatar la distribución geográfica de ciertas variantes lingüísticas (Chacón, 2016), de esta manera la presente investigación integra información espacial basada en las posiciones de las localidades junto con los resultados de las respuestas a los cuestionarios, representada en información léxica, proyectando tales resultados en isoglosas generadas en el software ArcGIS y resultado del proceso de estimación de la superficie dialectal continua.

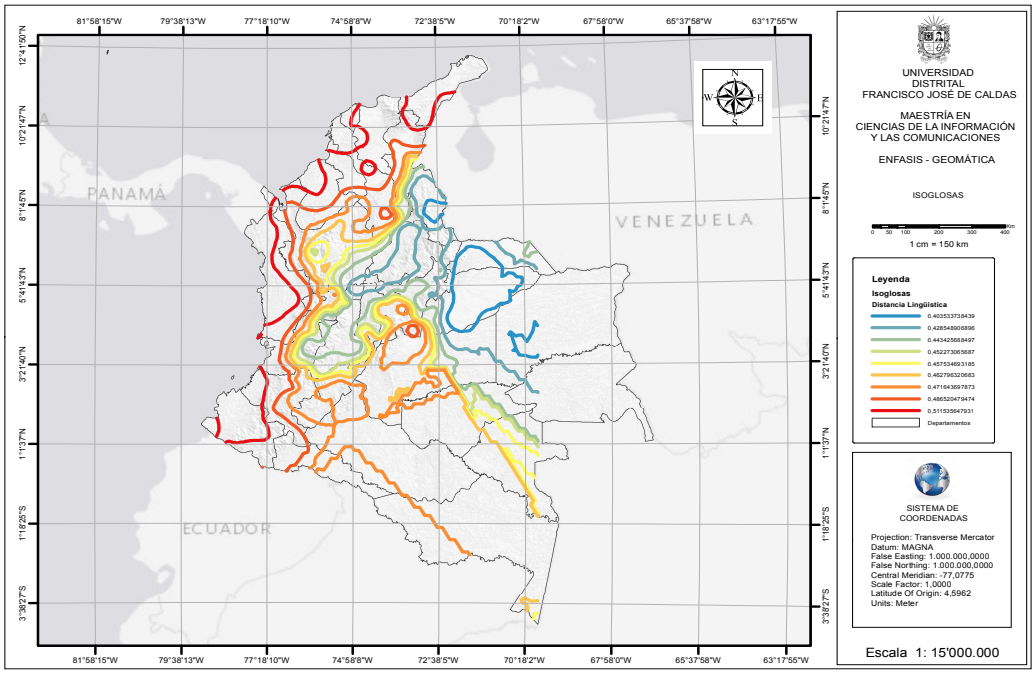


Figura 60: Isoglosas estimadas

Las isoglosas generadas son líneas que conectan geográficamente lugares con una diferencia lingüística similar, delimitando áreas concretas con intervalos definidos, por lo tanto, su generación se hizo a partir del supuesto de continuidad del fenómeno lingüístico, estas líneas en ámbitos geoespaciales son denominadas isolineas, sin embargo, dado que se está realizando la investigación sobre un termino especifico toman el nombre de isoglosas.

La figura 60 muestra las isoglosas resultantes a partir de la diferencia lingüistica estimada con el modelo espacial seleccionado (sección 7.5), la categorización simbólica muestra las lineas de menor diferencia en tonos claros, así mismo, resalta en tono oscuro aquellas zonas con mayor diferencia, en tal caso, se destaca la concentración de la menor diferencia lingüistica estimada en los departamentos del Casanare, Arauca y parte del Vichada con respecto a los demás departamentos, y la mayor diferencia en las regiones que hacen parte de la costa Caribe y del Pacífico Colombiano. Cabe mencionar que hay una isoglosa que se extiende entre ambas diferencias de forma vertical (color naranja), esta isolinea conecta ambos niveles de diferencia, siendo una zona

de transición que atraviesa departamentos como el Cauca, Valle del Cauca, Huila, Caquetá, Risaralda, Antioquia, entre otros, la importancia de tal transición será clave para la conexión de las categorías de clasificación que se definirán en la siguiente sección.

7.7. Clasificación dialectal

Según Montes (1982) la determinación de unidades dialectales, divisiones y subconjuntos que hacen parte de lenguas históricas es un problema que ha sido trabajado de forma constante sin tener aún la capacidad de resolverse satisfactoriamente, en esta sección se intentará delimitar los dominios específicos de determinado rango lingüístico (diferencia lingüistica) en un espacio geográfico, estableciendo como precedente los trabajos de clasificación mencionados atrás y teniendo como elementos base los resultados del modelo de regresión junto con las isoglosas generadas y la superficie dialectal continua estimada con técnicas geoestadísticas, adicionalmente se implementará un algoritmo de agrupamiento espacial (clustering) denominado "*Regionalization with dynamically constrained agglomerative clustering and partitioning*"(REDCAP) que traducido es "*Regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido*", una forma de dividir un conjunto de datos espaciales en una serie de regiones espacialmente contiguas en base a criterios alfanúmericos (distancia lingüística) y geográficos (pesos espaciales) (Guo, 2008).

7.7.1. Regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (REDCAP)

Este método de agrupamiento desarrollado por Guo (2008) hace parte de un conjunto de técnicas propuestas en minería de datos espaciales para la clasificación de información con un componente geográfico, la regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido **permite dividir un gran conjunto de objetos geográficos y optimizar una función objetivo (generalmente de homogeneidad) mejorando la agrupación en cada una de las múltiples iteraciones que realiza**, existen seis métodos con tres enfoques distintos de agrupamiento, un enfoque con un vínculo único, otro con un vínculo promedio (ALK) y finalmente uno con un vínculo completo (CLK), restringidos por contigüidad de primer y segundo orden en cada uno de los casos. En su investigación Guo

(2008) descubrió que los métodos de Full-Order-CLK y Full-Order-ALK son mejores significativamente a otros métodos (k-means, SKATER, entre otros.) en diferentes aspectos, principalmente en la heterogeneidad espacial total, el equilibrio en el tamaño de la región, la variación interna y la conservación de la distribución de la información (Guo, 2008).

La regionalización o agrupamiento es en esencia una forma de clasificación, en la que unidades espaciales se agrupan a partir de criterios o restricciones de contigüidad o adyacencia (Haining, 2003), en esta investigación se utilizarán los pesos espaciales (W) ya implementados en el modelo de retardo espacial, el método de agrupamiento aglomerativo será el agrupamiento de enlace completo (CLK) y su restricción de orden completo Full-OrderCLK, este método es señalado por Guo (2008) como el mejor en comparaciones de calidad con los otros métodos mencionados, con un tiempo de cálculo de $O(n^2 log(n))$.

Según Guo (2008) los tres métodos de agrupamiento difieren en la definición de distancia (disimilitud) entre grupos, el clustering de enlace único (SLK) define la distancia entre dos grupos como la diferencia entre el par más cercano de puntos de datos de cada grupo, este tipo de agrupamiento tiene una tendencia a agrupar puntos vinculados por una serie de puntos de datos intermedios cercanos, este fenómeno se conoce como el "efecto de encadenamiento", que puede resultar en grupos que contienen puntos de datos diferentes.

El agrupamiento de enlaces promedio (ALK) define la distancia entre dos grupos como la disimilitud promedio entre todos los pares de puntos de datos entre grupos, este tipo de agrupamiento depende de la escala numérica en la que se miden las diferencias; si se realiza una transformación monótona estrictamente creciente puede que cambie el resultado.

Finalmente está el método de agrupamiento de enlaces completo (CLK), este utiliza la disimilitud entre el par de puntos de datos más lejano como la distancia entre dos grupos, de la siguiente manera:

$$d_{CLK}(L, M) = max_{u \in L, v \in M}(d_{uv}) \tag{35}$$

Donde L y M son dos grupos, $u \in L$ y $v \in M$ son dos puntos de datos, y d_{uv} es la disimilitud entre u y v.

Para la agrupación por el método de enlace completo (CLK), dos agrupaciones se consideran similares solo si todas las observaciones en las dos agrupaciones son similares entre sí.

A diferencia de las definiciones de distancia los tres métodos utilizan el mismo procedimiento para el agrupamiento, inicialmente, cada objeto de datos individual es un grupo en sí mismo, luego, se van seleccionan los elementos similares (determinados por la definición de distancia) y se fusionan en un grupo, con las restricciones de contigüidad propuestas (W pesos), este proceso se repite hasta que los puntos de datos con mayor coincidencia estén en el mismo grupo.

Según Wang y Robert (2015) la implementación del método REDCAP es eficiente, ya que su construcción desde el primer paso busca determinar grupos espacialmente contiguos con base en la similitud de atributos bajo la restricción de adyacencia, **teniendo como principal objetivo un agrupamiento dinámico y heterogéneo**, razón principal por la que fue seleccionado en esta investigación; la ejecución del método es de forma determinista, produciendo un número de regiones establecidas desde el inicio del proceso, para esta investigación se utilizará el software GeoDA y su paquete de implementación para R (RGeoDa) con la determinación inicial de ocho regiones, verificando la última clasificación propuesta por Bonilla (2019), donde define tres superdialectos andino centro-oriental, costeño y andino centro-occidental, junto con ocho dialectos, todos distribuidos desde la parte norte de Colombia hacia el sur en forma vertical, es de resaltar que los datos lingüísticos utilizados son el producto del modelo de retardo espacial generado en secciones anteriores.

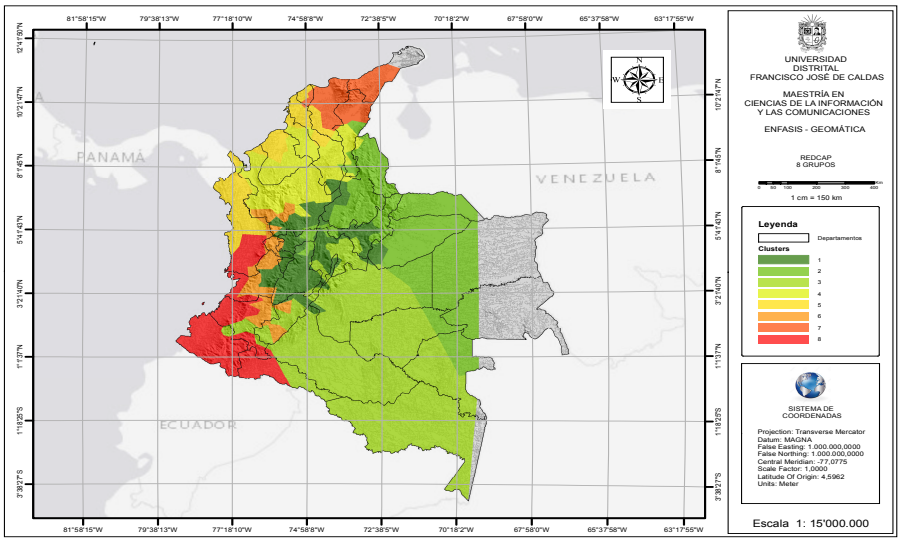


Figura 61: Clasificación método REDCAP (8 Grupos)

La clasificación ilustrada en la figura 61 muestra ocho regiones diferenciadas por color según su similitud lingüística estimada, todas ellas distribuidas verticalmente a lo largo del territorio, la primera se encuentra en la zona central, región Andina de Colombia, en color verde oscuro (número uno en la leyenda), de esta región hacen parte 53 localidades dentro del norte y oeste de Cundinamarca, sur y oeste de Boyacá, sur y oeste de Casanare, todo el Departamento del Tolima, zona central de Caldas, zona este del Valle del Cauca y parte sur del Departamento de Antioquia.

La segunda zona está representada en color verde en un tono más claro en comparación con la región anterior, de ella hacen parte 51 localidades, esta zona esta ubicada entre la región Andina y la Orinoquía dentro de los departamentos de Santander, Norte de Santander, zona oeste de Arauca, zona norte y este de Boyacá, y finalmente el este del departamento del Casanare.

La tercera zona se representa en tono verde claro, de ella hacen parte 30 localidades ubicadas en el centro-sur de Colombia, los departamentos que la conforman son, el sur de Cundinamarca, el oeste del Meta, la parte norte del Caqueta y el sur del Amazonas, hay también, algunas localidades en el sur del Valle.

El cuarto grupo está representado en color verde tenue, de el hacen parte 29 localidades, esta zona se encuentra entre la región caribe y la región andina, en lo que parece ser un área de transición, conectando las zonas de mayor distancia lingüistica con las de menor distancia, esto se menciona en la sección anterior (sección 7.6.2) donde se puede observar una isoglosa en color naranja (figura 60) que cruza verticalmente el país, enlazando áreas de variación lingüistica con alta diferencia; el cuarto grupo en mención ocupa la parte norte del departamento de Antioquia, el sur de Córdoba, Sucre y Bolivar, y la zona media del Cesar. El quinto grupo está representado en color amarillo, una zona que se encuentra entre la región Caribe y la región Pacífica, de este cluster hacen parte el norte del departamento del Chocó, de Sucre y de Bolivar, junto a ellos el sur del Atlántico, las localidades que componen este grupo son un total de 24.

La sexta zona puede denominarse como un área de transición (sección 7.6.2), está representada en color naranja tenue, de ella hacen parte una porción de distintos departamentos, el sur de Antioquia, el oeste de Risaralda y de Caldas, el este del Valle del Cauca y del Cauca, y finalmente el oeste del Huila, de esta región hacen parte un total de 19 localidades. En el séptimo grupo se encuentran la zona norte del Caribe Colombiano, de esta zona hacen parte el departamento de la Guajira, el norte del Cesar y del Magdalena, y el este del Atlántico, esta región esta compuesta por un total de 16 localidades, con una tendencia marcada en el norte del país.

Por último se encuentra el grupo octavo, representado en color rojo, este grupo esta compuesto por 16 localidades, ubicado en el suroeste del país, entre la región pacífica y andina, esta región tiene una distribución continua entre los departamentos del Nariño, el Putumayo, el oeste del Cauca, del Valle del Cauca y el sur del Chocó.

En general esta clasificación se ajusta a las generadas por otros autores, tal es el caso de Montes (1982), Mora et al. (2004), Avila et al. (2015) y Bonilla (2019), prevalece la existencia de zonas de transición entre las diferentes regiones del territorio Colombiano, sin embargo con una pequeña disimilitud en la región caribe, que en esta investigación tiene un área menor a la

señalada por ellos, con un total de 16 localidades; para mejorar esta clasificación, a continuación se realizará una regionalización con un número menor y mayor de grupos, pretendiendo verificar tanto los superdialectos, dialectos, subdialectos y léxicos regionales presentes en la bibliografía expuesta anteriormente.

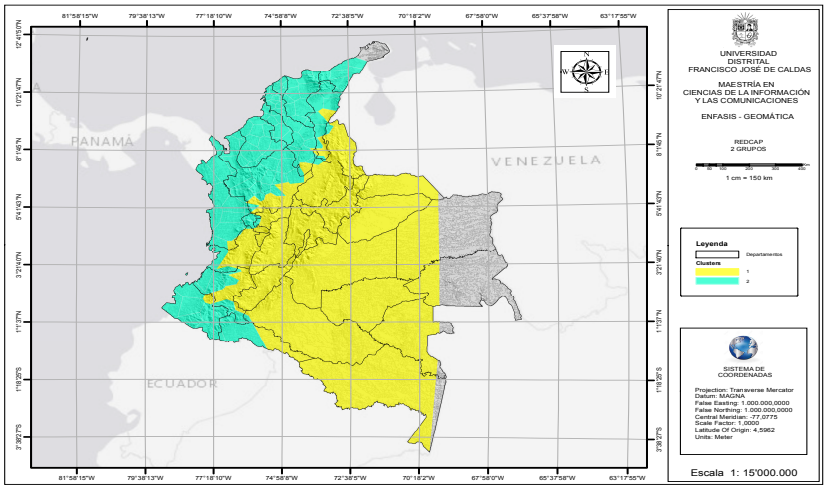


Figura 62: Clasificación método REDCAP (2 grupos)

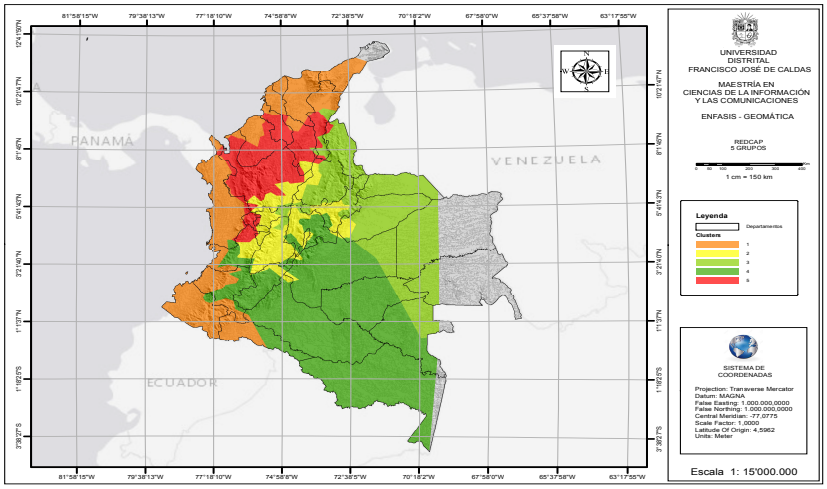


Figura 63: Clasificación método REDCAP (5 grupos)

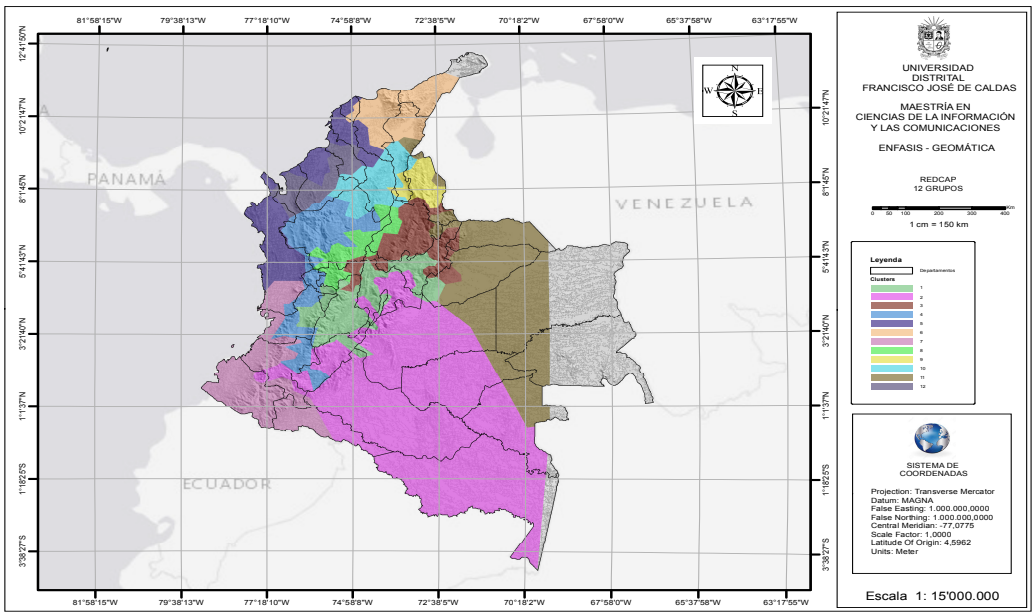


Figura 64: Clasificación método REDCAP (12 grupos)

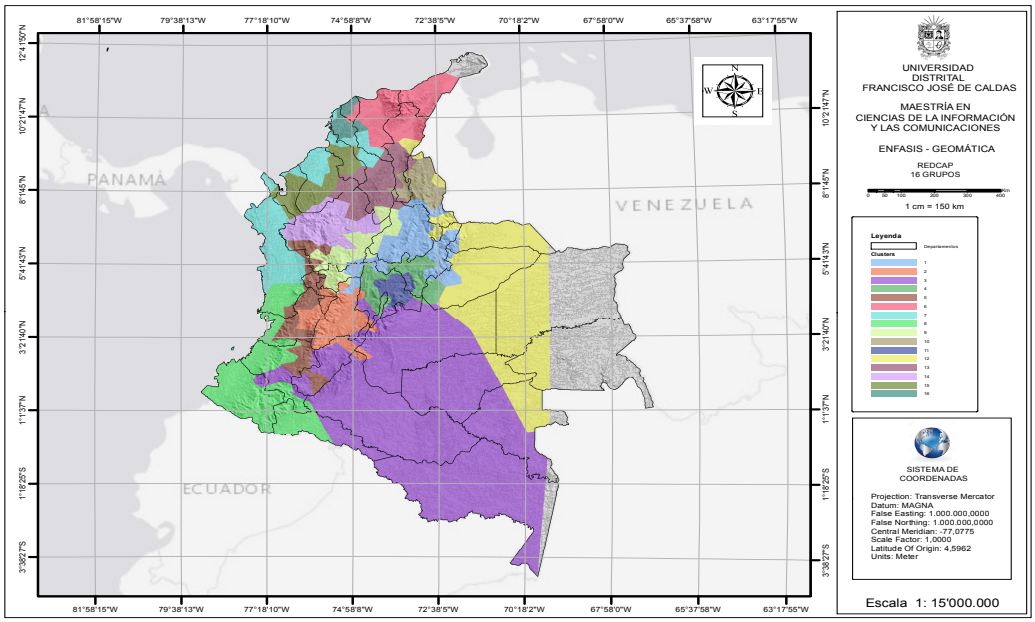


Figura 65: Clasificación método REDCAP (16 grupos)

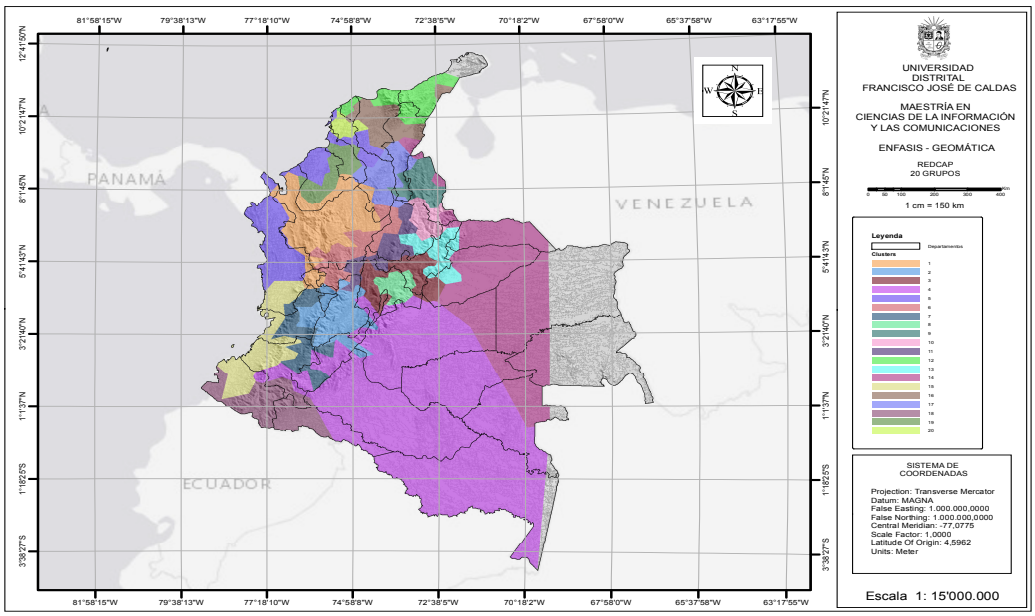


Figura 66: Clasificación método REDCAP (20 grupos)

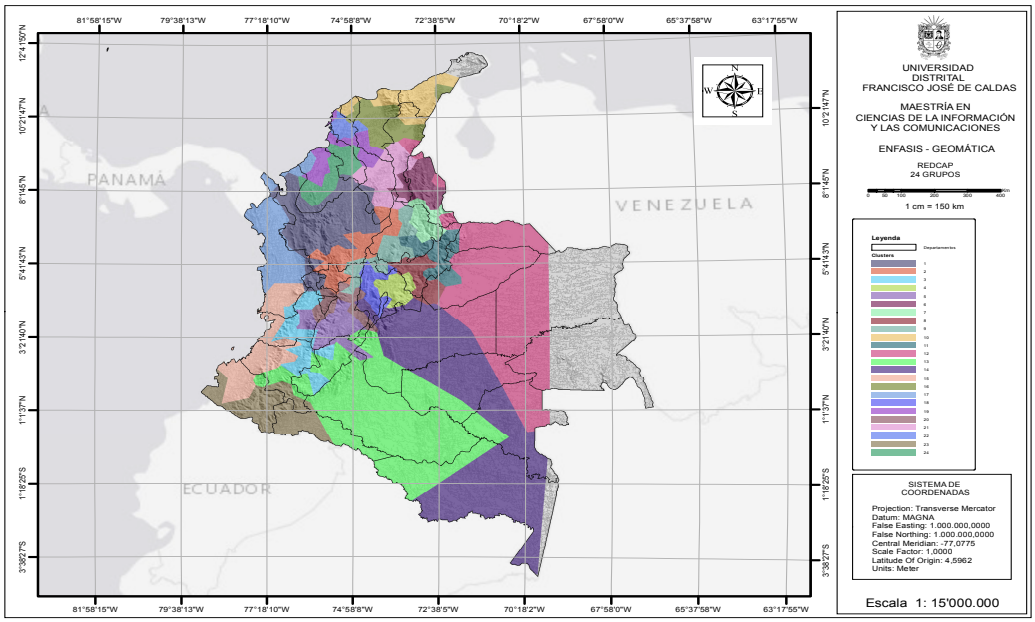


Figura 67: Clasificación método REDCAP (24 grupos)

7.7.2. Propuesta de clasificación dialectal

Teniendo presente las diferentes clasificaciones dialectales realizadas con el algoritmo de regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (REDCAP) se realizará la propuesta de clasificación dialectal, la figura 62 muestra como existe una división entre la región Pacífica junto con la región Caribe en relación con la región Andina muy marcada, con un leve cambio en el sur-oeste del territorio, en el Departamento del Putumayo, esta primera clasificación en dos grupos es muy relevante ya que muestra que las localidades encontradas en el centro oriente del país son en esencia distintas desde una perspectiva lingüística (diferencia lingüística) a aquellas localidades que están ubicadas cerca del Caribe y del Pacífico Colombiano.

La segunda división, representada en la parte inferior de la Figura 62 evidencia que hay cinco zonas con una similitud lingüística relacionada, el primer grupo es conformado por el Caribe y el Pacífico Colombiano, departamentos de La Guajira, Bolívar, Atlántico, Chocó, Nariño y parte occidental del Cauca, la segunda zona se encuentra en la región Andina, departamentos del Quindío, Tolima, Caldas, noroccidente de Cundinamarca y sur de Antioquia; el tercer grupo se encuentra en la zona nororiental del país, entre la región Andina y la región de la Orinoquia, de ella hacen parte departamentos como el Casanare, Arauca, Santander, Norte de Santander y noreste de Boyacá.

Los últimos dos grupos se encuentran ubicados en posiciones espaciales distintas, el primero de ellos está representado en color verde claro en la parte sur del territorio Colombiano, de el hacen parte las localidades ubicadas en los departamentos del Amazonas, Caqueta, Huila, Meta y el este de Cundinamarca, esta clasificación es relevante pues permite observar como la localidad en principio más distante (Leticia - Amazonas) tiene una relación con localidades contiguas ubicadas en otros departamentos; el último grupo se encuentra en el limite entre la region Andina y la región Pacífica señalado en color café, de este grupo hacen parte el departamento de Antioquia, el sur de Sucre, de Bolivar y de Córdoba, y por último el departamento de Risaralda .

En conclusión, a partir de las divisiones anteriores se pueden enunciar dos superzonas con las que se iniciara el desarrollo de la propuesta de clasificación dialectal, la primera, una zona netamente Costeña, que abarca el Pacífico y el Caribe Colombiano; y en segundo lugar, la región nororiental del país, grupo o zona que se denotará como Andina, siguiendo la clasificación de Montes (1982), estos dos grupos a su vez se van a segmentar de la siguiente manera:

Superdialecto	Dialecto	Subdialectos	Localidades
Andino	**Andino Oriental**	**Tolimense**	San Antonio, Pandi, Purificación, Ortega, Cunday, San Luis, Coello, Tocaima, Ibagué, Santa Isabel, Ambalema, El Líbano,Herveo, Mariquita
		Amazónico - Caqueteño	San Vicente del Caguán, Leticia, Florencia
		Huilense - Caqueteño	San Agustín, Timaná, Gigante
		Huilense	Santa María, Neiva, Baraya
		Huilense - Tolimense	Alpujarra, Planadas, Chaparral, Natagaima
		Llanero - Suroriental	Tauramena, Orocué, Puerto López, San Martín, Quetame, Gutiérrez, Restrepo, Maní, Medina
		Llanero - Nororiental	Arauca, Trinidad, Yopal, Pore, San Bernardo de Bata, Tame
		Cundinamarques	Caparrapí, Pasca, Gachalá, Fómeque, Guasca, Bogotá, La Mesa, Facatativá, Tenjo, Guaduas, Pacho, Paime
		Cundinboyacense	Manta, Villapinzón, Tenza, Turmequé, Gachetá, Nemocón, Ubaté, Simijaca, Muzo
		Boyacense	San Luis de Gaceno, Miraflores, Aquitania, Siachoque, Labranzagrande, Firavitoba, Monguí, Sotaquirá, Duitama, Socha, Ráquira, Moniquirá, Saboyá, Jericó, Güicán, Onzaga
		Santandereano	Guaca, Jesús María, Vélez, Cimitarra, Bocas del Rosario, Chitagá, Pamplona, Herrán, Cáchira, Charalá, Suaita, Málaga, Aratoca, San Vicente de Chucurí, Girón, Tona, Rionegro, Suratá, Sabana de Torres, Simacota
		Norte Santandereano	El Tarra, Bochalema, Salazar de las Palmas, Cúcuta,Ricaurte, Puerto Villamizar, Villacaro, Sardinata, Ábrego, Ocaña, San Calixto, El Carmen, Loma de Corredor
	Transición Andino - Costeño Caribe	**Transición Costeña Norte**	Majagual, Pailitas, Chiriguaná, Gamarra, Simití, San Martín de Loba, La Jagua de Ibirico
Costeño	**Costeño Caribe**	**Cesarense-Guajiro**	Riohacha, Valledupar, Fonseca, Carraipía, Uribia, Barbacoas, Palomino
		Bolivarense - Cesarense	Pivijay, Caracolí, El Difícil, Aracataca, Codazzi
		Magdalenense-Guajiro	Taganga, Ciénaga, Sitionuevo, Puerto Colombia
		Bolivarense - Atlanticense	La Boquilla, Turbaco, San Juan Nepomuceno, María La Baja, San Basilio de Palenque, Mahates, Villanueva, Manatí, Luruaco
		Bolivarense - Magdalenense	Mompós, Plato, Pedraza, El Carmen
		Bolivarense-Sucreño	San Benito Abad, Corozal, San Pedro, Sahagún, Tolú, San Onofre
		Cordobés	Tres Palmas, Planeta Rica, Cereté, Puerto Escondido, San Antero
	Costeño Pacífico	**Pacífico Norte**	Coredó, Acandí, Arusí, Cértegui, Tutunendo, Bahía Solano (o Ciudad Mutis), Riosucio
		Pacífico Medio	Timbiquí, San Juan de Mechengue, Buenaventura, Morales, Robles, Nóvita
		Pacífico Sur	Barbacoas, Tumaco, Iscuandé

Superdialecto	Dialecto	Subdialectos	Localidades
Costeño	**Transición Pacífico - Andino**	**Nariñense**	Potosí, Mocoa, La Cruz, Bolívar, La Florida, Ricaurte
		Caucano	Nátaga, La Sierra, Popayán, Silvia, Patía (El Bordo), Santander de Quilichao, Toro, Cartago
		Valluno	Palmira, Florida, Cali, Ginebra, Dagua, Buga, Darién (o Calima), Andalucía, Sevilla, Roldanillo
		Antioqueño Cordobés	Chigorodó, Turbo, Tierra Alta
Andino	**Andino Occidental**	**Eje cafetero**	Manzanares, Samaná, Calarcá, Manizales, Aranzazu, Pereira
		Antioqueño - Risaraldense	Santuario, Mistrató, Betania, Urrao
		Antioqueño - Caldense	Sonsón, Pácora, Jardín, Montebello, Concordia
		Antioqueño Occidental	Remedios, Dabeiba, Heliconia, San Pedro, Santa Fe de Antioquia, Yarumal, Frontino, Ituango, Santo Domingo, Amalfi, Zaragoza, Cáceres, Caucasia, Nechí, Ayapel
		Antioqueño Oriental	Cocorná, Rionegro, Barrancabermeja, Puerto Wilches, Puerto Berrío, Yolombó

Cuadro 42: Clasificación dialectal propuesta

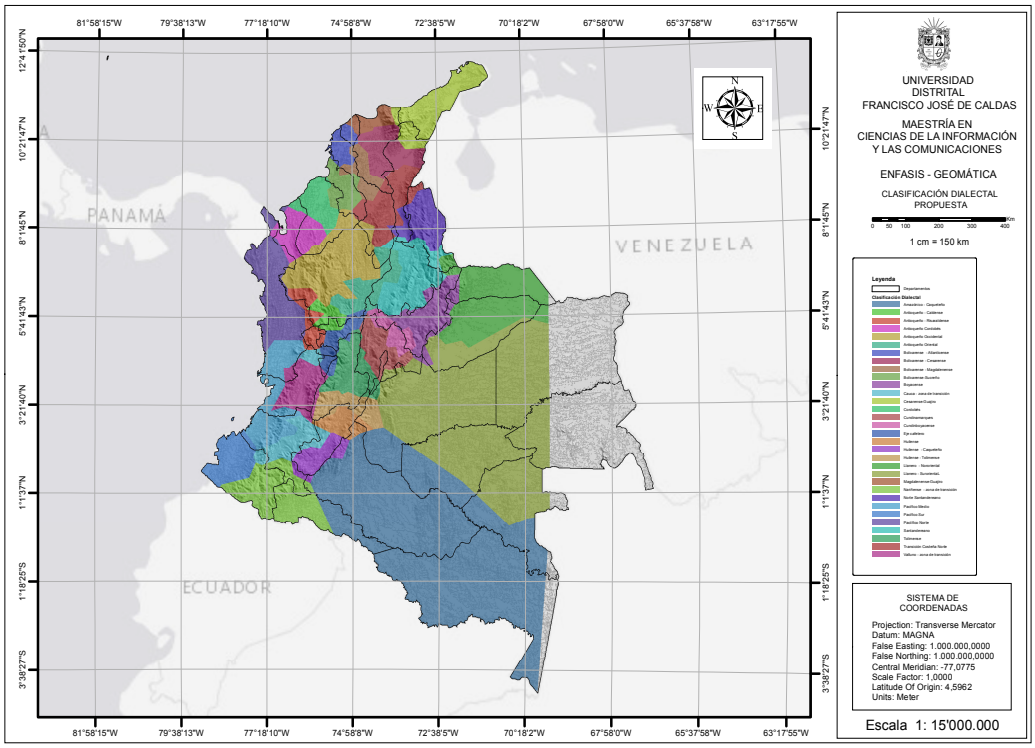


Figura 68: Clasificación dialectal propuesta

La propuesta de clasificación dialectal se enmarca dentro de los valores

producto del modelo de retardo espacial seleccionado en secciones anteriores, modelo en el cual se incluyen distintos fenómenos de tipo geográfico como variables que intervienen dentro de la variación de la lengua en el territorio, al igual que otras propuestas (Bonilla (2019), Montes (1982), Mora et al. (2004), Avila et al. (2015)) se incorporaron tres divisiones: superdialectos, dialectos y subdialectos, resaltando que los resultados se limitan exclusivamente al análisis léxico del ALEC (cien mapas).

Los superdialectos seleccionados surgieron del algoritmo de regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (Figura 62), esta división corresponde a dos grupos en esencia distintos que se distribuyen en la región Andina (en su mayor parte) y la zona pácifica y caribe, denominado en la bibliografía como Costeño, estos son representados de forma notoria, coincidiendo y verificando las propuestas de otros autores (Montes (1982), Mora et al. (2004), Avila et al. (2015)); cabe mencionar que existen zonas intermedias que se encuentran entre ambas divisiones, sin embargo estas son abarcadas en los siguientes niveles de la lengua propuestos.

Los dialectos propuestos después del análisis de clasificación en total son ocho, con similitud a algunas de las propuestas de autores como Montes (1982) y Mora et al. (2004) (4, 5 y 8 dialectos) pero con un número diferente agrupaciones, Montes (1982) expone cuatro zonas sin tener en cuenta algunas áreas de transición, sin embargo existe una región que se propaga de forma vertical entre la región Andina y Costeña, específicamente en sur del país, departamento del Nariño, hacia el departamento del Cauca y finalizando en el Valle del Cauca, por el costado Oriental se encuentra otra zona de transición, por los departamentos de Bolívar y Cesar, estas dos áreas de transición separan los dos superdialectos propuestos, de tal manera que existe coherencia y una gran similitud a las propuestas de otros autores; Bonilla (2019) es la propuesta que mayor se asemeja a esta clasificación, donde también se representan estas zonas, la figura 69 ilustra los dialectos propuestos, es de resaltar que la concentración de la información se encuentra en el norte del país, dado que existe ausencia de información en las localidades ubicadas en la zona sur, existe allí una mayor heterogeneidad espacial, es decir una menor variación.

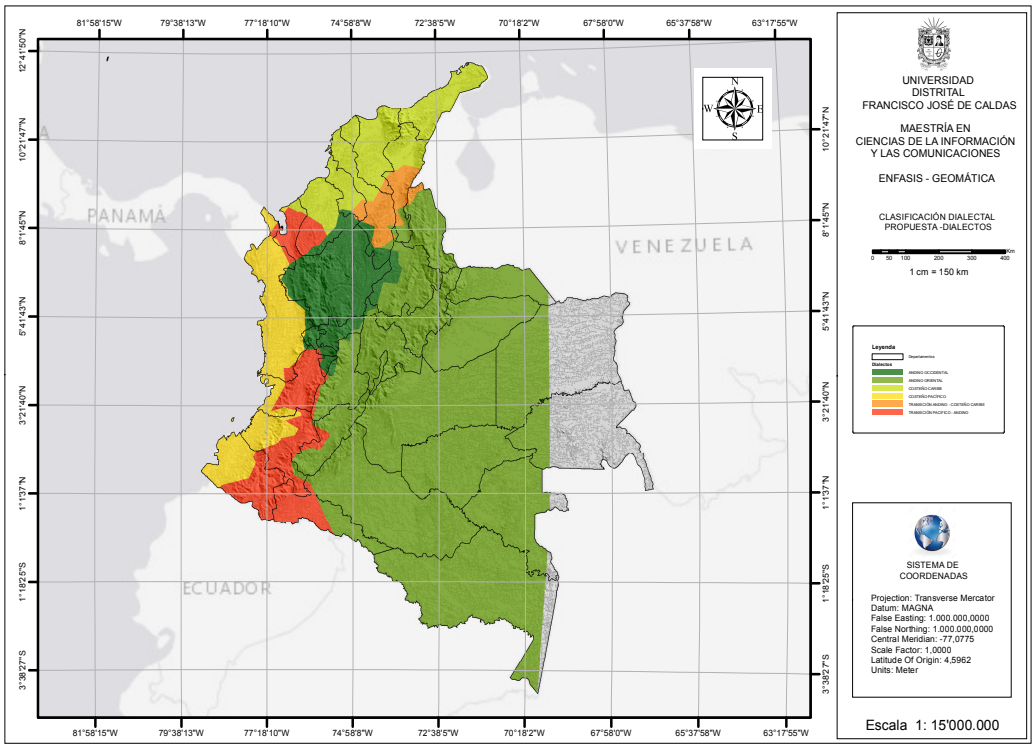


Figura 69: Clasificación dialectal propuesta - dialectos

Como subdialectos se propone la subdivisión del Andino Oriental en doce divisiones, que recogen principalmente la información referida a la distribución departamental, con transiciones entre diferentes unidades espaciales, en total son 112 localidades que hacen parte de estas subdivisiones dialectales, destacando la presencia del denominado Amazónico - Caqueteño, donde se encuentran localidades como San Vicente del Caguán, Leticia y Florencia; otra división importante es aquella que recoge la transición norte entre los superdialectos Andino y el Costeño Caribe, de ella hacen parte localidades como Majagual, Pailitas, Chiriguaná, Gamarra, Simití, San Martín de Loba, La Jagua de Ibirico, localidades ubicadas en los departamentos de Bolívar y Cesar.

Por otro lado se tiene el costeño caribe y el costeño pacífico, del primero hacen parte 40 localidades ubicadas en los departamentos del Bolívar, Córdoba, Magdalena, Cesar, Sucre, Atlántico y La Guajira, donde las divisiones

propuestas se encuentran entre los distintos departamentos, una de ellas es el Bolivarense - Magdalenense, de ella hacen parte localidades como Mompós, Plato, Pedraza y El Carmen, esta zona está representada claramente en la Figura 68, mostrando la variación de la lengua propuesta en esta investigación.

Ahora bien, se tomó el costeño pacífico como una zona dividida de sur a norte en tres regiones subdialectales, región norte, con siete localidades, Coredó, Acandí, Arusí, Cértegui, Tutunendo, Bahía Solano y Riosucio ubicadas en el Norte del Chocó, región media con localidades como Timbiquí, San Juan de Mechengue, Buenaventura, Morales, Robles y Nóvita, ubicadas en el sur del Chocó, occidente del Cauca y Valle del Cauca, y finalmente la región sur, con las localidades de Barbacoas, Tumaco e Iscuandé, en el occidente del departamento del Nariño, en total son 16 localidades, que se extienden a lo largo de la región pacífica y que desde el inicio del desarrollo de la clasificación se mostraban con diferencias notables con respecto a localidades ubicadas en la zona central del territorio.

Otra zona de transición propuesta fue evidente durante el proceso de clasificación, de ella hacen parte 27 localidades y 4 subdialectos denominados Nariñense, Caucano y Valluno, ellos en el suroccidente del territorio, el último en el occidente, entre la región caribe y la región pacífica, denominado Antioqueño Cordobés compuesto únicamente por tres localidades, Chigorodó, Turbo y Tierra Alta, región de transición norte-sur del costeño colombiano. Por último se propuso la división del andino en zona occidental, este dialecto está compuesto por 36 localidades y 5 subdialectos, Eje cafetero, Antioqueño - Risaraldense, Antioqueño - Caldense y Antioqueño Occidental y Oriental; la subdivisión dialectal total propone 32 clases, su ilustración se puede observar en la Figura 68 donde se evidencia el comportamiento de la lengua dentro de los departamentos.

Capítulo III
Resultados

8. Resultados

El resultado principal de esta investigación pone de manifiesto como la hipótesis y el planteamiento de autores como Córdoba (2011) coinciden, la variación de la lengua se puede estudiar y analizar desde una perspectiva dialectológica y con la ayuda de métodos proporcionados por la geográfica lingüistica o técnicas espaciales, donde en este sentido, la dimensión geográfica tiene un espacio especialmente relevante, el enfoque de esta investigación, un análisis espacial cuantitativo y estadístico de información lingüistica y geográfica permitió desarrollar distintos modelos para la delimitación de isoglosas y clasificación dialectal, incluyendo variables geográficas continuas y metadatos del ALEC demostrando que hay una relación cuantitativa significativa entre datos geográficos y léxicos.

8.1. Similitud lingüistica - IRI

Entre los resultados generados se encuentra el IRI (Índice Relativo de Identidad), medida de similitud lingüística propuesta por Goebl (1987), para su cálculo se realizó el análisis de los mapas léxicos del tomo III del ALEC dentro de las localidades cartografiadas, seleccionando con criterios de nulidad aquellos 100 mapas óptimos para el análisis dialectométrico propuesto, es decir, fueron escogidos aquellos mapas con mayor número de datos, los resultados del IRI se encuentran en una matriz cuadrada que relaciona localidad a localidad la diferencia lingüística, esta matriz fue ordenada verticalmente, escogiendo el nivel medio por cada punto geográfico de tal manera que cada localidad tuviese una medida única de similitud (Cuadro 43), es de resaltar que la distribución de la similitud lingüística y su media (ilustrados en la Figura 25) comparten una distribución normal, verificada con los test de Shapiro-Wilk y K^2 de D'Agostino y visualizada dentro del territorio en la

Figura 27.

Localidad	Similitud Lingüística
Acandí	0.5114845791875694
Riosucio	0.5484902070044664
Bahía Solano (o Ciudad Mutis)	0.5850693129212045
Coredó	0.5211848900010854

Cuadro 43: Similitud Lingüística media por localidad

Algunos estadísticos resumen para la similitud lingüística por localidad son los siguientes:

Localidad	Estadístico	Valor
Trinidad (Casanare)	Mínimo	0,348222
San Basilio de Palenque (Bolívar)	Máximo	0,617272
	Media	0,465078
	Suma	110,688665
	Desviación Estándar	0,048371

Cuadro 44: Estadísticos resumen de similitud lingüistica media

8.2. Distancias Geográficas

La distancia geográfica calculada para las ubicaciones espaciales de las localidades se hizo siguiendo la matriz de similitud lingüistica, como resultado se obtuvo la relación de distancia entre cada una de las ellas expresada en kilómetros (Km), la distancia para esta investigación representa una medida de longitud entre dos puntos sobre una superficie de revolución denominada elipsoide (WGS84), cuyo cálculo se realizó en Python con uso de la librería Geopy y la formula de Vincenty, para tener una medida única para cada localidad se tomó la media, esta puede visualizarse en la figura 28.

Según los estadísticos de resumen la localidad con menor distancia respecto a las otras es Sardinata en el departamento del Norte de Santander, con una longitud media de $224,79891\ Km$, seguida de Salazar de las Palmas y Villacaro ubicadas en el mismo departamento, la localidad con mayor longitud es Leticia en el departamento del Amazonas con una distancia media

aproximada de $1278,70889\ Km$ seguida de Orocué en el departamento del Casanare, en general la media es de $385,781878\ Km$ y la desviación estándar de $132,83443\ Km$.

8.3. Correlación espacial

Otro resultado interesante es el valor de dependencia espacial entre variables geográficas y lingüísticas que de forma global mostraron la relación espacial de los datos, para ello se utilizó el Índice de Moran, una medida para el diagnostico de correlación geográfica (Bohorquez y Ceballos, 2008), entre los resultados más llamativos se encuentran los valores adoptados por las actividades económicas, con autocorrelación positiva para la Minería con $I = 0{,}057$ y $p - valor = 0{,}054$, es interesante dado que toma un valor significativo cuando se introduce como variable respuesta dentro del modelo espacial seleccionado; al calcular el I de Moran bivariante, entre la distancia lingüística y la Minería el resultado fue $I = 0{,}003$ con un $p - valor = 0{,}4460$, es decir un nivel de significancia muy bajo o nulo.

La aptitud agroclimática es más coherente con el modelo seleccionado, en ella las categorías, extremadamente seco, ligeramente húmedo y moderadamente húmedo tienen un valor significativo, donde el I de Moran univariante y Bivariante toman los siguientes valores:

Aptitud Agroclimática	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Ligeramente húmedo	0.105	0.006	-0.113	0.001
Moderadamente húmedo	0.260	0.001	0.018	0.245
Extremadamente seco	0.659	0.001	0.063	0.015

Cuadro 45: I de Moran para Aptitud Agroclimática univariante y bivariante

Entre otras variables se encuentran la precipitación, donde los rangos entre 500 y 1000, y entre 2000 y 2500, fueron los más significativos dentro del modelo seleccionado, en ellos la autocorrelación espacial univariante y bivariante fue la siguiente:

Precipitación Rango	I de Moran Global	p-valor	I de Moran Bivariante	p-valor
Entre 500 y 1000	0.492	-0.128	-0.038	0.078
Entre 2000 y 2500	0.118	-0.013	-0.091	0.002

Cuadro 46: I de Moran univariante y bivariante - Precipitación

La temperatura tiene una alta autocorrelación, sin embargo dentro del modelo su significancia parece ser nula, no obstante las categorías que se incluyeron corresponden al rango de, entre 20 y 22 grados con un $I = 0{,}090$ y un $p{-}valor = 0{,}018$, y entre 26 y 28 grados con $I = 0{,}510$ y un $p{-}valor = 0{,}001$, cabe destacar que el I bivariante respectivamente es $I = -0{,}058$ con un $p - valor = 0{,}025$ y $I = 0{,}242$ con un $p - valor = 0{,}001$.

En el caso del Índice de disponibilidad hídrica las categorías incluidas dentro del modelo seleccionado son semihúmedo y semiseco, la de mayor autocorrelación y significancia es semiseco con un $I = 0{,}444$ y un $p - valor = 0{,}001$, en segundo lugar la categoría semihúmedo con un $I = 0{,}262$ y un $p - valor = 0{,}001$.

Por último, se presenta una de las variables tomadas a partir de los metadatos del ALEC, las vías de acceso, los resultados obtenidos para esta variable verifican la hipótesis realizada por Bonilla (2019) acerca de su relación intrínseca con la distancia lingüística, entre los valores de autocorrelación espacial se destacan los siguientes:

Vías de Acceso	**I de Moran Global**	**p-valor**	**I de Moran Bivariante**	**p-valor**
Via Camino de herradura	0.065	0.090	0.074	0.006
Via Carretera y Férrea	0.084	0.024	0.023	0.1820
Via Carretera	0.147	0.003	-0.041	0.070
Caminos carreteables	0.420	0.001	-0.206	0.001

Cuadro 47: I de Moran para las Vías de Acceso - Resultados

Dentro de las vías de acceso la categoría correspondiente a Via Carretera y Ferrea es la de menor significancia con respecto al Indice de Moran bivariante, sin embargo, dentro del modelo seleccionado tiene una alta relevancia, haciendo que aumente sustancialmente el valor de la distancia lingüística (mucho más que la distancia geográfica con $I = 0{,}798$ y un $p{-}valor = 0{,}001$), de esta manera los resultados señalan como algunas variables geográficas y alguna externas (vías de acceso, actividades económicas) tienen una notable importancia, así como una correlación espacial inherente a partir de la cual se estimaron algunos modelos espaciales, el resultado de este procedimiento se presenta a continuación.

8.4. Modelo estimado

Los modelos generados siguen la teoría propuesta en econometría espacial por Anselin (1988) donde las causas de la autocorrelación espacial (la distancia lingüística) son producto de la existencia de errores en las observaciones, o en este caso, de la interacción de fenómenos espaciales en unidades geográficas contiguas, fenómenos que se integraron dentro de modelos geográficos evidenciando la existencia de interdependencia y multidireccionalidad en los datos espaciales (Pérez, 2006). Estos modelos estimados siguieron la metodología propuesta por Anselin (2005) ilustrada en la Figura 37 de tal manera que inicialmente se generó el modelo lineal por mínimos cuadrados ordinarios (MCO) cuyas características más notables son las siguientes.

8.4.1. Modelo Lineal (Estimado por MCO)

El modelo lineal estimado por mínimos cuadrados ordinarios está compuesto por todas las variables objeto de análisis, tanto las geográficas como aquellas que hacen parte de los metadatos del ALEC, el resultado de este modelo hace posible tener una perspectiva inicial del comportamiento de todas las variables en relación con la distancia lingüistica, algunos de los datos más importantes se presentan a continuación.

Variable	Estimate	Std.Error	t value	Pr(>—t—)	Significancia
Intercept	0.2780479	0.0898982	3.093	0.00227	**
I(log(DistanciaG))	0.0507104	0.0113067	4.485	1.24e-05	***
Amenaza Sísmica					
Extremadamente Bajo	-0.1096399	0.0513014	-2.137	0.03383	*
Moderadamente Alto	0.0158319	0.0079435	1.993	0.04765	*
Temperatura					
Entre 24 y 26 grados	-0.0366812	0.0146446	-2.505	0.01307	*
IDH					
Muy seco	0.0423385	0.0150355	2.816	0.00536	**
Vías					
Caminos carreteables	-0.1450533	0.0247028	-5.872	1.83e-08	***
Carretera	-0.0445338	0.0163446	-2.725	0.00702	**
Carretera y Férrea	-0.0439274	0.0208021	-2.112	0.03599	*
Férrea	-0.0819993	0.0432531	-1.896	0.05947	.
Residual standard error: 0.03683 on 195 degrees of freedom **Multiple R-squared: 0.525,** **Adjusted R-squared: 0.4227** **F-statistic: 5.132 on 42 and 195 DF,** **p-value: 1.726e-15**					

Cuadro 48: Modelo lineal (estimado por MCO) variables mayor significancia

Las variables con mayor significancia hacen parte de las Vías de acceso, una constante dentro de la investigación, dentro de ellas la categoría Caminos Carreteables es la de mayor significancia, seguida de la categoría Carretera,

el modelo lineal plantea que cuando existen caminos carreteables la distancia lingüística disminuye en 0,14, esto es, una mayor similitud y por lo tanto una menor variación, lo que no es coherente, teniendo presente que estás vías implicaron una mayor dificultad de acceso a las localidades en el momento de la realización de encuestas y por lo tanto la hipótesis de disminución de propagación de la lengua de un territorio con estas condiciones; este modelo adicionalmente confirma la teoría que expresa, a mayor distancia existe una mayor variación, con un alto grado de significancia, **manifestando que la distancia lingüística aumenta en 0.05 por cada unidad del logaritmo de distancia (Km)**, este modelo inicial fue fundamentales para la creación de los modelos espaciales siguientes debido al componente geográfico esencial no incluido en el.

Contrastes de dependencia espacial: estos contrastes permitieron detectar y analizar el tipo de dependencia espacial dentro del modelo estimado por mínimos cuadrados ordinarios, la hipótesis nula para cada test indicaba que el tipo de dependencia era nula junto a ella una hipótesis alternativa que en contrapunto señala la existencia de la misma (Anselin, 2005), los resultados de los test en orden de significancia fueron los siguientes:

Prueba	Valor	Grados de Libertad	p-valor
SARMA	60.876	2	6.04e-14
LM - LAG	49.895	1	1.622e-12
LM - LE	43.379	1	4.509e-11
LM - EL	10.98	1	0.0009209
LM - ERR	17.496	1	2.879e-05

Cuadro 49: Resultados contraste de dependencia espacial

El test de mayor significancia es el SARMA, este incorpora el rezago espacial y el error espacial en un solo modelo, es de resaltar que este fue el último que se realizó, siguiendo la metodología propuesta de Anselin Anselin (2005), en segundo lugar están los test robustos de rezago y error espacial, con un nivel de significancia similares, por lo tanto, era posible que estos tres modelos se ajustaran a los datos de similitud lingüística, de tal manera que se realizaron los tres modelos y se analizaron para seleccionar el mejor de ellos, el modelo final seleccionado se presenta a continuación.

8.4.2. Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial

Este modelo incorporó la influencia de las variables geográficas y los metadatos del ALEC a través de la variable dependiente (distancia lingüística media) espacialmente retardada, es decir a partir de los valores que tomaban las variables independientes en localizaciones vecinas, el modelo de retardo espacial fue seleccionado dado que cumple con todos los criterios estadísticos en los términos estimados y en los términos del error, en el se incluyen las variables con mayor significancia (VMS); las características más importantes de este modelo se obtienen del Cuadro 50 y son las siguientes.

Variable	Estimate	Std.Error	z value	Pr(>—z—)
Intercepto	-0,0041025	0,0490852	-0,0836	0,9333909
Distancia Geográfica				
log(Distancia Geográfica)	0,0236311	0,0082084	2,8789	0,0039908
Actividades Económicas				
Minería	0,025458	0,012917	1,9709	0,0487367
Aptitud Agroclimática				
Extremadamente seco	0,0091526	0,0052458	1,7447	0,0810289
Ligeramente húmedo	-0,0068969	0,0063624	-1,084	0,2783598
Moderadamente húmedo	0,0259769	0,0125158	2,0755	0,0379371
Precipitación				
Entre 2000 y 2500	-0,0110223	0,0062273	-1,77	0,0767298
Entre 500 y 1000	0,0064258	0,0077946	0,8244	0,4097157
Amenaza Sísmica				
Moderadamente Alto	0,0068347	0,0061211	1,1166	0,2641779
Temperatura				
Entre 20 y 22 grados	-0,0058762	0,0098401	-0,5972	0,5503967
Entre 26 y 28 grados	0,0021984	0,0050846	0,4324	0,6654811
IDH				
Semihúmedo	-0,0203181	0,0108448	-1,8735	0,0609942
Semiseco	-0,0086323	0,0050384	-1,7133	0,0866584
Vías				
Camino de herradura	0,0711757	0,0202489	3,5150000	0,0004397
Caminos carreteables	0,0413851	0,0130986	3,1595000	0,0015804
Carretera	0,0443168	0,0165841	2,6722000	0,0075345
Carretera y Férrea	0,0683297	0,0162778	4,1977000	0,0000270

Cuadro 50: Modelo de retardo espacial para variables significativas

El intercepto, un valor que muestra el valor de la variable dependiente cuando las variables independientes son cero es equivalente a $-0,0041$, un valor cercano a cero, lo cual es congruente con los datos de la investigación; en el caso de la distancia lingüistica, **esta aumenta en 0,0236 por cada unidad del logaritmo de la distancia (Km)**, lo que verifica de forma reiterada la hipótesis de relación proporcional entre distancia lingüística y distancia geográfica, en cuanto a las actividades económicas la minería es la más importante, el modelo manifiesta que si hay una localidad que cumpla con esta característica la distancia lingüística en ella (con influencia en localidades vecinas) aumentara en un $0,025458$, existen variables adicionales

referentes a la Aptitud Agroclimática, la Precipitación, la Amenaza Sísmica, la temperatura, el índice de disponibilidad hídrica y las vías, cuyos resultados se muestran en la tabla 36 y que serán analizados de una forma más profunda más adelante.

Por ahora se hablará acerca de los test realizados a este modelo, las principales localidades que intervienen dentro de los residuos y la estimación final, base para su selección (sobre los otros modelos estimados), para la delimitación de isoglosas, así como para la clasificación dialectal propuesta; el test primordial de verificación de ajuste es el coeficiente de determinación de Nagelkerke (R^2) (Nagelkerke, 1991), una medida de bondad o éxito de la estimación, el cual fue 0,58062, es decir, en un 58 % el modelo se ajusta a la distancia lingüística presente en cada localidad, un valor alto en consideración de un gran número de variables, junto a el se encuentra ρ, 0,61495, coeficiente que midió la dependencia espacial intriseca de los datos, de tal manera que desde la estimación inicial se obtuvo una bondad de ajuste y una dependencia espacial aceptables al propósito de la investigación.

Al analizar los términos del error del modelo los contrastes de normalidad de Shapiro-Wilk, Anderson-Darling y la prueba D'Agostino's K^2 no rechazaron la hipótesis nula de normalidad y por lo tanto fue pertinente concluir que los residuos probablemente tuviesen una distribución normal, por otro lado se realizó la prueba para la detección de heteroscedasticidad espacial de Breusch-Pagan cuyo estadístico calculado fue 25,583 con un p-valor de 0,06018, la hipótesis nula para este test plantea que existe homocedasticidad, al ser el valor de α mayor a 0,05 ($\alpha > 0{,}05$) se aceptó la hipótesis nula, por lo tanto se determinó concluir la existencia de homocedasticidad en este modelo.

Las localidades con mayor error presentado para este modelo se pueden ver en la Figura 46, estas se encuentran principalmente en las zonas donde hay una baja densidad de observaciones, con intervalos $(-0{,}08, -0{,}06)$ y $(0,06 - 0,09)$, las localidades en total son 14, representando un 5 % aproximadamente del total, en general se encuentran distribuidas a lo largo del territorio y dentro de los siguientes departamentos:

Número	Localidad	Departamento
1	Carraipía	La Guajira
2	**San Basilio de Palenque**	Bolivar
3	Pailitas	Cesar
4	Arauca	Arauca
5	Jericó	Boyacá
6	Güicán	Boyacá
7	Tauramena	Casanare
8	Yopal	Casanare
9	Bahía Solano	Chocó
10	Tutunendo	Chocó
11	Novita	Chocó
12	Buenaventura	Valle del Cauca
13	San Juan de Mechengue	Cauca
14	Aranzazu	Caldas

Cuadro 51: Localidades con mayor error en la diferencia Valor estimado - Valor Real

La mayor parte de estas localidades se encuentran ubicadas en los límites del territorio nacional y dentro de diez Departamentos, así mismo es importante destacar que la localidad San Basilio de Palenque en la investigación realizada por Avila et al. (2015) se diferencia también de las demás localidades, para este desarrollo la causa partícular radica en el error de la estimación, sin embargo es menester de próximas investigaciones ahondar en el porque de este fenómeno, finalmente es de resaltar que los términos estimados de distancia lingüistica se acercaron un poco más a los valores reales para las localidades restantes, por lo que este fue el modelo seleccionado, cumpliendo con todos los criterios estadísticos de estimación señalados.

Una vez seleccionado el modelo, este se uso para la generación de los límites lingüísticos, teniendo presente los dos paradigmas claves de la dialectología, una, que plantea que el lenguaje se representa en forma de isoglosas (líneas imaginarias que separan áreas dialectales concretas) y otra, a través de superficies continuas (Jeszenszky y Weibel, 2015), que se distribuyen de forma gradual en el territorio, el resultado de este proceso se muestra en la siguiente sección.

8.5. Limites Lingüísticos

8.5.1. Superficie dialectal continua

La superficie dialectal continua fue generada utilizando el método geoestadístico de interpolación denominado Kriging (ordinario), herramienta que utiliza el análisis de semivarianza para la construcción de una estimación acertada, en este caso el semivariograma verificó nuevamente el criterio bajo el cual a menor distancia entre las localidades mayor similitud espacial entre las observaciones (Giraldo, 2015), el resultado del uso de esta herramienta es la superficie ilustrada en la Figura 70.

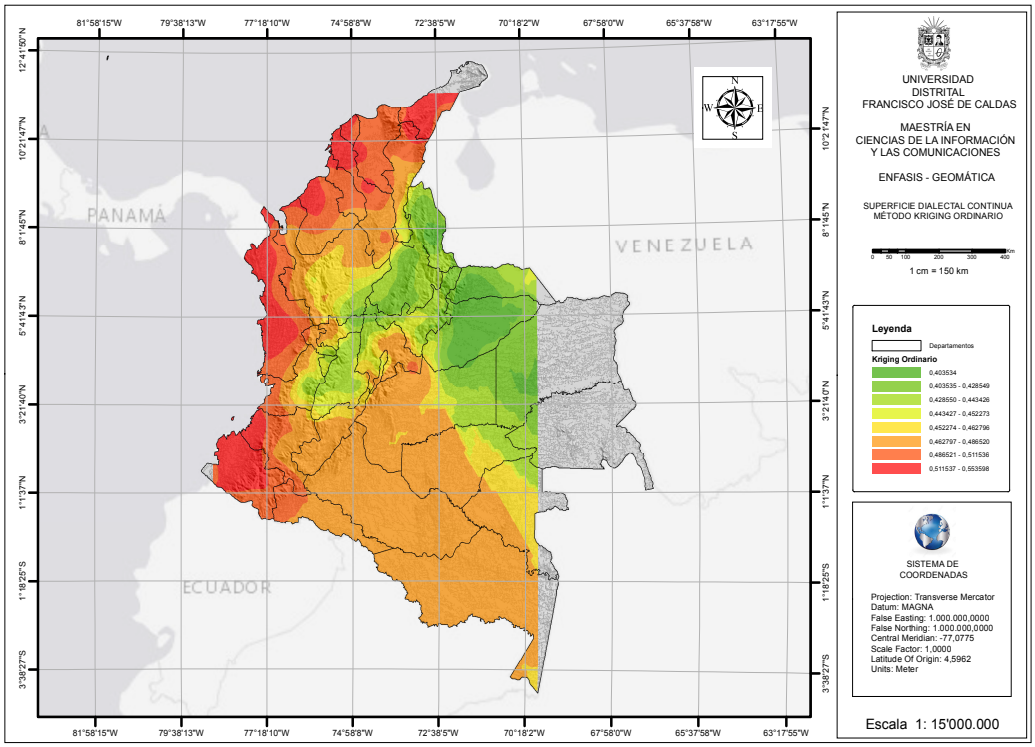


Figura 70: Resultado Superficie dialectal continua

8.5.2. Isoglosas

Las isoglosas generadas son líneas que conectan geográficamente lugares con una diferencia lingüística equivalente, delimitando áreas concretas con

intervalos definidos, estas líneas en ámbitos geoespaciales son denominadas isolineas, sin embargo, dado que se está realizando la investigación sobre un termino especifico toman el nombre de isoglosas, la ilustración de estas es la siguiente:

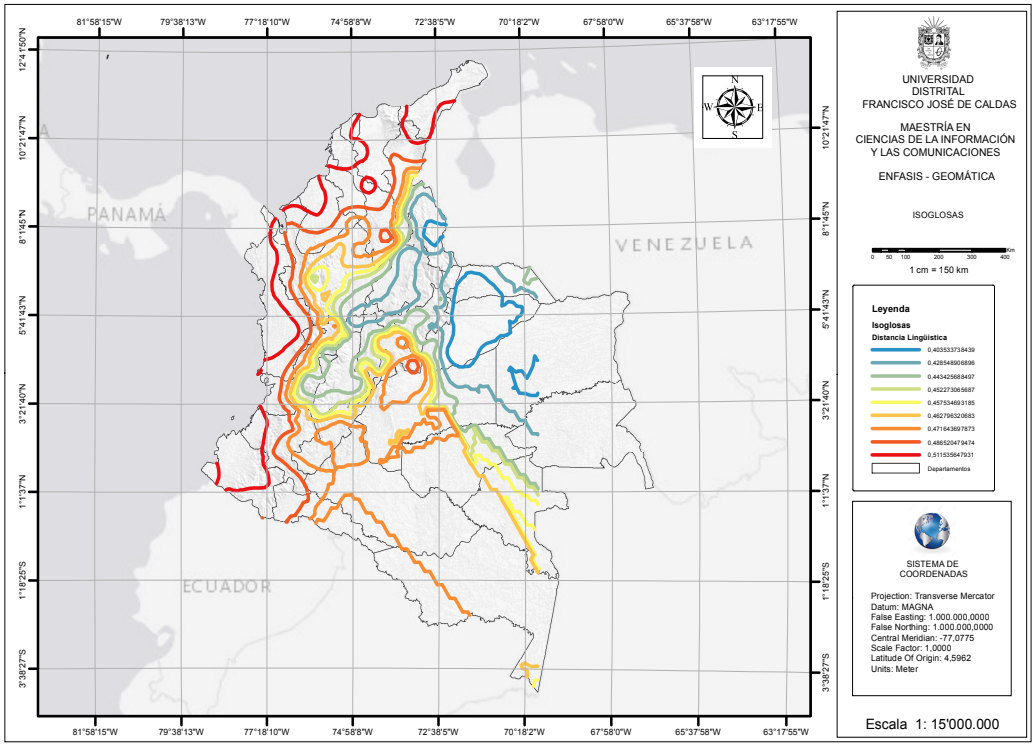


Figura 71: Resultado Isoglosas

8.6. Clasificación dialectal

Para la clasificación dialectal propuesta se utilizó como conjunto de datos base aquellos estimados con el modelo de retardo espacial (modelo seleccionado), el método de agrupamiento empleado fue desarrollado por Guo (2008) en Minería de datos Espaciales y tiene por nombre, regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (REDCAP), este algoritmo permite dividir un gran conjunto de objetos geográficos y optimizar una función objetivo (generalmente de homogeneidad o heterogeneidad) mejorando la agrupación en cada un de las múltiples iteraciones

que realiza, aunque existen seis métodos con tres enfoques de agrupamiento, se utilizó el Full-Order-ALK, método que según Guo (2008) es significativamente mejor a otros métodos (k-means, SKATER, entre otros.) en diferentes aspectos, principalmente en la heterogeneidad espacial total, el equilibrio en el tamaño de la región, la variación interna y la conservación de la distribución de la información; la clasificación dialectal propuesta está basada en la clasificación generada por este algoritmo (REDCAP), la componen en total dos superdialectos, seis dialectos y treinta y dos subdialectos, estos últimos se ilustran a continuación.

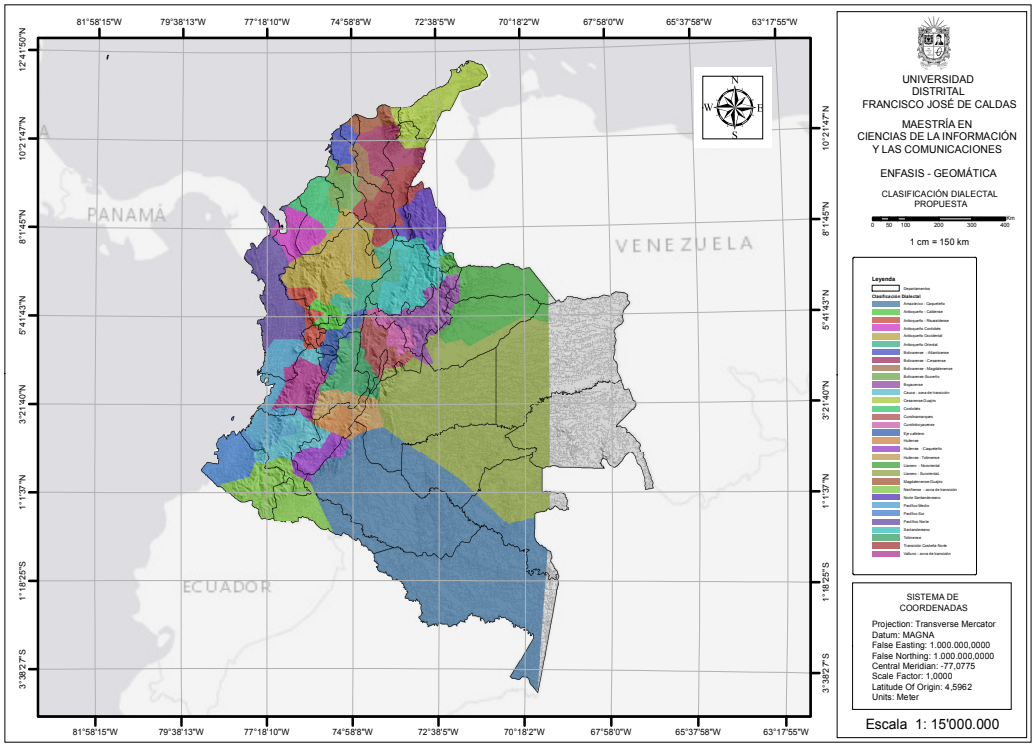


Figura 72: Resultado clasificación dialectal propuesta

9. Análisis de Resultados

9.1. Distancia lingüistica media

El resultado inicial a analizar es la diferencia lingüistica media calculada con el índice relativo de identidad, la principal observación debe hacerse teniendo en cuenta su distribución a lo largo del territorio y es que este recoge la información principal de la disimilitud lingüística entre las localidades, donde la distancia más alta (en una agrupación por clases - 6 clases) está en el intervalo $(0,539638 - 0,617272)$, de ella hacen parte un total de 16 localidades, representando el 6.7 % del total (Cuadro 52).

Localidad	Código	Actividad Económica	Temperatura	Precipitación
Riosucio	Ch10_	Agricultura	Entre 26 y 28 grados	Entre 2500 y 3000
Bahía Solano (Ciudad Mutis)	Ch30_	Agricultura	Entre 24 y 26 grados	Entre 5000 y 7000
Tutunendo	Ch04_	Agricultura	Entre 26 y 28 grados	Entre 7000 y 9000
Arusí	Ch05_	Agricultura	Entre 24 y 26 grados	Entre 5000 y 7000
Cértegui	Ch06_	Minería	Entre 26 y 28 grados	Entre 7000 y 9000
Nóvita	Ch60_	Minería	Entre 24 y 26 grados	Entre 4000 y 5000
San Juan de Mechengue	Ca10_	Agricultura	Entre 16 y 20 grados	Entre 4000 y 5000
Potosí	N_06_	Agricultura	Entre 12 y 16 grados	Entre 1000 y 1500
Puerto Escondido	Co10_	Agricultura	Entre 26 y 28 grados	Entre 1000 y 1500
Cereté	Co11_	Ganadería	Entre 26 y 28 grados	Entre 1000 y 1500
Tres Palmas	Co03_	Agricultura	Entre 26 y 28 grados	Entre 1000 y 1500
San Pedro	Su02_	Agricultura	Entre 26 y 28 grados	Entre 1000 y 1500
San Basilio de Palenque	B_13_	Agricultura	Entre 26 y 28 grados	Entre 1500 y 2000
Pedraza	M_30_	Agricultura	Mayor a 28 grados	Entre 1000 y 1500
Carraipía	G_40_	Comercio	Entre 26 y 28 grados	Entre 500 y 1000
Fonseca	G_05_	Agricultura	Entre 26 y 28 grados	Entre 1000 y 1500

Cuadro 52: Localidades con mayor diferencia lingüística media

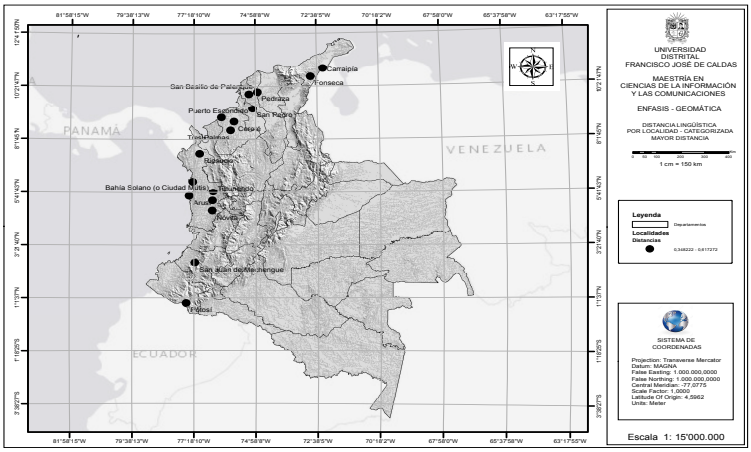


Figura 73: Localidades mayor distancia lingüística media

Entre otros datos, la densidad espacial de los valores más altos de la distancia lingüística media del total de localidades es cerca del Caribe y el Pacífico, lo que en la propuesta de clasificación dialectal fue denominado Costeño, estas 16 localidades están ubicadas específicamente en los departamentos de La Guajira, Bolívar, Sucre, Córdoba, Chocó, Cauca y Nariño, la actividad económica principal es la agricultura (75 %), seguida de la minería (12.5 %) (ver figura 74), un importante dato, dado que es un indicativo de mayor diferencia dentro del modelo seleccionado.

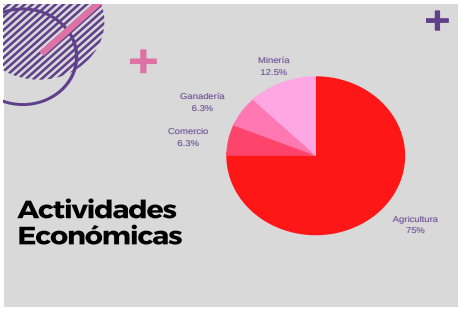
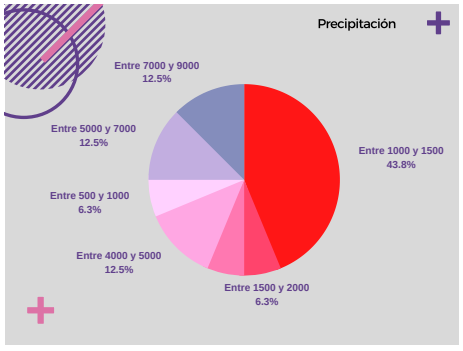
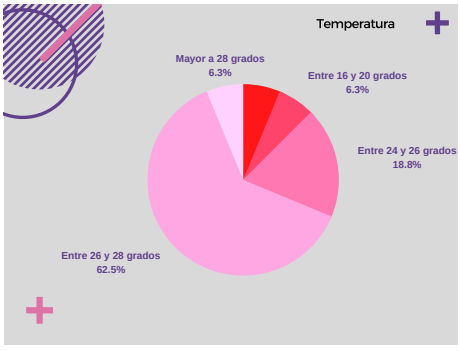


Figura 74: Localidades mayor distancia lingüística - Actividades Económicas

Las actividades económicas constituyen un indicador importante en relación con la alta distancia lingüística en esta zona costeña y según el modelo la minería es una de ellas, de manera equivalente es pertinente mencionar la temperatura y la precipitación (Figura 75) como variables significativas de la siguiente manera:



(a) Precipitación



(b) Temperatura

Figura 75: Temperatura y precipitación mayor distancia lingüística

La precipitación (Figura 75) para estas localidades es la variable con el mayor número de categorías, sn embargo el intervalo Entre 1000 y 1500 representa la mayor parte con un 43 % del total, así mismo la temperatura es alta, constituyendo el 62.5 % en el intervalo, entre 26 y 28 grados, es decir, **es probable que cuando haya una distancia lingüistica alta exista una temperatura entre 26 y 28 grados**, esto fue en parte verificado por el modelo seleccionado sin embargo la significancia de esta variable fue baja.

En el caso de las variables con menor distancia lingüistica, las localidades principales se encuentran en la zona centro-este del territorio, dentro de la región lingüística denominada Andina, los departamentos que contienen estas localidades son Tolima, Cundinamarca, Boyacá, Norte de Santander, Arauca y Casanare (su ilustración se puede ver en la figura 76), en total son 21 localidades de las cuales algunas de las más importantes son Yopal, Sardinata, Arauca, Tauramena, Tame, Orocué y Cúcuta, con distancia lingüística en el intervalo (0,348222 - 0,404176).

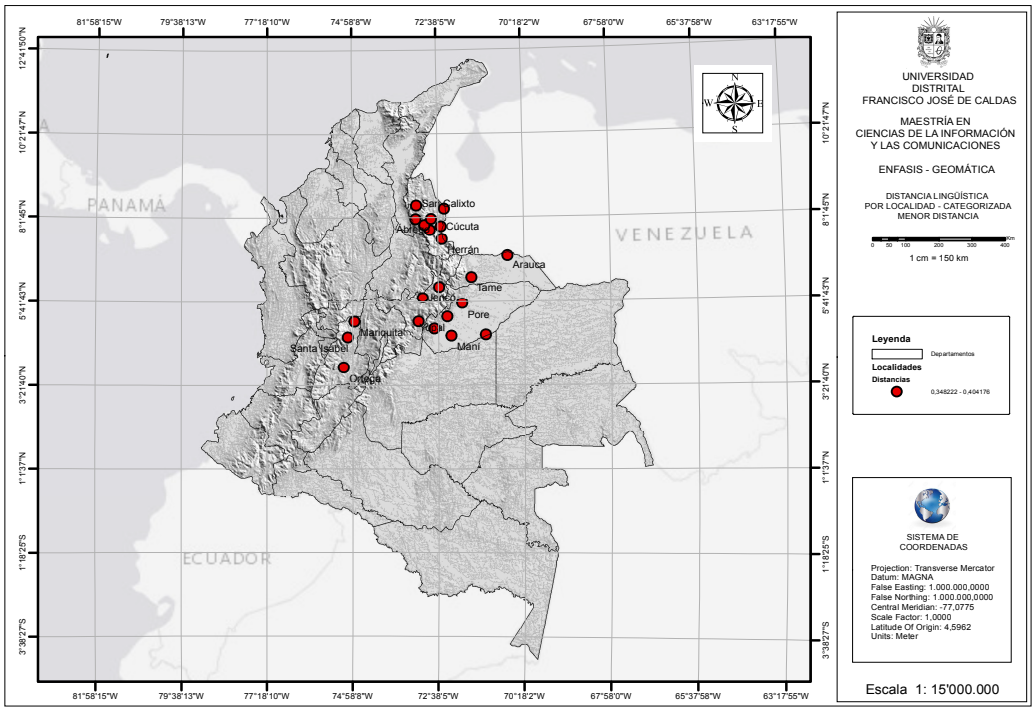
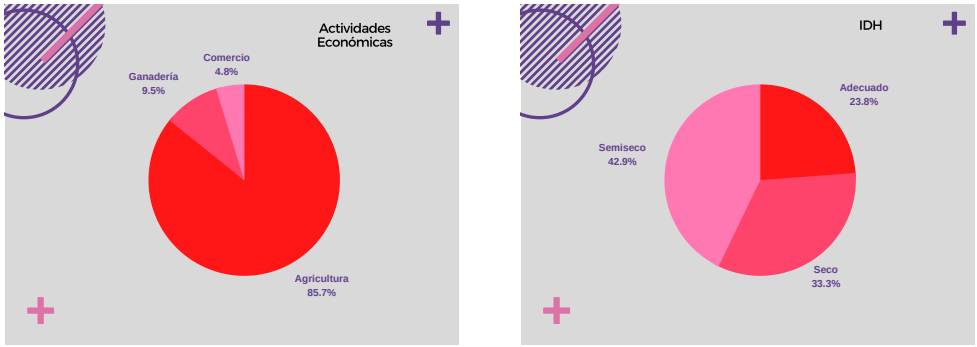


Figura 76: Localidades menor distancia lingüística media

La actividad económica principal de estas localidades es la Agricultura, representando el 85.7 %, seguida de la Ganadería (9.5 %) y el Comercio (4.8 %), dado que la agricultura está presente en la mayoría de localidades independientemente de su distancia lingüística, es probable que el algoritmo para la generación del modelo no la haya tenido en cuenta, otra variable importante es el Indice de Disponibilidad Hídrica, el 5 % está ubicada en una zona categorizada como Adecuada, el 33.3 % en una zona seca y el 42.85 % en una región semiseca, verificando de esta manera la estimación del modelo, el cual plantea que la distancia lingüística disminuye en un -0,0086 cuando la región tiene estas condiciones.



(a) Actividades económicas (b) IDH

Figura 77: Actividades Económicas e IDH menor distancia lingüística

Por último se encuentra la distancia lingüística media con valores centrales, de ella hacen parte las localidades restantes, 198 (83 %), estas se ubican entre las región costeña y la mayor parte de la región Andina, su distancia lingüística media se encuentra en el intervalo (0,4041 - 0,5396); cabe mencionar que ninguna de estás localidades está en el departamento del Chocó, esto se debe a que la distancia lingüistica media para las localidades de este departamento es elevada.

9.2. Distancia geográfica

Una de las maneras de verificar que el valor de la distancia lingüística es un indicador idóneo dentro de la estimación de un modelo espacial y en general para hacer una clasificación adecuada, puede resultar al compararlo

con la distancia geográfica, es decir, si a mayor distancia geográfica hay mayor distancia lingüística puede que el calculo del índice base para posteriores análisis sea el apropiado.

Ahora bien, ya calculados ambos indices (lingüístico y geográfico), el análisis apropiado es verificar que esto se cumple en las localidades mencionadas atrás, el resultado para las localidades con mayor distancia lingüistica se representa en el cuadro 53, la media es de 430 ,637831 km, con un valor mínimo de 363,14617 Km para la localidad Tutunendo y un valor máximo 762,067565 km para Potosí en los departamentos de Chocó y Nariño respectivamente.

Localidad	**Código**	**Distancia Geográfica**
Tutunendo	Ch04_	363,1461703
San Pedro	Su02_	363,7145312
Pedraza	M_30_	370,5040151
Cértegui	Ch06_	370,5385189
Tres Palmas	Co03_	372,8783892
Nóvita	Ch60_	380,3168856
Cereté	Co11_	381,7464853
Fonseca	G_05_	400,4594184
Puerto Escondido	Co10_	417,9557483
Bahía Solano (o Ciudad Mutis)	Ch30_	418,1384836
Riosucio	Ch10_	421,7096394
Carraipía	G_40_	433,1712404
Arusí	Ch05_	433,8953948
San Basilio de Palenque	B_13_	439,7852567
San Juan de Mechengue	Ca10_	560,1775572
Potosí	N_06_	762,0675649

Cuadro 53: Localidades con mayor diferencia lingüística - distancia geográfica

Las localidades que se encuentran en el oeste del departamento del Chocó son algunas de las localidades con mayor distancia geográfica, es llamativo ver como la localidad más apartada como Leticia, en el departamento del Amazonas, no aparece dentro de ellas, esto se puede ver en la visualización de la estimación del modelo seleccionado donde los valores son similares a

aquellos de la zona centro-sur del territorio colombiano.

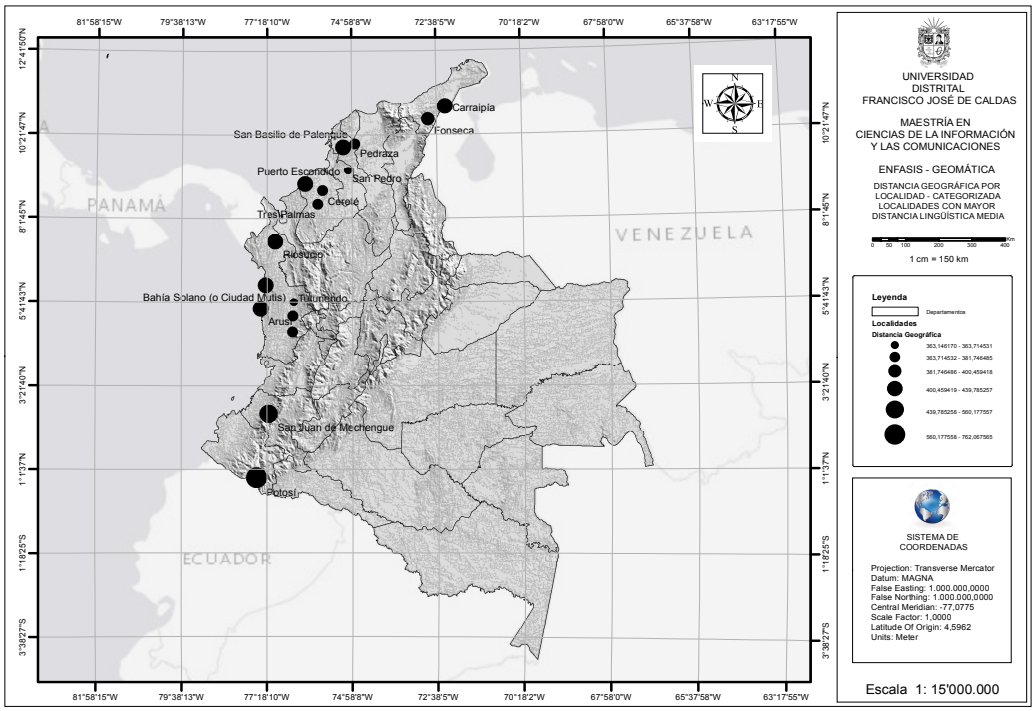


Figura 78: Localidades mayor distancia lingüística media - distancia geográfica

En cuanto a las localidades con menor diferencia lingüística media, se podría decir que de acuerdo a los modelos generados y al estudio de autores como Goebl (2006), Nerbonne y Kleiweg (2007) o Séguy (1973) su distancia geográfica debería ser menor, esto es así para la mayor parte de localidades, sin embargo, existe un caso particular: la localidad de Orocué en el departamento del Casanare, esta localidad cuenta con una distancia geográfica promedio de 1000.12 km con respecto a las demás (distancia más alta que aquellas localidades con mayor diferencia lingüistica media), no obstantes su distancia lingüística media es baja: 0.376, esto hace que el histograma de la Figura 79 tenga un valor atípico en la parte derecha.

A pesar de encontrar la localidad de Orocué, una localidad con una distancia geográfica particular, las demás localidades con distancia lingüistica

media baja presentan distancia geográfica baja, algo acorde a lo esperado, la media para estos lugares es de 341,15 km, de ellas el valor mínimo es de 224,79 km para la localidad Sardinata en el Norte de Santander y excluyendo a Orocué la concentración de valores para las localidades restantes está por debajo de 300 km; de esta manera se podría concluir que la distancia geográfica media guarda una relación inherente con la distancia lingüistica media, proporcional en la mayor parte de la región objeto de estudio (Figura 79).

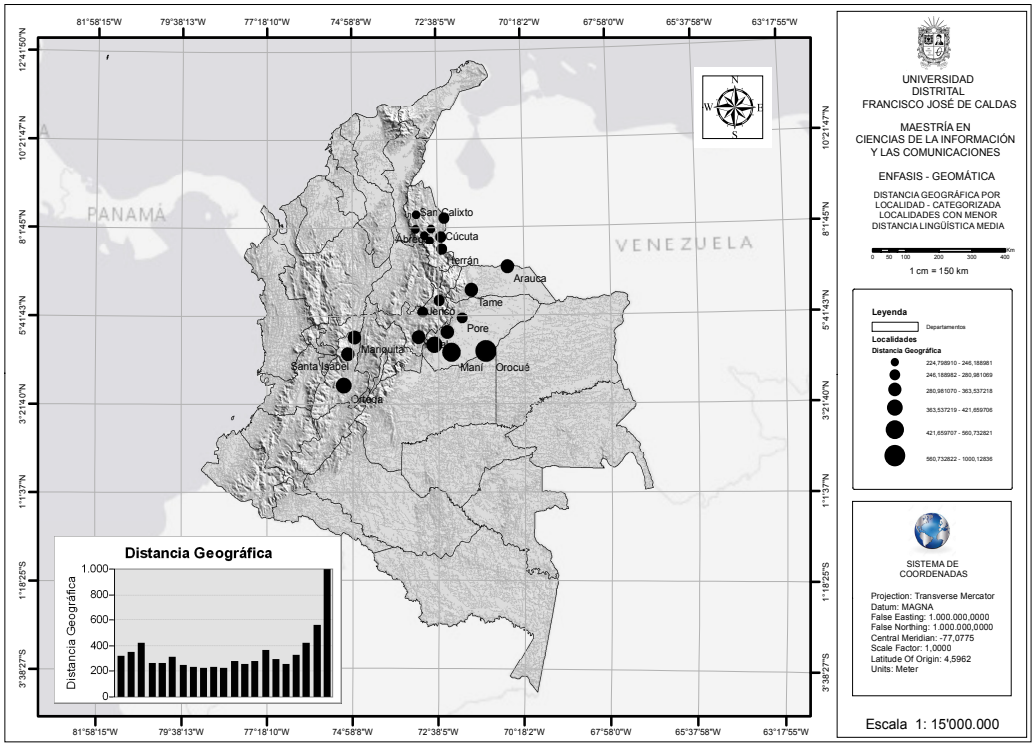


Figura 79: Localidades menor distancia lingüística - distancia geográfica

9.3. Correlación espacial

La correlación espacial univariante y bivariante se presentó como un indicador de dependencia espacial entre las variables que representan a las observaciones dentro de cada localidad, el caso más llamativo como se men-

cionó es el de las actividades económicas, esto debido a que la categoría minería según esto ejerce gran influencia cuando la diferencia lingüistica es elevada, esto se puede ilustrar en la figura 80 donde se presenta esta actividad en localidades como Remedios y Zaragoza en el departamento de Antioquia, Cértegui y Nóvita en el departamento del Chocó e Iscuandé y Barbacoas en el Nariño, adicionalmente se expone la superficie dialectal continua estimada por el método de interpolación Kriging Ordinario (a partir de los datos estimados con el modelo de retardo espacial).

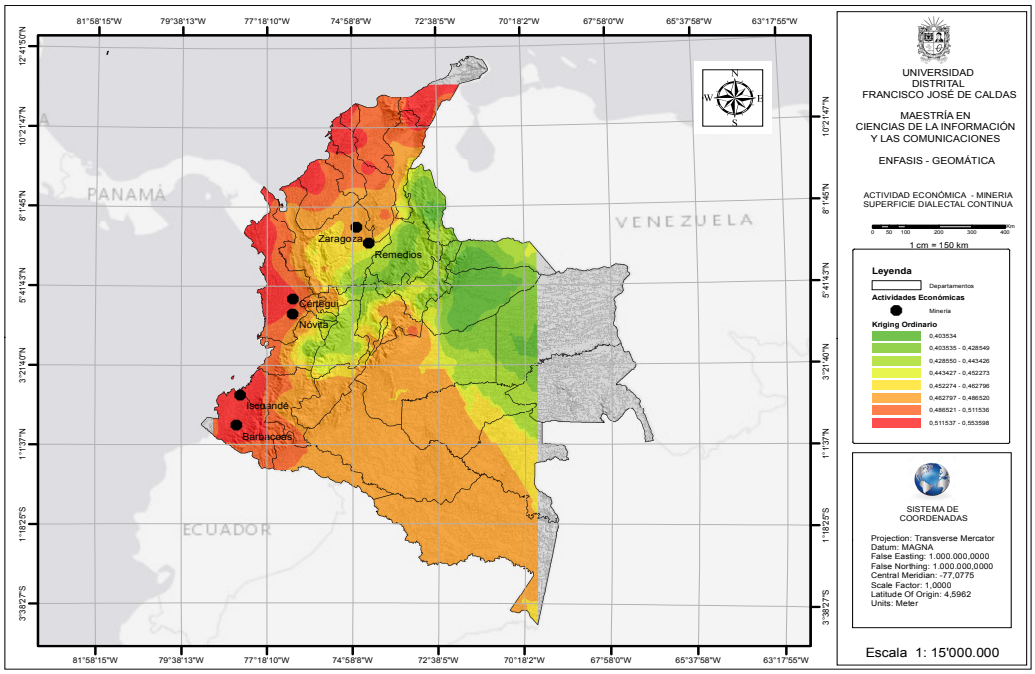


Figura 80: Actividad económica minería y superficie dialectal continua

Es de destacar que las localidades se ubican en el noroeste, centro-oeste y suroeste del territorio, en lugares donde existe una distancia lingüística media alta o cerca de estas regiones (Pacífico Colombiano), de tal manera que se puede plantear la hipótesis de la relación entre la distancia lingüistica alta y zonas con alta influencia afrodescendiente (Barbary y Urrea, 2003) (Figura 81) , así mismo se puede afirmar que el modelo responde y verifica en sus resultados tal hipótesis, sin embargo también se debe mencionar que la muestra es pequeña para tomar este resultado como decisivo.

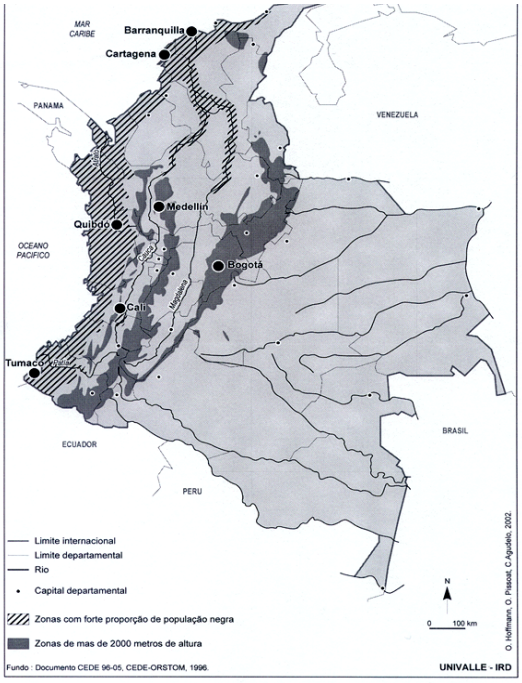


Figura 81: Zonas de población afrodescendiente (Barbary y Urrea, 2003)

Por otro lado se tiene la aptitud agroclimática, donde categorías como extremadamente seco y ligeramente húmedo tienen un valor significativo, según la prueba de correlación espacial bivariante existe una dependencia negativa cuando se trata de la relación entre la categoría ligeramente húmedo y la distancia lingüistica media, de ella hacen parte un total de 33 localidades localizadas en 10 departamentos, la mayor parte de ellas en la zona central, denominada en la clasificación dialectal propuesta como región Andina, la distancia lingüística media para estas localidades se encuentra en el intervalo (0.359, 0.548) y a pesar de ser un intervalo en esencia amplio la mayor parte de valores se encuentra en distancias relativamente bajas, esto supone que el estadístico calculado es correcto; en el caso de la distancia geográfica media pasa algo similar, los valores son en su mayoría bajos, de tal manera que existe una alta congruencia entre el indice de correlación espacial, el modelo estimado y la superficie dialectal continua.

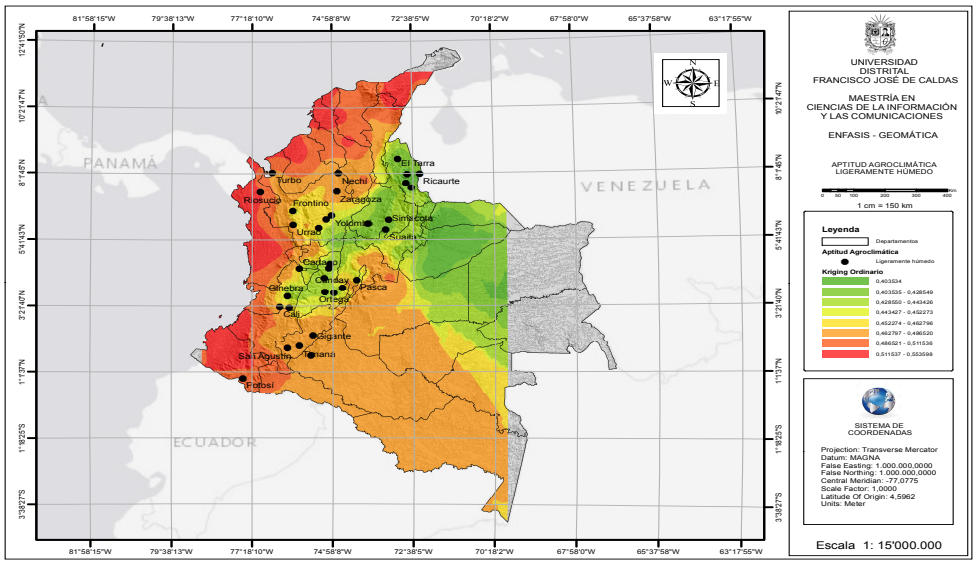


Figura 82: Aptitud agroclimática - Ligeramente húmedo

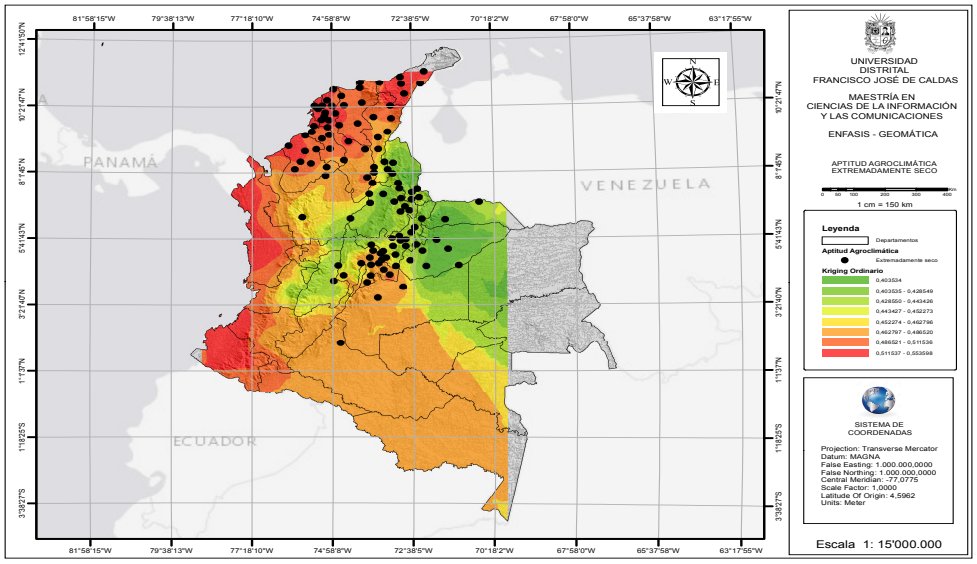


Figura 83: Aptitud agroclimática - Extremadamente seco

Las localidades con aptitud agroclimatica en el rango extremadamente seco según el test de diagnostico de correlación espacial bivariante (I de Moran)

tienen relación positiva y proporcional a la distancia lingüística, sin embargo dentro de la visualización existe una parte de ellas que se encuentra en la región andina, una zona con una distancia lingüística baja, de tal manera que no corresponde del todo con el resultado de la prueba.

La precipitación en el rango "*entre 2000 y 2500*" fue aquel que tuvo mayor índice de correlación espacial bivariante significativo (Sección 8.3), según el modelo seleccionado la **relación es inversamente proporcional**, por lo que se esperaría que los valores en este rango se encuentren en una zona con distancia lingüística media baja, en la figura 84 se observa como la mayor parte de las localidades cumple con esta premisa ubicándose en la superficie dialectal continua en zonas con tonos claros(verde, amarillo), lo cual es un indicador de la coherencia del modelo y la correlación espacial estimada.

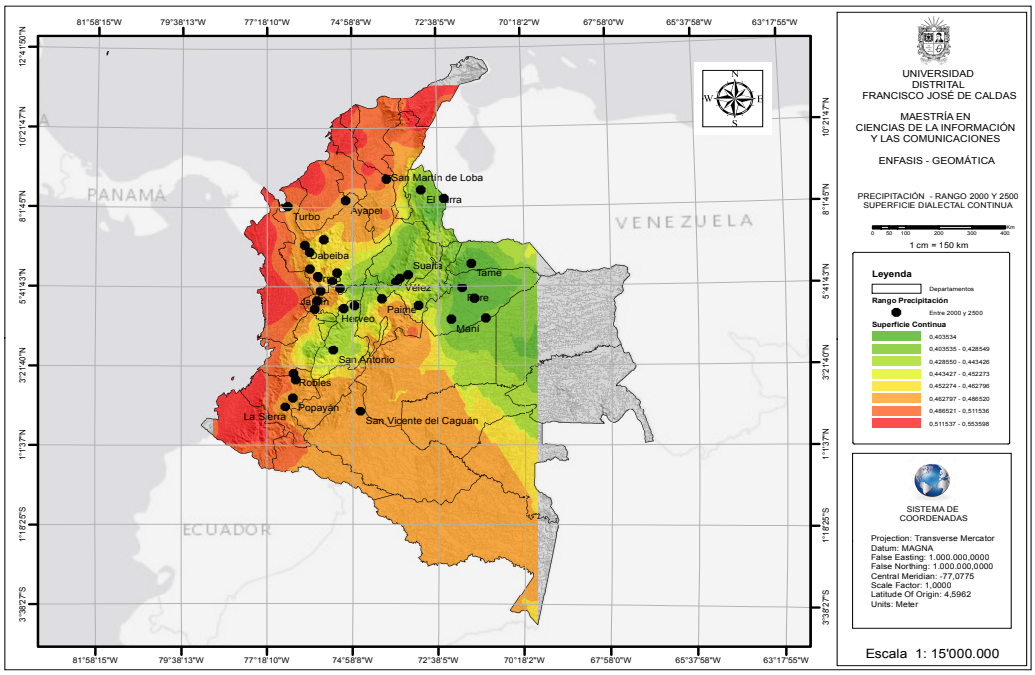


Figura 84: Precipitación - Rango 2000- 2500

El índice de disponibilidad hídrica (IDH) presenta la mayor correlación espacial en la categoría semiseco con un p-valor = 0.001, lo cual supone una relación significativa, dentro del modelo seleccionado, en forma equivalente

se suscita que una localidad con estas condiciones puede presentar menor distancia lingüística media, para verificar dicha afirmación se expone a continuación la ilustración respectiva (Figura 85), en ella la concentración de localidades con este IDH es mayor en la zona central del territorio coincidiendo con la premisa planteada y los test calculados.

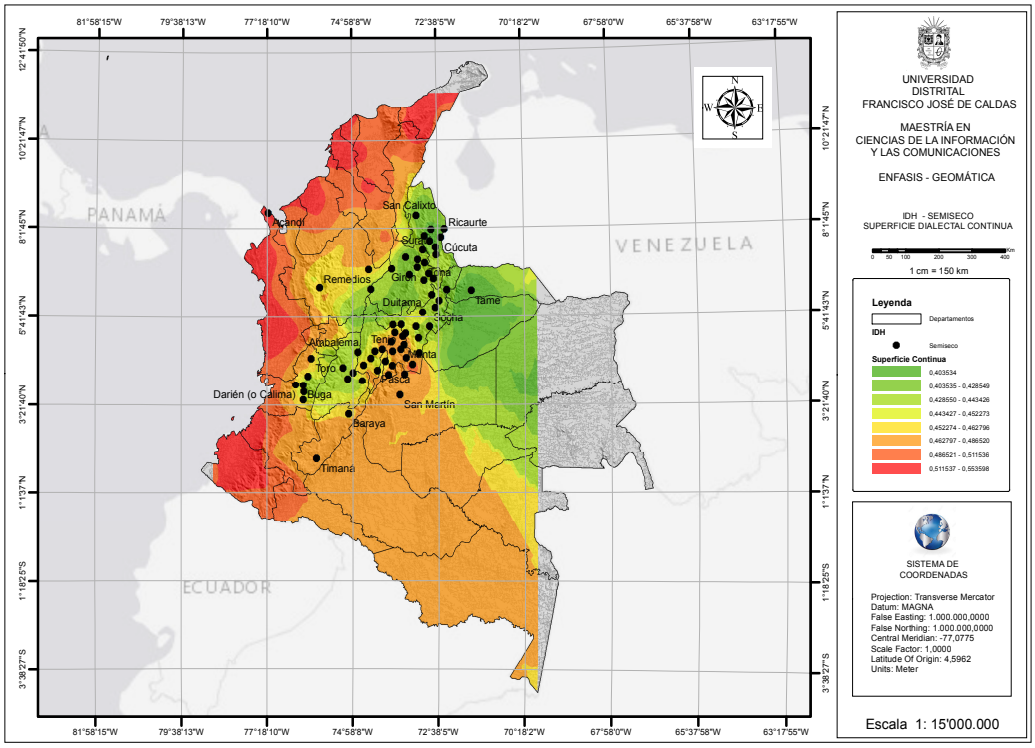


Figura 85: IDH - Semiseco

En total son 66 localidades con esta condición geográfica, representando el 27 % del conjunto seleccionado, lo que hace que sea significativa esta relación en el plano espacial, es importante resaltar que la distancia lingüística como geográfica se concentra en los niveles bajos lo que comprueba nuevamente lo mencionado atrás acerca de esta correspondencia.

Por último se tienen las variables categorizadas dentro del grupo Vías, su significancia representada en el índice de correlación espacial expone que la mayor parte de ellas son importantes en la relación bivariante con la distan-

cia lingüística, en ese sentido es valioso mencionar que el 87 % del total están en la categoría carreteras y el 13 % restante hacen parte de 6 categorías, esto es revelador pues supone que los datos con menor densidad espacial son los más representativos dentro del modelo seleccionado y los de mayor influencia, una de las categorías más representativas es el **Camino de Herradura**, de el hacen parte 4 localidades, Medina en el oriente de Cundinamarca, San Juan de Mechengue y Morales en el departamento del Cauca, y finalmente Pedraza en el departamento del Magdalena; las distancias lingüísticas para estas localidades rondan el intervalo $(0, 459291, 0, 587989)$, lo que implica que están por encima de la media o en zonas medias-altas de la superficie dialectal continua, esta distribución espacial precisa gráficamente lo mencionado en el modelo, **una probable mayor distancia lingüística cuando las localidades (basadas en la configuración espacial - pesos espaciales) presentan esta vía de acceso**.

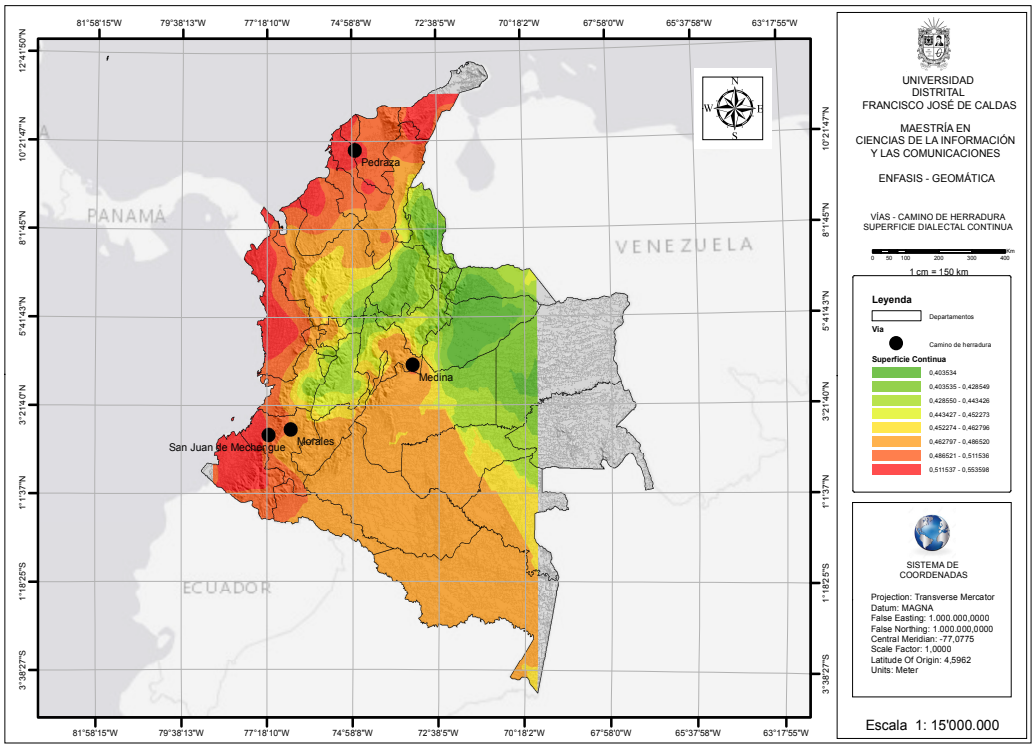


Figura 86: Vías - Camino de herradura

9.4. Modelo estimado

9.4.1. Modelo lineal no espacial

El modelo lineal estimado por mínimos cuadrados ordinarios fue la base para la generación de los modelos espaciales, entre las variables más destacadas dentro de esta estimación no espacial se encuentra la distancia geográfica, que según los resultados **aumenta en 0.05 la distancia lingüística media por cada unidad (utilizando su algoritmo)**, así mismo integra una variable que dentro del modelo espacial no fue suficientemente significativa, regionalizada en todo el territorio y denominada *Amenaza por riesgo sísmico*, las localidades con condiciones *Extremadamente Baja* son en general aquellas con una incidencia mayor, sin embargo de esta categoría solo hace parte una localidad, Leticia en el departamento del Amazonas, lo cual no es determinante como se asevera en el modelo; por otra parte la categoría Moderadamente Alto cuenta con 30 localidades y aunque el modelo supone que hay menor distancia lingüística esta condición está distribuida a lo largo de los departamentos de Antioquia, Bolívar, Boyacá, Caldas, Cesar, Córdoba, Caqueta, Casanare, Cundinamarca, Magdalena, Sucre y Magdalena, doce departamentos, que en general tienen un comportamiento variado de la lengua en el territorio y que en el momento de compararse con los datos lingüísticos originales y los estimados son en esencia distintos.

La temperatura es una variable importante dentro del modelo lineal no espacial, especialmente en el rango que va *entre 24 y 26 grados celsius*, de ella hacen parte un total de 23 localidades, sin embargo al igual que la *Amenaza Sísmica* están distribuidas a lo largo de numerosos departamentos (13), entre ellos Antioquia, Amazonas, Boyacá, Cauca, Chocó, Caquetá, Casanare, Cundinamarca, Meta, Nariño, Norte de Santander, Tolima y Valle del Cauca, exponiendo así que no existe una continuidad espacial homogénea y por lo tanto que su correlación espacial con la distancia lingüistica media es baja.

Otra variable destacada dentro del modelo de regresión lineal es el Índice de Disponibilidad Hídrica (IDH) cuya categoría *Muy seco* tiene una alta significancia, según el resultado la distancia lingüística aumenta en 0,0423 cuando una localidad presenta esta condición, al verificar esta información se pueden encontrar un total de 25 localidades, ubicadas en los departamentos

del Caribe y Magdalena, con un rango de distancia lingüística en el intervalo $(0{,}52, 0, 56)$, cabe resaltar que dentro del modelo espacial no fue relevante, esto puede deberse a que no cumple con los parámetros de contigüidad requeridos.

Finalmente se tienen las variables categorizadas dentro del grupo Vías de Acceso, de ellas las categorías más significativas son los Caminos Carreteables, las Carreteras, las Carreteras y Vía Férrea, y finalmente la Vía Férrea, las localidades con estas categorías se pueden ilustrar de la siguiente manera:

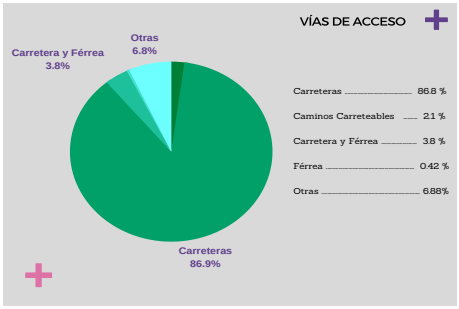


Figura 87: Vías de acceso - modelo estimado por MCO

El 86.8 % del total de localidades tienen como vía de acceso principal la Carretera, esta es una variable bastante significativa según el modelo lineal estimado, haciendo que en general la distancia lingüística dependa en gran parte de ella, lo que resulta interesante dado que la muestra (208 localidades) es numerosa, de esta manera se puede concluir que el modelo lineal recoge algunas de las variables que pueden intervenir en el fenómeno objeto de estudio, sin embargo, dado que la esencia de los datos es espacial hay algunas variables que no son tenidas en cuenta y que pueden resultar valiosas en el momento de realizar una mejor o correcta estimación.

A partir de lo anterior, el modelo lineal de regresión estimado por mínimos cuadrados ordinarios fue objeto de algunos contrastes de dependencia espacial, cuyo resultado arrojó que el tipo de dependencia más significativa era aquella referida a la influencia de variables geográficas y metadatos del ALEC en localidades vecinas, teniendo presente que la distancia lingüística sería retardada espacialmente, esto se puede verificar en la tabla 49 con el

estadístico del test de LM - LAG, una prueba robusta con un valor significativo, a continuación siguiendo la teoría de Anselin (2005) se tomó como un modelo probable para lograr una correcta estimación, el modelo de retardo espacial, no obstante se realizaron otros adicionales y con ayuda de algunos estadísticos de prueba adicionales se logró obtener un modelo consistente denominado Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial para variables significativas (VMS) el cual se analiza a continuación.

9.4.2. Modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial (VMS), protocolo de pruebas y validación

Los resultados más destacados del modelo seleccionado muestran las variables más importantes dentro de aquellas escogidas inicialmente, a continuación se señalan las localidades que tienen mayor inferencia dentro del modelo espacial, ya sea por sus condiciones geográficas o por sus metadatos.

Cuando aumenta la distancia lingüística media las variables de mayor significancia son, la actividad económica Minería, la aptitud agroclimática, *moderadamente húmedo* y la mayor parte de las categorías del grupo Vías de acceso; puede esperarse que algunas localidades tengan estas condiciones de forma simultanea, sin embargo no es del todo así, localidades como las siguientes (Figura 54), aunque presentan la actividad económica Minería, no cumplen la mayor parte de ellas, esto puede deberse a las restricciones de contigüidad del modelo así como el pequeño número de localidades que presenta esta condición.

Localidad	Código	Dist. Geográfica	Dist. Lingüística	Vías de Acceso	Temperatura
Barbacoas	N_30_	716,91	0,489	Carretera	Entre 24 y 26 grados
Cértegui	Ch06_	370,54	0,582	Carretera	Entre 26 y 28 grados
Iscuandé	N_01_	638,79	0,504	Fluvial	Entre 26 y 28 grados
Nóvita	Ch60_	380,32	0,586	Carretera	Entre 24 y 26 grados
Remedios	A_41_	276,65	0,458	Carretera	Entre 24 y 26 grados
Zaragoza	A_40_	297,85	0,428	Carretera	Entre 26 y 28 grados

Cuadro 54: Localidades con actividad económica - Minería

El cuadro muestra como las localidades con la actividad económica Minería tienen una distancia lingüística media alta, como se mencionó atrás (Figura 80) se encuentran localizadas en el centro-oeste del territorio, en los

departamentos de Antioquia, Chocó y Nariño, departamentos relativamente cercanos, así mismo aunque en el modelo la temperatura no es una variable determinante se puede observar como la temperatura para estas localidades se encuentra entre 24 y 28 grados centigrados, una temperatura promedio para zonas costeras, la vía de acceso principal como se verificó es la Carretera, ella según el modelo hace que la distancia lingüística aumente en 0,0443168 por cada localidad que tenga esta condición (según la matriz de pesos).

Dentro de la revisión adicional de estas localidades es importante mencionar el año de realización de la respectiva encuesta, ya que puede facilitar los procesos de análisis de la información en un contexto histórico, de tal manera, sus respectivos años son los siguientes:

Localidad	**Año de encuesta**
Zaragoza	1975
Remedios	1962
Cértegui	1974
Nóvita	1974
Iscuandé	1973
Barbacoas	1961

Cuadro 55: Localidades con actividad económica - Minería (año encuesta)

Es de resaltar que el promedio de encuesta sea el año 1969 para estas localidades, ya que en un contexto histórico y desde una perspectiva económica es trascendental, los resultados en este sentido son acordes a algunos de las menciones realizadas por otros autores acerca del desarrollo de la economía en el territorio, según Kalmanovitz (2015) desde la época colonial una de las características de nuestro territorio ha sido el enfoque hacia la explotación de materias primas (en este caso la minería de carbón, oro, entre otros), en principio por esclavos negros traídos del África y desde el siglo XVII y XVIII por los colonos y campesinos, lo cual implicó un proceso de migración interna hacia regiones como el Cauca, Antioquia y Chocó, regiones que en el momento de hacerse la encuesta tenían aún esta actividad como fuente de desarrollo y por lo tanto llevaban cerca de 300 años ejecutándola de forma continua, de esta manera se puede concluir que los resultados del modelo no solo dan cuenta de los procesos económicos del año de encuesta sino que

abarcan procesos de desarrollo desde principios de la colonización.

Las variables geográficas utilizadas dentro del modelo son, la distancia geográfica, la aptitud agroclimática, la precipitación, la amenaza por riesgo sísmico, la temperatura y por último el índice de disponibilidad hídrica, de ellas ya se analizaron las localidades y su distancia geográfica, con base en los resultados se verificó la hipótesis de otros autores con respecto a su relación con la distancia lingüística; en cuanto a la aptitud agroclimática, se verificó la correlación espacial en dos de las tres categorías, donde la distancia lingüística media suele ser menor en zonas ligeramente húmedas y mayor en extremadamente secas, tal es el caso de la zona andina que agrupa el mayor número de localidades con rango de distancia lingüística media similar y menor con respecto a aquellas ubicadas en la zona costeña.

La categoría faltante es la *aptitud agroclimática-moderadamente húmedo*, la cual dentro del modelo es la más significativa, de ella hacen parte un total de doce localidades, es decir el 5 % del total, dado que la distancia lingüistica media está espacialmente retardada, cuando hay localidades contiguas con esta condición se puede esperar que esta aumente aproximadamente en 0.026, esto se puede verificar gráficamente en la Figura 88.

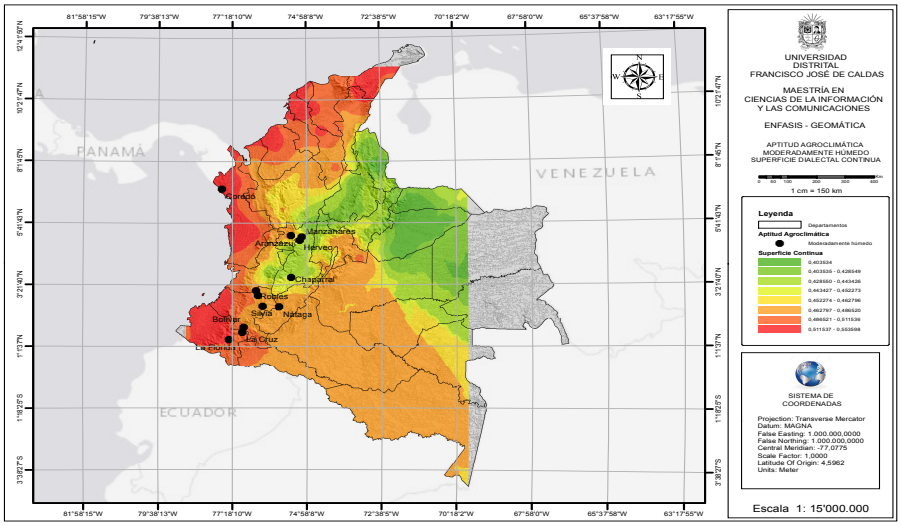


Figura 88: Aptitud agroclimática - moderadamente húmedo

Los departamentos que presentan esta condición agroclimática (Figura 88) son, Caldas, Cauca, Chocó, Huila, Nariño, Tolima y Valle del Cauca, departamentos de la zona suroccidental del territorio, en este lugar existe una frontera de transferencia lingüística que se extiende de la zona andina hacia la zona costeña; el resultado y análisis de esta investigación permite concluir que existe una relación significativa entre esta variable geográfica (aptitud agroclimática) y la distancia lingüística media, verificada en forma cuantitativa e ilustrada gráficamente.

Con respecto a la precipitación el rango *entre 2000 y 2500* permitió verificar que la región del territorio con esta condición tiende a disminuir la distancia lingüística media, o las localidades contiguas tienen una distancia lingüística media menor; con respecto al rango que está *entre 500 y 1000* las localidades están dispersas en el territorio, algunas en la zona del caribe en los departamentos del Atlántico y la Guajira, y otras en la zona central en los departamentos de Cundinamarca y Boyacá, por lo que la información y la relación en este rango presenta una alta variación; con respecto a la amenaza sísmica, los valores de distancia lingüistica media son más altos cuando la zona tiene condiciones de riesgo *moderadamente altas* según el modelo, no obstante hay algunas que se encuentran en zonas de transición y en la zona central donde esta distancia es menor, tal como se ilustra en la figura 89.

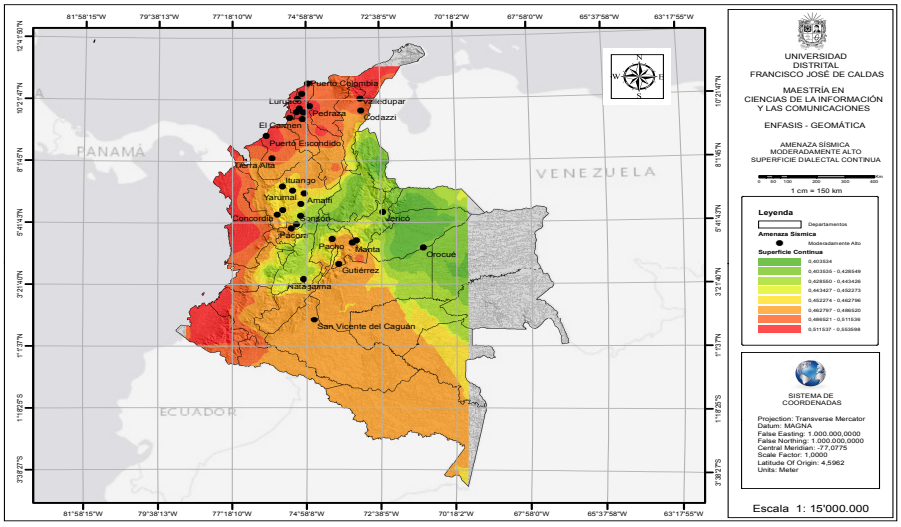


Figura 89: Amenaza Sísmica - Moderadamente Alto

En total de la variable amenaza sísmica con riesgo *moderadamente alto* hacen parte 30 localidades, la mayor parte de ellas en zonas con distancia lingüística media alta, por lo que se puede concluir que existe una relación con la misma, no obstante esta no es significativa, por lo que el modelo se adecua en tal sentido; la temperatura es una de las variables que dentro del modelo lineal no espacial se mencionó, en el se denotaron algunas categorías como influyentes en el fenómeno objeto de estudio, de forma equivalente en este modelo dos categorías aparecen, sin embargo la influencia de las mismas es mínima, tales categorías se muestran en las Figuras 90 y 91.

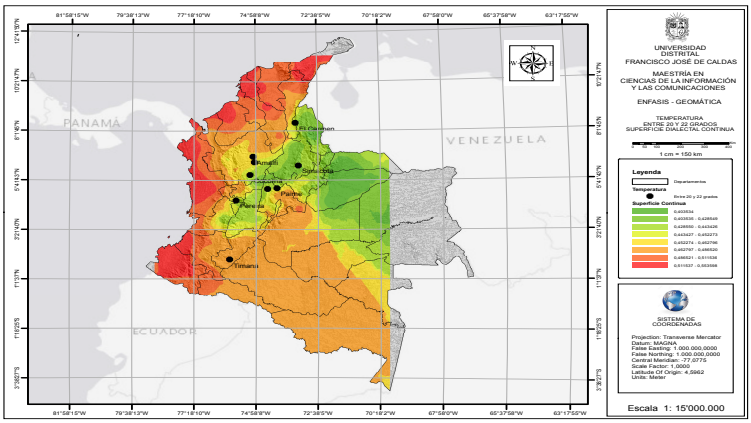


Figura 90: Temperatura - Entre 20 y 22 grados

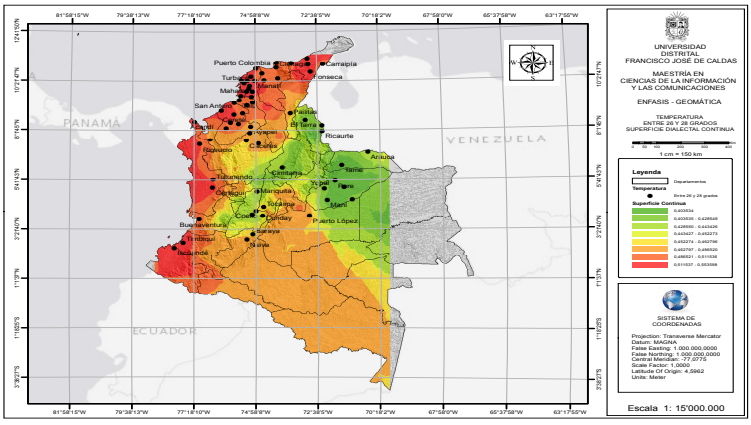


Figura 91: Temperatura - Entre 26 y 28 grados

El rango de mayor influencia es el que se encuentra *entre 20 y 22 grados*, de el hacen parte localidades ubicadas en el centro del territorio, donde disminuye la distancia lingüística media, el segundo rango es un poco mas variado y aunque la temperatura es alta (*entre 26 y 28 grados*), está distribuido tanto en zonas costeras como centrales, lo que permite conocer su variación cuando se visualiza gráficamente; la relación con el rango de temperaturas bajas (*entre 20 y 22 grados*) existe y aunque no sea significante dentro del modelo, hace que disminuya la diferencia lingüística media en ciertas regiones centro-occidentales del país; por último se tiene la variable geográfica IDH, donde el rango *semiseco* y *semihúmedo* son los más significativos, ambos tienen una relación inversa con la distancia lingüística, esto se puede ver en las Figuras 92 y 93 donde los valores se concentran en el centro, sur y oeste del territorio.

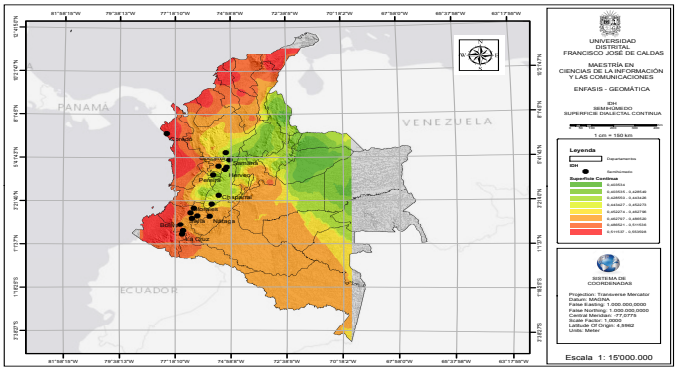


Figura 92: IDH - Semihúmedo

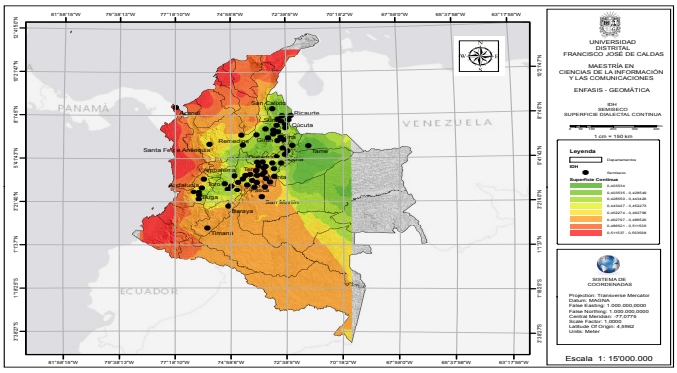


Figura 93: IDH - Semiseco

Las dos estructuras tienen algo en común (Figuras 92 y 93), la mayor parte de localidades se encuentran en la zona central del territorio, la particularidad para cada una es que, la zona centro-sur comprende condiciones geográficas *semihúmedas* y la zona centro-norte condiciones geográficas *semisecas*, no obstante ambas regiones tienen la distancia lingüística media menor al promedio del total de observaciones lo cual supone que existe una relación significativa con el fenómeno lingüístico objeto de estudio y por lo tanto se puede concluir que si hay una relación de las variables geográficas con la distancia lingüística, de tal manera que la hipótesis de la investigación es fehaciente.

Por último se tiene la variable tomada de los metadatos del ALEC y denominada Vías de Acceso, dentro del modelo hay cuatro categorías significativas, los *Caminos de Herradura*, los *Caminos Carreteables*, las *Carreteras* y aquellas localidades con *Carreteras y vía férrea*, según el modelo seleccionado todas aumentan significativamente la distancia lingüística media, con influencia especial de los *Caminos de Herradura*; para determinar las localidades que tienen estas condiciones y donde están ubicadas, en primer lugar se verán el porcentaje específico que hacen parte de cada una de las categorías y luego su respectiva ilustración dentro del territorio.

Vías de Acceso	**Localidades**	**Porcentaje**
Camino de herradura	4	1,68
Carretera y férrea	9	3,78
Caminos carreteables	5	2.1
Carretera	208	87,39

Cuadro 56: Localidades con vías de acceso - mayor influencia

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 la categoría *Camino de Herradura* es la más influyente sobre las demás teniendo solo un 1.68 % de localidades en su haber.

Ahora, la ubicación de estas localidades es diferente a las otras variables ya que en su mayoría no tienen una relación con la superficie dialectal, es

el caso de la categoría *Caminos carreteables* que se distribuye en el oriente del territorio entre los departamentos de Casanare y Arauca, una zona con valores de distancia lingüistica media baja, otra categoría que se comporta de forma similar es *Carreteras*, cuya distribución abarca todo el territorio a excepción del Amazonas, cabe resaltar que esta es la de menor significancia entre las demás categorías; finalmente los grupos restantes, tienen una ubicación mucho más cercana a los resultados que supone el modelo y su significancia se ajusta a tal fín lo que permite concluir que existe una correlación espacial alta cuantificable entre las vías de acceso y la distancia lingüística media dentro del territorio colombiano.

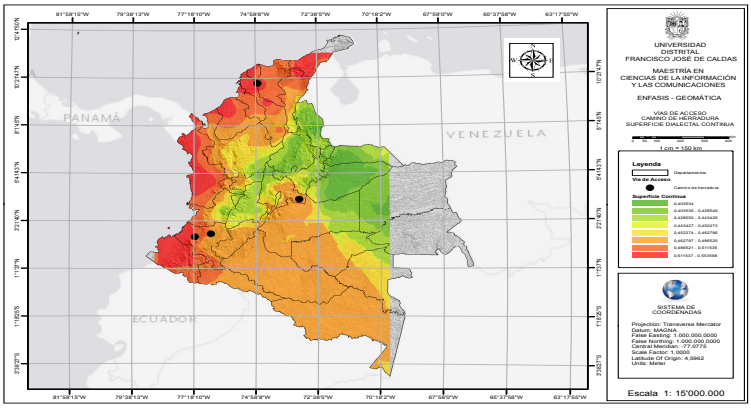


Figura 94: Vías de acceso - Camino de herradura

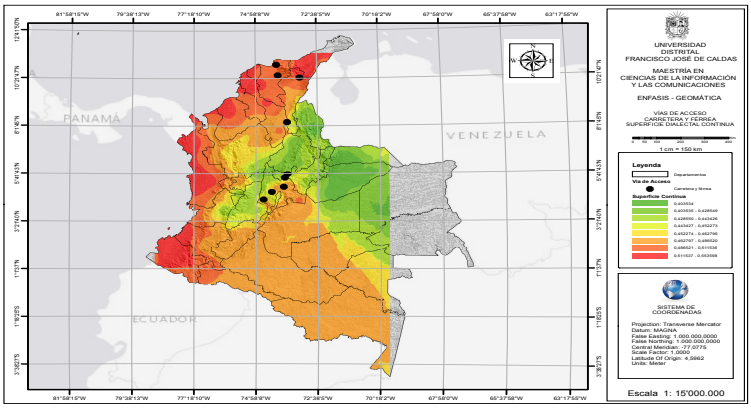


Figura 95: Vías de acceso - Carretera y férrea

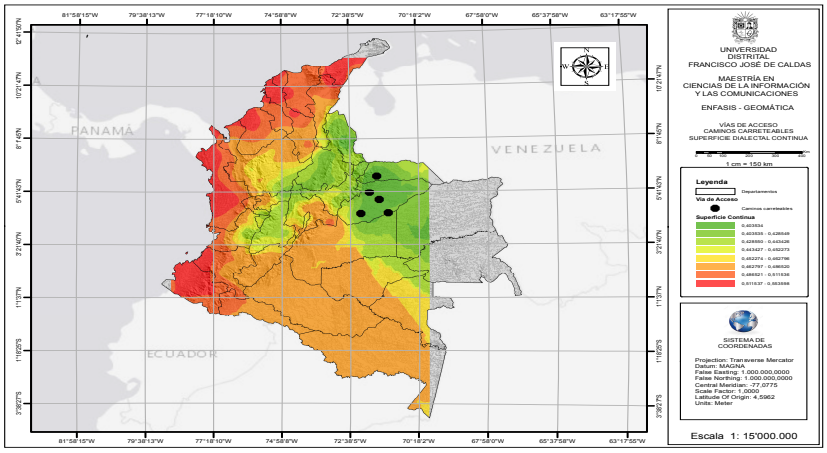


Figura 96: Vías de acceso - Caminos carreteables

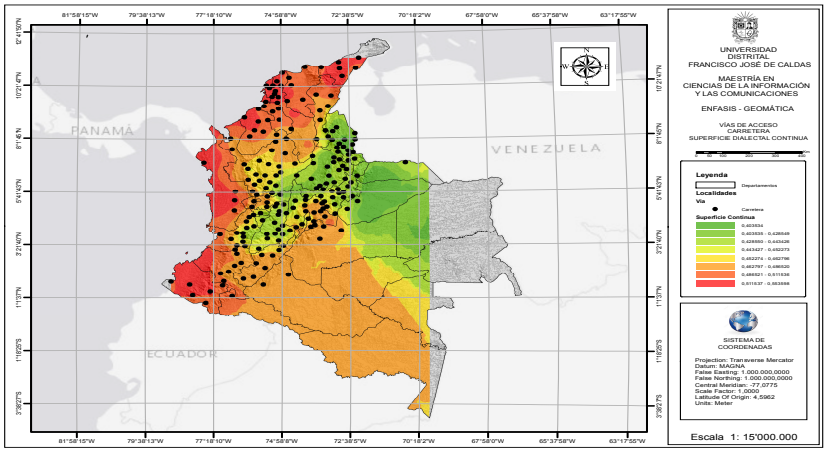


Figura 97: Vías de acceso - Carretera

Según Pérez (2005) durante la primera mitad del siglo XX, el desplazamiento por carretera entre los distintos **centros urbanos era compleja**, esto debido al lento desarrollo de infraestructura a lo largo de todo el territorio, no obstante hacia finales de los años cuarenta y principios de los cincuenta en el periodo de la posguerra inició una etapa en el desarrollo vía del país, con la construcción de nuevas redes de transporte, la causa interna según Pérez (2005) era la necesidad de movilizarse hacia los centros de oferta de bienes y servicios, concentrada en la región central, es decir la Sabana de

Bogotá y sus alrededores; en general otros autores como Pachón (2006) plantean que la concentración espacial de la red nacional de vías fue en la región andina y sólo a partir de los años 60' la proporción de carreteras pavimentadas aumentó en departamentos que no habían sido tenidos en cuenta, como el Cauca, Huila, Caquetá, Meta, Nariño, Magdalena y La Guajira, es decir, aunque existian caminos y/o vías, su dificultad de acceso y su lejanía de la región central dan prueba de la concentración no solo de la lengua sino de esta relación espacial con actividades económicas y procesos de movilización y/o comunicación vial (Figura 98).

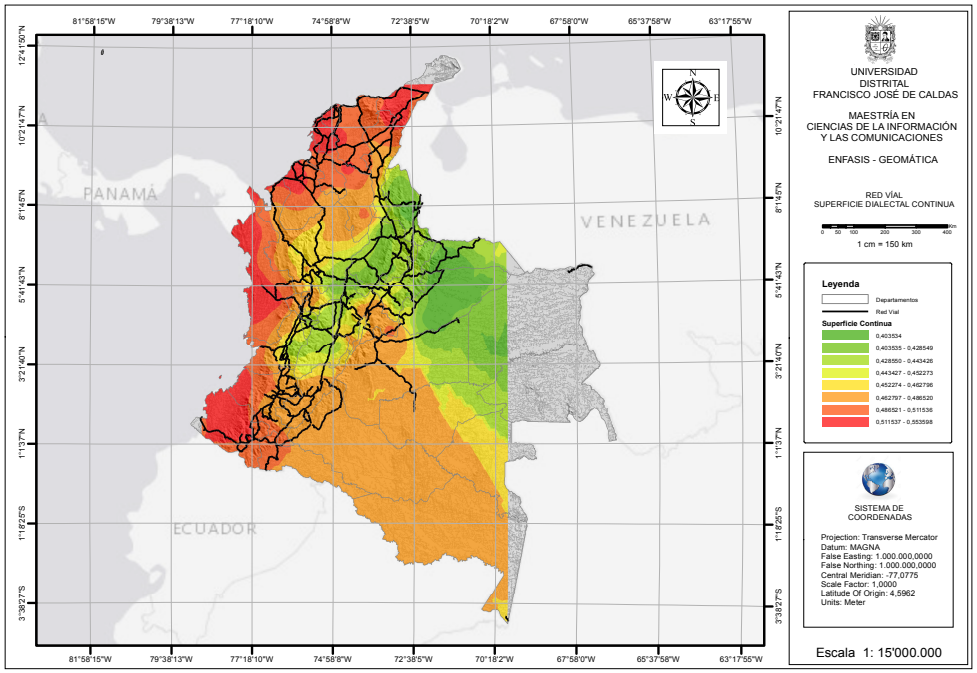


Figura 98: Red Vial (INVIAS) - superficie dialectal continua

9.5. Limites Lingüísticos

Los límites lingüísticos generados con el modelo espacial seleccionado fueron las isoglosas y la superficie dialectal continua, ambos límites muestran en forma gráfica la distancia lingüistica entre las distintas localidades donde se realizó el cuestionario del ALEC, dentro de este capítulo se ha trabajado especialmente con esta última, sin embargo es preciso resaltar que ambas

proceden del mismo origen, de ellas se pueden referenciar zonas comunes, tales como, la zona central que recoge los departamentos del Tolima, noroccidente de Cundinamarca, Boyacá, Casanare, Arauca, Santander y Norte de Santander, la zona oeste que recoge todos los departamentos del Pacífico y el Caribe como el oeste de Nariño, Cauca, Valle del Cauca, Chocó, Antioquia, Córdoba, Súcre, Bolívar, Magdalena, Cesar, Atlántico y La Guajira, y finalmente una zona suroriental, con los departamentos del Meta, Caqueta, Huila, Putumayo y Amazonas.

La zona con menor distancia lingüística media con respecto al total es la zona central, seguida de la suroriental y la occidental, como se mencionó, existen tanto variables geográficas como grupos de metadatos que pueden influenciar dicha distancia, a nivel histórico se mencionó como las actividades económicas y la facilidad o dificultad de movilización podrían tener algún papel en dicho fenómeno, no obstante es menester de investigaciones posteriores abarcar este análisis de una forma más profunda, esto no impide sin embargo que esta investigación de luces acerca de fenómenos sociales adicionales implicados, tal es el caso del fenómeno demográfico y de migración interna registrada en el siglo XX y es que según Mejía (2013) es el principio de la construcción de un país urbano basado en el crecimiento y asentamiento poblacional en las cabeceras municipales principales.

Mejía (2013) expone que hacia el año 1.912 la población colombiana no superaba los cinco millones de colombianos, predominando la ubicación del 75 % de la población fuera de las cabeceras municipales, es decir, Colombia era por entonces un país rural, resaltando que el 25 % restante tenía el control real del territorio, dominando las relaciones de mercado, las instituciones estatales y dinámicas socioculturales, por todo esto las cabeceras municipales se convirtieron en los centros de dirección del crecimiento y asentamiento de la población, estas dinámicas dieron lugar a fenómenos tales como la expansión de la frontera interna y otra la migración hacia las grandes ciudades, esta última a razón de políticas gubernamentales y al modelo de desarrollo capitalista adoptado.

Otro fenómeno mencionado por Mejía (2013) es la expansión de la frontera agrícola, fenómeno que por causas de migración interna tuvo que expandirse

hacia tierras cálidas en los valles interandinos, especialmente hacia las tierras selváticas de la Amazonía y Orinoquía (Figura 99), en las ciudades por otra parte, dada la consolidación del estado y de la adopción del capitalismo en forma paralela se incrementó la riqueza de unos pocos y así mismo aumento el número de pobres, Mejía (2013) manifiesta que 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 Mejía (2013) son de reciente colonización, por último es de precisar que desde tiempos de conquista la región de mayor preferencia para el asentamiento de la población fue la región Andina, seguida de la región Caribe, cuyas diferencias culturales son marcadas, de tal manera que el modelo generado y la superficie dialectal continua se ajustan a ello.

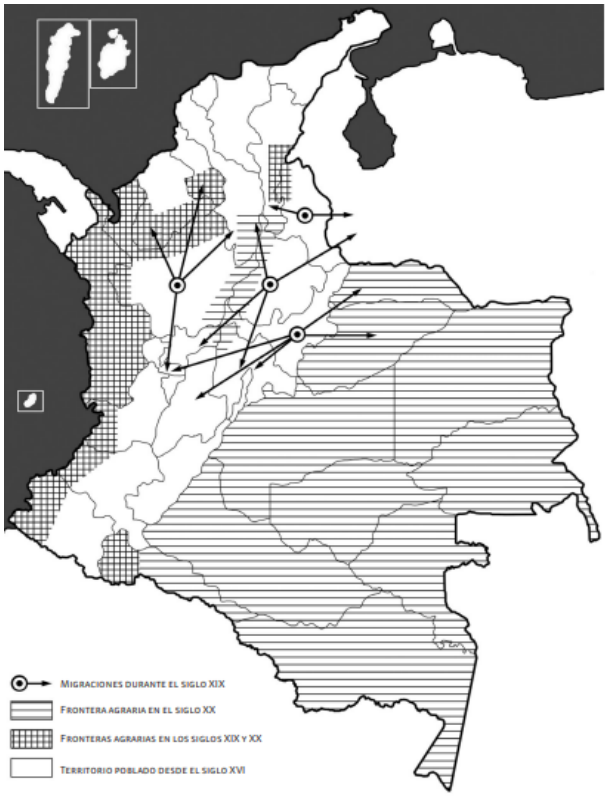


Figura 99: Movimientos de población siglo XIX (Mejía, 2013)

Sin lugar a dudas la representación de los movimientos se ajusta a algunos de los límites lingüísticos trazados dentro de la presente investigación, quizás una de las migraciones internas más llamativa es aquella que se encuentra al centro-oriente del territorio hacia los departamentos de Casanare y Arauca, departamentos donde la distancia lingüística es menor, esto puede de alguna manera ser explicado por la explotación de minerales como el petroleo (Granados, 2010) en la década de los 70's y la iniciativa gubernamental para mejorar las vías de comunicación con el altiplano andino (Herrera, 2017), Herrera (2017) señala que bajo la presidencia de Alfonso López Michelsen se construyó el eje vial entre los Andes y la Orinoquía, por la cuenca del río Upía en el sur-occidente de Casanare, conectando municipios como Guateque en Boyacá y Sabanalarga en Casanare, razón que puede entre otras ser la causa para tal similitud lingüística media entre estos dos departamentos.

9.6. Clasificación dialectal

En el apartado anterior se hizo una contextualización breve de la relación entre los límites lingüísticos y algunos fenómenos geográficos, históricos y sociodemográficos, en este aparte se pretende determinar la fiabilidad del modelo a partir de la implementación de un protocolo de pruebas sobre los resultados, teniendo presente los estadísticos de prueba ejecutados así como el análisis y comparación de la propuesta de clasificación dialectal de esta investigación y la realizada por otros autores, integrando los límites lingüísticos generados como complemento determinante en el agrupamiento propuesto.

En principio, es de resaltar que el modelo seleccionado fue sometido a pruebas estadísticas sobre los valores estimados como sobre los términos del error, siendo el válido el modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial para las variables con mayor significancia (VMS), una vez adoptado el modelo se generaron los límites dialectales (la superficie dialectal continua y las isoglosas), no obstante fue preciso realizar un agrupamiento entre las distintas localidades y su distancia lingüística, para ello se utilizó el algoritmo de clasificación desarrollado en minería de datos espaciales llamado regionalización con agrupamiento y partición aglomerativa con restricciones dinámicas (REDCAP) un método que según Guo (2008) es una ampliación de los métodos de agrupación jerárquica de clústeres de vinculo único (SLK), vínculo promedio (ALK), vínculo completo (CLK) y

de Ward, reforzando la contigüídad espacial de los clústeres y permitiendo obtener un conjunto de regiones que optimizan la heterogeneidad general, en el caso particular se utilizó el agrupamiento de enlace completo (CLK) y su restricción de orden completo Full-OrderCLK cuyas pruebas en relación con los otros métodos resultaron ser sobresalientes.

La implementación del algoritmo REDCAP, base para la clasificación dialectal propuesta, se realizó en el software GeoDa, a partir de este algoritmo se generaron particiones de 2, 5, 8, 12, 16, 20 y 24 grupos, dentro de ellos se verificó que la cota mínima tuviese un mínimo porcentaje de localidades, en el caso de 2 grupos el 41 %, para 5 el 16 %, para 12 el 5 % y de esta manera para los demás grupos, garantizando mayor uniformidad en la clasificación, adicionalmente se calculó la medida de homogeneidad de cada grupo, entre los resultados se destacan los valores de grupos con mínima y máxima variabilidad de la siguientes forma:

REDCAP - 2 Clusters	**Grupo**	**Valor**	**Total**
Min SSD	1	63.5764	*103,95*
Max SSD	2	40.3741	
REDCAP - 5 clusters	**Grupo**	**Valor**	**Total**
Min SSD	4	2.89634	*38,6829*
Max SSD	3	16.7546	
REDCAP - 8 clusters	**Grupo**	**Valor**	**Total**
Min SSD	7	0.513213	*37,9579*
Max SSD	1	16.7546	
REDCAP - 12 clusters	**Grupo**	**Valor**	**Total**
Min SSD	6	0.426529—	*33,7477*
Max SSD	3	0.520584	
REDCAP - 16 clusters	**Grupo**	**Valor**	**Total**
Min SSD	14	0.156865	*28,5488*
Max SSD	12	9.59849	

Cuadro 57: SSD mayor y menor por particiones REDCAP

En el cuadro 57 la regionalización cuyo resultado fueron dos grupos muestra que la concentración de la variabilidad en su mayoría está en el grupo dos, región ubicada en el pacífico y el caribe, y clasificada dentro de la propues-

ta como zona Costeña, sin embargo, cuando de la regionalización resultan 5 grupo la mayor variabilidad se concentra en el nororiente en los departamentos de Arauca, Casanare, Santander y Norte de Santander, regiones con baja distancia lingüística, es preciso resaltar que el SSD disminuye cuando existe un mayor número de grupos, lo cual es coherente al ser esta una medida de homogeneidad.

Cuando el resultado del algoritmo REDCAP son ocho grupos la variación máxima se concentra en el centro del territorio, departamentos de Cundinamarca, Tolima, Caldas, entre otros, por lo que el resultado supone hasta el momento que la varición de los clusters cambia de región cuando hay un mayor número de grupos, sin permanecer en un sitio particularmente, no obstante los grupos de 12 y 16 clusters se concentran en la región nororiental, y dado que los resultados son diversos no es preciso hacer manifestar alguna conclusión.

El algoritmo REDCAP marcó la base para la clasificación dialectal propuesta cuya validación se realizará teniendo presente algunas propuestas realizadas por otros autores, es importante aclarar que los datos son netamente léxicos y hacen parte únicamente del tomo III del ALEC, esto difiere con algunas de las otras propuestas, por lo que se relacionará de la forma más coherente dependiendo cada una de ellas.

Luis Flórez, 1961: La clasificación realizada por Florez (1961) tiene presente seis divisiones dialectales, la primera de ellas agrupa el caribe y el pacífico, una región similar a la zona que en esta propuesta se denominaría región Costeña (Figura 100), no obstante la región propuesta es de menor tamaño dado que se presentan zonas de transición en el norte cerca a los departamentos de Antioquia, Bolívar, Córdoba y Cesar, y en el Sur en los departamentos de Nariño, Cauca y Valle del Cauca.

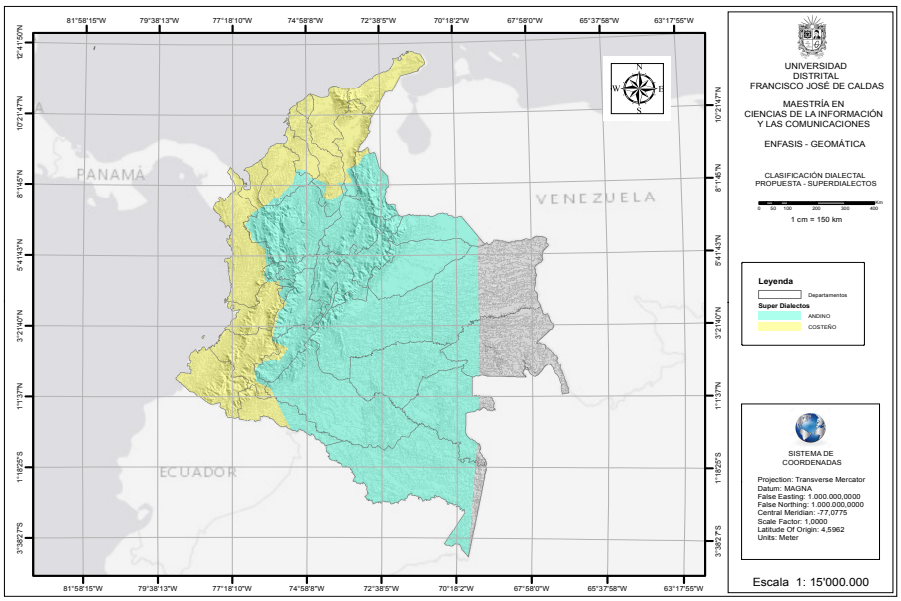


Figura 100: Superdialectos

La clasificación propuesta por Florez (1961) es en esencia diferente puesto que se centra en un agrupamiento no jerárquico y en el análisis de un número menor de localidades, para entonces 53, de tal manera que la comparación solo puede hacerse en el nivel dialectal de esta investigación, un nivel aún general; en los resultados del desarrollo de este estudio la región Andina Oriental recoge la mayor parte de localidades, un total de 112, el 47 % del total, así mismo abarca la mayor parte de los dialectos propuestos por Flórez, el principal, aquel calificado como Andino, entre los departamentos de Boyacá y Cundinamarca, que en la jerarquía de esta investigación se encuentra en el nivel Subdialectal y nombrado como Cundinamarques, y Cundiboyacense, de igual forma existe coincidencia en las regiones dialectales restantes, tal es el caso de la región de Transición Pacífico Andino en el sur del país en el oriente de los departamento de Nariño, Cauca y Valle del Cauca, la última relación equivalente reside en la zona Andina Occidental, una región sin denominación en la clasificación de Flórez y ubicada entre los departamentos de Antioquia, Caldas y parte de los departamentos de Risaralda y Quindio, cuyo nombre propuesto en el nivel subdialectal es el Eje Cafetero, Antioqueño Risaraldense y Antioqueño Caldense; en conclusión, la propuesta de esta investigación se ajusta a algunas regiones de aquella realizada por Flórez en el año 1961,

en niveles que pueden entenderse como semidetallado y detallado, a pesar de sólo utilizar datos léxicos difiriendo con Flórez, quien utilizó fenómenos gramaticales, lingüísticos y etnográficos.

José Joaquín Montes, 1982: La segunda propuesta, realizada por Montes (1982), propone de forma equivalente la bipartición del español de Colombia en dos superdialectos, Costeño y Central o Andino, planteados como el producto de factores históricos de propagación de las distintas hablas de España y estudiados a niveles léxicos, fonéticos, entre otros, permitiendo concluir en primera instancia la existencia de similitud en ambas propuestas (la de Montes (1982) y la resultante de la presente investigación) en el nivel general, sin embargo es posible visualizar una menor área en la de propuesta de este estudio, especialmente en el norte del departamento de Antioquía así como una disminución de esta misma propiedad en el sur del territorio, departamentos de Nariño, todo esto como producto de la comparación de ambos agrupamientos.

En el nivel Dialectal o semidetallado Montes (1982) define cuatro grupos, en la zona Costeña el dialecto Pacífico y el Caribe, clusters equivalentes a aquellos definidos dentro de esta investigación, no obstante dentro de este estudio se mencionan dos áreas dialectales de transición ubicadas tanto en el Caribe como en el Pacífico, que dentro de la clasificación de Montes (1982) no están; en la zona Andina o central al igual que la presente clasificación definió una zona Oriental y una zona Occidental, zonas que en el nivel detallado son similares en ambas propuestas, por lo que se puede concluir que en esencia la configuración espacial y la demarcación dialectal son equivalentes en ambos planteamientos.

Cabe resaltar que dentro de la clasificación de Montes (1982) aparece la primera definición del subdialecto llanero puesto que como se advirtió atrás las investigaciones del ALEC visitaron por su época departamentos como Arauca, Casanare y Meta, según Montes (1982) en los llanos orientales hay ciertos rasgos del Costeño que en nuestro caso aparecerían en el Llanero - Nororiental y desaparecen en localidades como Restrepo y Tauramena, Llanero - Suroriental dentro de nuestra propuesta, no obstante las isoglosas generadas tanto como la superficie dialectal continua no parecen ilustrar dicha

afirmación.

Mora et al., 2004: La propuesta de Mora et al. (2004) estableció el precedente inicial de clasificación mediante la determinación de isoglosas en 1523 mapas del ALEC, la agrupación de dichas isolíneas permitió la clasificación del idioma español de Colombia en dos superzonas, la Costeña y la Andina, la Costeña dentro de los departamentos del Chocó, Córdoba, Sucre, Bolívar, Atlántico, Magdalena, Cesar y Guajira en la parte norte y la aparición de parte de algunos otros como el Santander, Santander, Antioquia, Casanare, Arauca, Valle del Cauca y Cauca, zonas de transición dentro de este estudio, y la región Andina con departamentos como Antioquia, Caldas, Valle del Cauca, Boyacá, Tolima y Cauca, estas dos regiones son equivalentes a la propuestas de esta investigación, sin embargo Mora no presenta algunas regiones como zonas de transición pero manifiesta que hay departamentos que tienen alguna intervención parcial en este fenómenos objeto de análisis.

Es de resaltar que la propuesta de Mora et al. (2004) se basa en el trazo de isoglosas léxicas, no obstante se presentaron algunas criticas al respecto, Lancheros Redondo (2018) por ejemplo menciona que si bien la propuestas tiene como fundamento el nivel léxico de la lengua, Mora justifica algunas categorías de su clasificación con aspectos de tipo gramatical y etnográfico, elementos que no son idóneos para el trazo de isolíneas, un ejemplo de ello es la clasificación que hace del costeño pacífico, donde incluye el mapa ≪juegos con bolas≫, según Lancheros al parecer las observaciones recolectadas para este mapa no se refieren siempre al mismo juego y por lo tanto puede que exista algún tipo de error en tal clasificación.

A nivel dialectal Mora et al. (2004) mantiene en la región Costeña los niveles Caribe y Pacífico y en el Andino las categorías Oriental, Occidental, adicionalmente agrega el Llanero, dialecto que a su vez se divide en Llanero Norte y Sur, categorías similares a las desarrolladas en este estudio en un nivel más detallado, sin embargo de nuevo, al igual que Montes, no propone alguna zona de transición entre los dos superdialectos, lo cual difiere de este estudio.

Finalmente, en el nivel subdialectal Mora et al. (2004) hace referencia a

14 categorías, que son en esencia más limitadas que las propuestas en esta investigación, no obstante se encuentran entre ellas zonas como el Pacífico Norte, Pacífico Sur, Antioqueño, Santandereano, Norte Santandereano, regiones definidas de forma equivalente en este estudio, quizás la diferencia notable es que el Nariñense dentro de nuestros resultados se encuentra en una zona de transición entre el superdialecto Pacífico y Andino, región dialectal que Mora ubica únicamente en la parte Andina del territorio.

Ávila, 2015: Avila et al. (2015) realizó una nueva propuesta de clasificación dialectal del español de Colombia, para su análisis incluyó métodos dialectométricos sobre cien mapas con información léxica del ALEC, de forma similar al desarrollo de esta investigación, el análisis de valores de similitud lingüística le permitió validar las propuestas de Mora et al. (2004), Montes (1982) y Florez (1961), no obstante dentro de ellas incluyó dos superdialectos: superdialecto Caribe y superdialecto Andino (el resto del país), es decir incluye dentro de la zona andina el pacífico colombiano, su justificación se basa en el estudio de los metadatos de los informantes tales como lugar de proveniencia, edad oficio, una diferencia notable con este desarrollo, ya que la presente investigación muestra que la zona costeña tanto del Pacífico como del Caribe constituyen una superzona dialectal, siendo disímil en esencia a la región central o Andina, de tal manera que en conclusión no se ajustan en nivel general ambas propuestas.

La propuesta de Avila et al. (2015) afirma que se pueden replantear en el nivel léxico los dialectos de Mora et al. (2004) y Montes (1982) (oriental y occidental) en forma vertical, según Avila et al. (2015) la región norte estaría compuesta por las hablas Antioqueño-Caldense y Santandereano, y la parte sur por hablas como el caucano-valluno, andina, tolimense-huilense, cundiboyacense y llanera; la clasificación de Avila et al. (2015) entonces permite concluir que el nivel superdialectal y dialectal difiere de aquellas presentadas por Mora et al. (2004) y Montes (1982), y por lo tanto de la desarrollada en este estudio, no obstante en el nivel subdialectal existen hablas similares que se encuentran ubicadas en las mismas zonas y dejan ver la existencia de cierta correspondencia siendo así el nivel detallado el de mayor similitud entre ambas clasificaciones.

Johnatan Bonilla, 2019: Una de las propuestas más reciente es la de Bonilla (2019), como se mencionó su investigación fue realizada sobre doscientos mapas léxicos del ALEC escogidos de forma aleatoria y procesados en el software dialectométrico Gabmap, tomando como medida de similitud lingüística entre las variables léxicas el índice relativo de identidad y e plasmando finalmente los resultados cartográficamente empleando métodos de agrupamiento difuso.

Entre los resultados más relevantes de su investigación Bonilla (2019) presenta como la división del español de Colombia se puede hacer en tres superzonas, andina centro-oriental, andina centro-occidental y costeño, zonas que en general se ajustan a la propuesta dialectal de este estudio (Figura 102), sin embargo en nuestra investigación fue el algoritmo REDCAP el que arrojó una marcada división entre la región central u Andina y la región costeña con zonas de transición (Figura 101), por lo que los resultados aunque similares difieren en esencia, la justificación de Bonilla en este caso es sustentada por el resultado del algoritmo de agrupamiento utilizado y por la extensión de cada grupo en el espacio, justificación equivalente a la de este desarrollo.

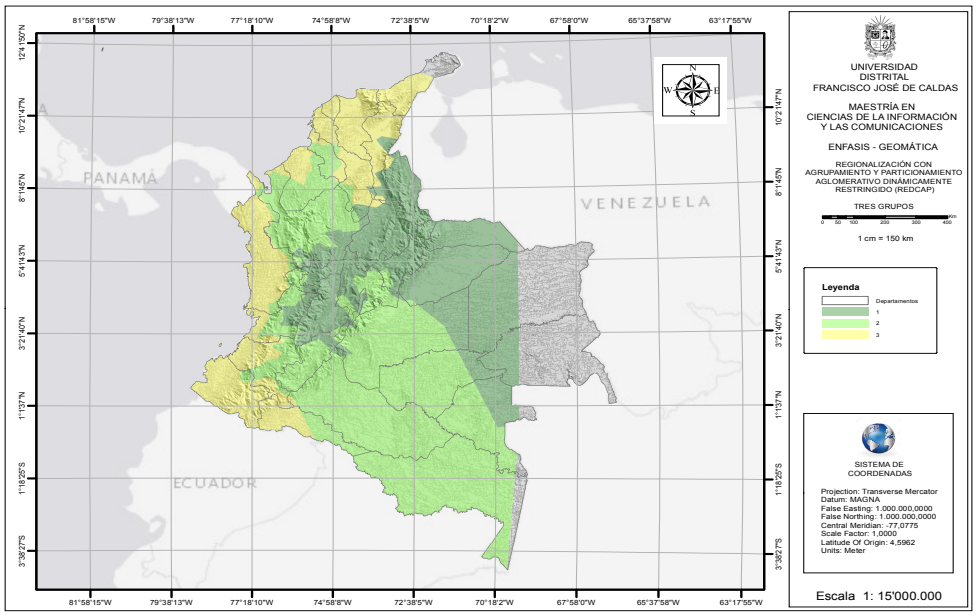


Figura 101: Tres divisiones mediante agrupamiento REDCAP

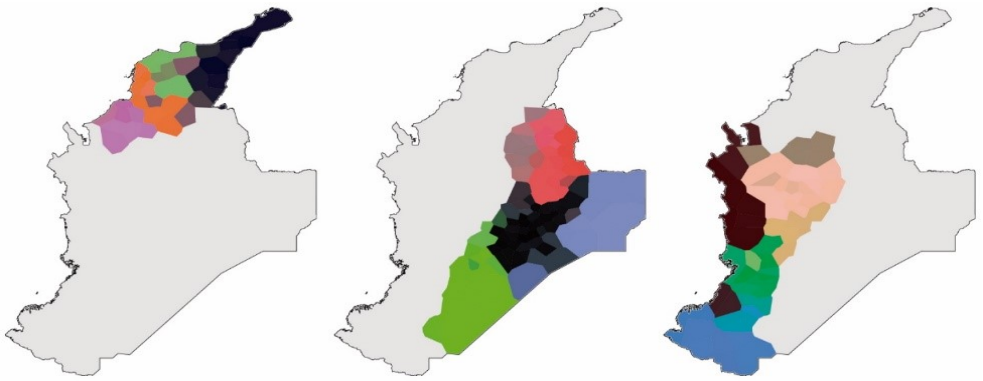


Figura 102: Tres divisiones mediante agrupamiento difuso (Bonilla, 2019, pg. 19)

Uno de los planteamientos más llamativos de Bonilla (2019) es la hipótesis que realizó acerca de la relación entre las vías históricas de Colombia y su uso como rutas de difusión lingüística, hipótesis que se verificó y planteo sobre las vías de acceso a cada localidad en este desarrollo, probando su significancia en un modelo de mixto autorregresivo espacial, aunque se saben son influyentes y se prueba dicha hipótesis en esta investigación es necesario aún realizar una revisión más profundo acerca de su desarrollo a nivel histórico.

En el nivel dialectal la propuesta de Bonilla (2019) redistribuye cada uno de los dialectos de acuerdo con los relieves observados en los resultados de su agrupamiento, con una notable diferencia en número de aquellas que proponen Mora et al. (2004), Montes (1982) y la presente investigación, con cuatro, cinco y seis dialectos respectivamente, Bonilla (2019) adiciona dos regiones, con zonas de transición entre el Costeño Pacífico y el Andino, de la misma forma como se hace en esta investigación, es importante resaltar que algo adicional en común es que no encuentra los denominados costeñismos parciales relacionados entre el Caribe y otras regiones propuestos por Montes (1982), mencionando que hay una clara distinción entre estas regiones y que el algoritmo no permite tal agrupamiento; en conclusión el nivel superdialectal y dialectal de Bonilla (2019) son más detallados que las demás propuestas y que esta investigación, no obstante no se menciona particularmente la norma o estructura espacial mínima de extensión (en área) que permite considerar una región como superdialecto o dialecto respectivamente, en esta investigación al contrario se mencionaron cotas mínimas para el agrupamiento en

función de la medida de heterogeneidad espacial que fueron implementadas en el algoritmo REDCAP y permitieron dicha clasificación propuesta.

Por último en el nivel subdialectal Bonilla (2019) presenta 22 grupos con sus respectivas localidades que en contraste con otros autores es en cierto sentido más detallado, la diferencia notable con la de este estudio es que Bonilla (2019) no incluye dentro de su clasificación la localidad del Amazonas y por lo tanto la relación con esta observación es nula, lo llamativo quizás de nuestro estudio y la clasificación de Bonilla (2019) es el análisis de continuidad, es decir la perspectiva bajo la cual la lengua se distribuye de forma continua en el territorio y por lo tanto permite la existencia de zonas de transición, consecuentes con la superficie dialectal generada; en conclusión la propuesta de Bonilla (2019) es la más cercana a los resultados obtenidos y a la propuesta realizada, sin embargo difiere en el nivel de detalle y en el número de observaciones utilizadas que generaron un resultado diferente en el sur del país, especialmente en los departamentos de Nariño, Cauca, Amazonas, Huila y Caquetá.

Néstor Ruíz, 2020: La clasificación dialectal más reciente es la de Ruíz (2020) realizada con la integración de datos lingüísticos y dialectologícos, su propuesta se fundamente en la elaboración e incorporación de un modelo espacial basado en el concepto de escala geográfica, incluyéndola en la noción de superdialecto, dialecto y subdialecto, la escala toma como base una distribución normal, distribución que quizás pueda deberse a la forma como se distribuye la diferencia o distancia lingüística dentro del territorio colombiano (tal como se verifico en esta investigación), no obstante Ruíz (2020) manifiesta que el intervalo de esta escala (0, 100 %) representa el total de la superficie del territorio y a partir de ella puede designar subintervalos para cada grupo dialectal, sin embargo no especifica el porque selecciona dicha distribución para tal análisis más que para relacionar la variación espacial con un modelo estadístico básico.

El resultado de Ruíz (2020) muestra tres superdialectos, el primero de ellos es el Español Antillano que se extiende en la región Caribeña, segundo el Español andino en el sur del territorio y tercero el denominado Español Neogranadino en el centro del país (Figura 7), es preciso resaltar que dentro

del Español Neogranadino se encuentra el Pacifico, similar a lo que expone Avila et al. (2015), de esta manera los superdialectos definidos por Ruíz (2020) son en esencia distintos a los propuestos en este estudio y aunque resalta que su investigación se basa en clasificaciones anteriores propuestas como la de Bonilla (2019), Avila et al. (2015) y Mora et al. (2004), no hace referencia al método espacial o de análisis que permite la incorporación de dicha información en el trazado de los límites dialectales desde una cartográfica.

En el nivel dialectal Ruíz (2020) define cinco clases, el Caribeño colombiano que es a su vez es el único subgrupo de Español Antillano, el Español Neogranadino Occidental, Español Neogranadino Oriental y el Español Neogranadino Amazónico del superdialecto Español Neogranadino y finalmente en la parte sur el Andino colombiano del superdialecto Español Andino, resaltando que la única subdivisión para la clasificación superdialectal se hizo sobre el superdialecto central o andino, de tal manera se puede concluir que el sur y el norte del territorio comparten la misma área en nivel dialectal y superdialectal, así mismo esta clasificación constituye en el nivel general y semidetallado la de menor similitud en comparación con las propuestas de otros autores señaladas en esta investigación, no obstante hay un acercamiento mayor en el nivel subdialectal, ya que Ruíz como base para este nivel tomo datos léxicos del ALEC e investigaciones como la de Bonilla (2019), Mora et al. (2004), Avila et al. (2015), entre otros, de las cuales extrajo los límites para estas unidades espaciales.

Capítulo IV
Conclusiones

En virtud de la reciente divulgación del ALEC Digital mediante el uso de herramientas geoespaciales se llevó a cabo un análisis cualitativo y cuantitativo de la información lingüística (léxica) contenida en el tomo III del ALEC, de algunos metadatos e información geográfica; haciendo uso de técnicas dialectométricas, de econometría espacial, geoestadísticas y de minería de datos y en relación al objetivo general, se logró diseñar un modelo espacial para la estimación de isoglosas y clasificación dialectal que permitió integrar de forma cuantitativa la interrelación del dialecto con el espacio geográfico, validando así las hipótesis de clasificación propuestas por autores como Florez (1961), Montes (1982), Mora (1996) y Bonilla (2019) en el nivel superdialectal, dialectal y en los dos últimos casos, Mora (1996) y Bonilla (2019), en el nivel subdialectal.

Dentro del desarrollo metodológico se evaluó la información contenida en el tomo III del ALEC, logrando incluir dentro de la investigación variables referentes a los metadatos, tales como altura sobre el nivel medio del mar, temperatura, vías de acceso y actividades económicas, así mismo se incorporaron variables de tipo geográfico, teniendo presente el planteamiento hecho por Hernández Campoy (1999) que explica como factores físicos de tipo espacial intervienen de forma fundamental en la diferenciación de la lengua, su propagación y su cambio, en tal sentido fueron seleccionadas las variables aptitud agroclimática, índice de disponibilidad hídrica, riesgo por amenaza sísmica, precipitación y distancia geográfica (distancia geográfica media entre localidades), siendo en total nueve variables las seleccionadas inicialmente, todas ellas precisadas en un periodo temporal cercano a aquel en que se realizaron las encuestas y regionalizadas en el espacio geográfico referente a cada una de las localidades.

La esencia lingüística dentro de esta investigación fue la medida de proximidad isocrática o medida de similitud no ponderada propuesta por Goebl

(1987), el Índice Relativo de Identidad (IRI), medida utilizada por Bonilla (2019) para la generación de su propuesta de clasificación dialectal; a diferencia de Mora et al. (2004), quienes presenta una a una las variables léxicas y expone su distribución espacial a partir de isoglosas, este estudio realizó el cálculo automático de una matriz cuadrada de distancias a partir del análisis de nulidad sobre 100 mapas estrictamente léxicos del tomo III del ALEC; el análisis de esta variable dentro del territorio permitió determinar que la configuración espacial de los valores mas altos se encuentran cerca de la Costa Caribe y Pacífica, lo que en la propuesta de clasificación dialectal fue denominado Costeño, en tal sentido la menor distancia lingüística se extiende a lo largo de la zona Central, incluyendo la localidad ubicada en Leticia, departamento del Amazonas y denominado a nivel superdialectal dentro de la clasificación propuesta como Andino; de esta manera y luego de analizar la distancia geográfica en función de esta medida de similitud se corroboro la relación de proporcionalidad inherente entre ambas variables en la mayor parte de la región objeto de estudio.

La evaluación principal explicada por el índice de autocorrelación espacial (I de moran) sugirió que las variables: actividad económica, aptitud agroclimática, precipitación, índice de disponibilidad hídrica 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 variable (en este caso la variable lingüística) en espacios vecinos (Flint et al., 2000), el resultado de este diagnóstico de dependencia geográfica fue verificado de forma cuantitativa dentro de los múltiples modelos espaciales generados.

Las causas de la correlación espacial se identificaron a partir de la existencia de varios fenómenos de interacción geográfica, según Anselin (1988) todo aquello que ocurre en un espacio geográfico afecta espacios vecinos, por ello se consideró el diseño una matriz de pesos espaciales y/o matriz de retardo espacial basada en contigüidad, el tratamiento de esta interacción para su análisis y observación fue incorporada dentro de un modelo geográfico, para esta investigación se hizo uso de tres variantes de modelos espaciales,

el primero de ellos, el modelo de regresión con dependencia espacial en la perturbación aleatoria o modelo de error espacial, en segundo lugar el modelo mixto autorregresivo de regresión espacial o modelo del retardo espacial y finalmente el modelo mixto de regresión espacial con perturbaciones aleatorias autorregresivas (SARAR); el diseño de cada uno de ellos se hizo en virtud de la teoría propuesta en econometría espacial, teniendo presente las pruebas de hipótesis de normalidad realizadas sobre la distancia lingüística y con el propósito de tener una perspectiva multidireccional en el momento de seleccionar el modelo idóneo, la implementación y desarrollo de cada uno de ellos en etapas posteriores permitió identificar patrones de significancia sobre distintas variables, revelando así características comunes que posibilitaron realizar una mejor selección.

Aunque existen múltiples aplicaciones informáticas con herramientas estadísticas integradas para el desarrollo, diseño e implementación de un modelo espacial para la estimación de isoglosas y clasificación dialectal, dentro de esta investigación se utilizaron seis de ellos: Gabmap, una aplicación web de libre acceso para el análisis de información lingüística, que permitió el calculo de la matriz de diferencias lingüísticas, Python, un lenguaje de programación que posibilitó la automatización de tareas y validación de procesos, GeoDa, una herramienta de código abierto para el procesamiento y modelamiento de datos geográficos , R, un lenguaje y a la vez un entorno de computación estadística que facultó la implementación del modelamiento espacial, ArcGIS, un sistema informático para la recopilación, organización, administración, análisis y divulgación de información espacial y finalmente SpaceStat, un software desarrollado para el análisis de información geográfica dentro de econometría espacial; todas estas herramientas facultaron no solo el desarrollo e implementación de un modelo espacial sino también la validación constante de procesamientos y análisis, de tal manera que los resultados corresponden a una prueba fehaciente del mismo desarrollo en múltiples plataformas informáticas.

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 según los contrastes de dependencia espa-

cial propuestos por Anselin (2005) permitió rechazar la hipótesis nula de inexistencia de correspondencia geográfica, en él se incluyen un total de ocho variables, siendo las de mayor influencia, las categorías de la variable Vías de acceso, validando parcialmente la hipótesis de Bonilla (2019) acerca de estas rutas como medio de difusión lingüística, en este sentido cabe resaltar que la concentración de infraestructura nacional vial fue en la región Andina, y sólo hacia los años 60' la proporción de carreteras aumentó en departamentos que no habían sido tenidos en cuenta (Pachón, 2006), de esta manera tanto el modelo propuesto como hechos históricos son prueba de la concentración no solo del lenguaje en el centro del territorio sino de su relación espacial inherente con actividades económicas y procesos de movilización y/o comunicación vial.

Otro de lo fenómenos históricos estudiados dentro del desarrollo y análisis de resultados de la investigación fue aquel relacionado con las vías de comunicación (vías de acceso), aquel relativo a los movimientos demográficos internos a principios del siglo XIX, la verificación al respecto constato que existieron movimientos de migración interna en la zona central especialmente, región que para entonces tenía el control real del territorio, dominando las relaciones de mercado, las instituciones estatales y dinámicas socioculturales (Mejía, 2013) , tales dinámicas dieron lugar a fenómenos como la expansión de la frontera interna y a la migración hacia las grandes ciudades, esta última a razón de políticas gubernamentales y al modelo de desarrollo capitalista adoptado; la representación cartográfica de Mejía (2013) de estas manifestaciones se ajustó en forma general a algunos de los límites lingüísticos trazados dentro de la presente investigación.

Si bien se mencionó como 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, permitiendo concluir que existe una relación medible y/o cuantificable entre variables léxicas y variables geográficas, cuyas principales categorías corresponden a variables tales como la aptitud agroclimática, la precipitación, el índice de disponibilidad hídrica y la distancia geográfica, posibilitando así, la estimación de un modelo óptimo para la generación de isoglosas, y la delimitación precisa y detallada de zonas dialectales.

Los límites lingüísticos generados con el modelo espacial seleccionado fueron las isoglosas y la superficie dialectal continua, ambas muestran de forma discreta y continua respectivamente la distancia lingüística entre las distintas localidades cartografiadas del ALEC, es preciso resaltar que ambas proceden del mismo origen, el método de interpolación Kriging Ordinario implementado sobre los valores estimados del modelo seleccionado; los trazos de estas dos superficies permitieron referenciar zonas con alta similitud lingüística , tal es el caso de la zona central, dentro de los departamentos del Tolima, noroccidente de Cundinamarca, Boyacá, Casanare, Arauca, Santander y Norte de Santander, la zona oeste en la mayor parte de los departamentos del Pacífico y el Caribe como el oeste de Nariño, Cauca, Valle del Cauca, Chocó, Antioquia, Córdoba, Súcre, Bolívar, Magdalena, Cesar, Atlántico y La Guajira, y finalmente una zona suroriental en los departamentos del Meta, Caqueta, Huila, Putumayo y Amazonas.

La clasificación dialectal propuesta se llevó a cabo a través de la implementación del algoritmo de regionalización con agrupamiento y particionamiento aglomerativo dinámicamente restringido (REDCAP) propuesto por Guo (2008) en minería de datos espaciales, siendo una ampliación de métodos de agrupación jerárquica, sin embargo a diferencia de ellos este refuerza la contigüidad espacial de los clusters y permiten obtener un conjunto de regiones con mayor heterogeneidad geográfica, así como equilibrio en el tamaño de la región y menor variación interna Guo (2008); la clasificación final considera un total de dos superdialectos: Costeño y Andino, seis dialectos, que involucran zonas de transición entre las dos regiones de mayor jerarquía: Transición Andino – Costeño Caribe y la Transición Pacífico – Andino, y así mismo incluye treinta y dos subdialectos, resaltando la inclusión de la relación entre las localidades de Leticia y Caquetá en un subdialecto denominado Amazónico – Caqueteño; es importante resaltar que regiones como la Orinoquía y la Amazonía carecen de datos, por lo tanto su agrupación en tal instancia representa en su mayor parte una conjetura espacial la cual puede verificarse en el estudio dialectal detallado.

El diseño e implementación del protocolo de pruebas del modelo espacial contempló en su primera instancia el análisis de los valores estimados y la evaluación de los términos del error a través de la implementación de

estadísticos diagnóstico, una vez seleccionado el modelo, la segunda prueba se ejecutó en virtud del contraste de los resultados de la propuesta de clasificación dialectal y aquellas realizadas por otros autores, resaltando que para ello se integraron los limites lingüísticos generados como complemento determinante en dicho proceso, de esta segunda evaluación se logró concluir que las propuestas de mayor semejanza son la de Florez (1961), Montes (1982), Mora et al. (2004) y Bonilla (2019), todas ellas manifestando la bipartición del español en términos generales en Costeño (Costeño Pacífico y Costeño Caribe) y Andino, con variaciones mínimas en el nivel dialectal y en el caso de Bonilla (2019), con una coincidencia notable en el nivel subdialectal; con respecto a las propuestas presentadas por Ruíz (2020) y Avila et al. (2015), las diferencias son notables en el nivel superdialectal, con pequeñas coincidencias en niveles más detallados, esto puede deberse a la naturaleza de la información utilizada así como el tipo de análisis realizado.

Finalmente, el modelo espacial para la generación de isoglosas y clasificación dialectal desarrollado dentro de esta investigación se ajusta a fenómenos de tipo económico y sociodemográficos presentados a principios del siglo XIX, así como a la mayor parte de resultados de clasificación dialectal realizados por otros autores; la incorporación de mecanismos geoespaciales y algoritmos de ciencias de la computación (minería de datos) posibilitó la generación de nuevo conocimiento, quedando demostrada la relación cuantitativa entre el espacio geográfico y fenómenos lingüísticos, no obstante es menester de investigaciones posteriores 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

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Springer.

Anselin, L. (2001). A companion to theoretical econometrics. *Spatial econometrics*, pages 310–330.

Anselin, L. (2005). Exploring spatial data with geoda: A workbook. *Urbana*, 51.

Anselin, L., Bera, A. K., Florax, R., y Yoon, M. J. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26(1):77 – 104.

Anselin, L. y Florax, R. J. G. M. (1995). *New Directions in Spatial Econometrics*, pages 3–18. Springer Berlin Heidelberg.

Anselin, L. y Rey, S. (2014). *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Anselin, L. y Smirnov, O. (2006). Efficient algorithms for constructing proper higher order spatial lag operators*. *Journal of Regional Science*, 36(1):67–89.

ArcGIS (2020). Arcgis resources. *ESRI*.

Avila, Y., Mendieta, F., Álvarez, A. C. R., Rodríguez, C., y Silva, F. (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: Axac*, 1:107 – 125.

Barbary, O. y Urrea, F. (2003). La población negra en la Colombia de hoy: dinámicas sociodemográficas, culturales y políticas. *Estudos Afro Asiáticos*, 25:9 – 21.

Barreiro, J., Ruzo, E., y Losada, F. (2004). Modelo logit multinomial y regresion con variables ficticias: una aplicacion regional al sector lacteo. *University of Santiago de Compostela*.

Bohorquez, I. y Ceballos, E. (2008). Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. *Ecos de Economía*, 12.

Bonilla, J. E. (2019). Propuesta de división dialectal del español de colombia con base en el análisis dialectométrico de datos léxicos del alec (tesis de maestría). *Instituto Caro y Cuervo*, 1:1–35.

Burridge, J. (2017). Spatial evolution of human dialects. *Physical Review X*, 7.

Burridge, P. (1980). On the cliff-ord test for spatial correlation. *University of Kent at Canterbury*, 42:107–108.

Burrough, P. y McDonnell, R. (1998). Principle of geographic information systems. *OUP Oxford*.

Chacón, C. (2016). Análisis demolingüístico del léxico hispánico: Estudio aplicado a las "nociones específicas" del plan curricular del instituto cervantes. *Universidad Nacional de Educación a Distancia Escuela de Doctorado*.

Chambers, J. K. y Trudgill, P. (1998). *Dialectology*. Cambridge University Press, 2 edition.

Coro, Y. (2005). Análisis exploratorio de datos espaciales al servicio del geomarketing. *III Seminario sobre Nuevas Tecnologías en la Investigación - Universidad Autónoma de Madrid*.

Cressie, N. (2015). *Statistics for Spatial Data*, pages 1–26. John Wiley Sons Inc.

Córdoba, G. A. (2011). Geografía, lingüística y geolingüística. una propuesta para comprender el contacto dialectal. *Forma y Función*, 24(1):47–60.

Dubert, F. y Sousa, X. (2016). On quantitative geolinguistics: An ilustration from galician dialectology. *Instituto da Lingua Galega - Universidade de Santiago de Compostela*, 2016.1:1–31.

Emery, X. (2013). Geoestadística. *Universidad de Chile*, 1:4 – 149.

Esenbuğa, O. y Çolak, E. (2016). Comparison of principal geodetic distance calculation methods for automated province assignment in turkey. *Istanbul Technical University*.

Everett, C., Blasi, D., y Roberts, S. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences of the United States of America*, 112.

Fernandez, J. (2017). Diseño e implementación de un sistema de información geográfica en linea para el atlas lingüístico-etnográfico de colombia, su manual y suplemento (tesis de pregrado). *Universidad Distrital Francisco José de Caldas*, pages 1–139.

Flint, C., Harrower, M., y Edsall, R. (2000). But how does place matter? using bayesian networks to explore a structural definition of place. *The New Methodologies for the Social Sciences Conference*.

Florez, L. (1961). Nota informativa del atlas lingüístico-etnográfico de colombia (alec). *THESAURUS*, 1(1):77–125.

Florez, L. (1983). *Manual del Atlas Lingüistico-Etnográfico de Colombia*. Manual del ALEC. Instituto Caro y Cuervo.

Geoda (2020). Introducing geoda. *University of Chicago*.

Giraldo, R. (2015). Introducción a la geoestadística. *Universidad Nacional de Colombia - Departamento de Estadística*.

Goebl, H. (1982). Dialektometrie: Prinzipien und methoden des einsatzes der numerischen taxonomie im bereich der dialektgeographie. *Wien: Osterreichische Akademie der Wissenschaften*, 1:4–12.

Goebl, H. (1987). Points chauds de l'analyse dialectométrique : pondération et visualisation. *Revue de linguistique romane*, 1:1 – 57.

Goebl, H. (1991). Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data. *Revue de linguistique romane*, 1:277 – 324.

Goebl, H. (2006). Recent advances in salzburg dialectometry. *Literary and Linguistic Computing*, 21:1–25.

Goodchild, M. F. (1987). A spatial analytical perspective on geographical information systems. *International Journal of Geographical Information Systems*, 1(4):327–334.

Granados, J. (2010). Las migraciones internas y su relación con el desarrollo en colombia: Una aproximación desde algunos estudios no clasificados como migración interna de los últimos 30 años. *Pontificia Universidad Javeriana - Facultad de estudios ambientales y rurales, Maestría en desarrollo rural*.

Grieve, J. (2018). Spatial statistics for dialectology. *The handbook of dialectology*, 1:415 – 434.

Grieve, J., Speelman, D., y Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change*, 23(2):193 – 221.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (redcap). *International Journal of Geographical Information Science*, 22(7):801–823.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge.

Heeringa, W. (2004). Measuring dialect pronunciation differences using levenshtein distance. *Groningen Dissertations in Linguistics*, 46:20–48.

Heeringa, W. y Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, 13:375 – 400.

Hernández Campoy, J. M. (1999). *Geolingüística: modelos de interpretación geográfica para lingüistas*. Universidad de Murcia.

Herrera, C. (2017). Poblamiento histórico de casanare: reflexiones para una agenda de acción colectiva de ordenamiento democrático del territorio. *Universidad Nacional de Colombia*.

Herrera Gomez, M., Paz, J., y Cid, J. C. (2012). Introducción a la econometría espacial. una aplicación al estudio de la fecundidad en la argentina usando r. *Munich Personal Repec Archive MPRA*.

Hoch, S. y Hayes, J. (2010). Geolinguistics: The incorporation of geographic information systems and science. *Geographical Bulletin - Gamma Theta Upsilon*, 51:23–36.

Instituto Caro y Cuervo (2020). Alec. *Instituto Caro y Cuervo*.

Instituto Caro y Cuervo (2021). Misión y visión - instituto caro y cuervo. *Instituto Caro y Cuervo*.

Jeszenszky, P. y Weibel, R. (2015). *Measuring boundaries in the dialect continuum*. International Conference on Geographic Information Science.

Jiménez, A. (2004). Aptitud agroecológica para el cultivo de la caña de azúcar en el sur de Tamaulipas, México. *Investigaciones geográficas*, pages 58 – 74.

Kalmanovitz, S. (2015). Breve historia económica de colombia. *Universidad Tadeo Lozano*.

Kessler, B. (1995). Computational dialectology in irish gaelic. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 7:60–67.

Labov, W. (1995). Principles of linguistic change. *Journal of Linguistics*, 31(2):435–439.

Lancheros Redondo, H. (2018). Los indigenismos léxicos en las variedades diatópicas del español colombiano. *Forma y Función*, 31:9–29.

Leinonen, T., Çağrı Çöltekin, y Nerbonne, J. (2016). Using gabmap. *Lingua*, 178:71 – 83. Linguistic Research in the CLARIN Infrastructure.

Lichtenstern, A. (2013). Kriging methods in spatial statistics. *Department of Mathematics - Technische Universitat Munchen*.

Matheron, G. y Blondel, F. (1962). Traite de geostatistique appliquee. *Paris : Technip*.

Mejía, G. (2013). *Historia concisa de Colombia (1810-2013)*. Universidad del Rosario.

Montes, J. (1982). El español de colombia: Propuesta de clasificación dialectal. *THESAURUS*, 31(1):23–92.

Mora, S. (1996). Dialectos del español de colombia - caracterización léxica de los subdialectos andino-sureño y caucano-valluno. *THESAURUS*, L1(1):1–26.

Mora, S., Lozano, M., Ramírez, R., Espejo, M. B., y Duarte, G. E. (2004). Caracterización léxica de los dialectos del español de colombia según el alec. *Instituto Caro y Cuervo*, L1(1):1–235.

Nagelkerke, D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692.

Nerbonne, J. (2006). Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, 21.

Nerbonne, J. (2010). Measuring the diffusion of linguistic change. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365:3821–8.

Nerbonne, J. y Kleiweg, P. (2007). Toward a dialectological yardstick. *Journal of Quantitative Linguistics*, 14:148–166.

Nerbonne, J., Prokic, J., y Wieling, M. (2019). Dialectology for computational linguists. *Dialectology - Oxford Research Encyclopedia of Linguistics*, 1:1–22.

Nerbonne, J. y Wieling, M. (2017). Statistics for aggregate variationist analyses. *The Handbook of Dialectology*, pages 400–414.

Pachón, A. (2006). *La infraestructura de transporte en Colombia durante el siglo XX*. Banco de la Republica - Colombia.

Pascale, A. J. (2010). Investigaciones agrometeorológicas para el cultivo de la soja realizadas en la cátedra de climatología y fenología agrícolas de la fauba. *Revista de la Facultad de Agronomía - UBA*, page 1–73.

Python (2020). Python. Last accessed 18 March 2020.

Pérez, J. (2005). La infraestructura del transporte vial y la movilización de carga en colombia. *Documentos de trabajo sobre economía regional*, 1(64).

Pérez, J. (2006). Econometría espacial y ciencia regional. *Investigación Económica*, 65:129–160.

R (2020). The r project for statistical computing. *The R Foundation*.

Rodriguez, A. y Duarte, C. (2017). Diseño e implementación del atlas lingüístico-etnográfico de colombia web. *Universidad Distrital Francisco José de Caldas*, pages 1–78.

Rodríguez, C., Jimenez, S., Dueñas, G., Bonilla, J., y Gelbukh, A. (2018). Dialectones: Finding statistically significant dialectal boundaries using twitter data. *Computacion y Sistemas*, 22.

Ruíz, N. (2020). El español de colombia. nueva propuesta de división dialectal. *Lenguaje*, 48(2):160–195.

Salvatore, M. (2009). Metodología de evaluación de aptitud de tierras. *FAO*, page 11–19.

Servicio Geológico (2020). Sistema de consulta de la amenaza sísmica de colombia. Fecha de actualización - 27 de julio de 2020.

SpaceStat (2020). Spacestat. *BioMedware - Geospatial Research and Software.*

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory For Kriging.* Springer.

Séguy, J. (1973). La dialectometrie dans l'atlas linguistique de gascogne. *Revue de Linguistique Romane*, 33(2):1–24.

Vayá, E. y Moreno, R. (2000). La utilidad de la econometría espacial en el ámbito de la ciencia regional. *Madrid*, pages 1–49.

Vilalta, C. (2005). Cómo enseñar autocorrelación espacial. *Economía, Sociedad y Territorio.*

Wang, F. y Robert, L. K. (2015). Constructing geographic areas by redcap and mlr for analysis of homicide rates: A case study of new orleans, louisiana. *Papers in Applied Geography*, 1(4):295–306.

Webster, R. y Oliver, M. (2007). Geostatistics for environmental scientists: Second edition. *Geostatistics for Environmental Scientists: Second Edition.*

Wieling, M. (2012). A quantitative approach to social and geographical dialect variation. *PhD thesis, University Groningen*, 1:1–178.

Wieling, M. y Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics*, 1:243–264.

Wieling, M., Valls, E., Baayen, H., y Nerbonne, J. (2018). *Border Effects Among Catalan Dialects*, pages 71–97. Springer.

Yrigoyen, C. C. (2003). *Econometría espacial aplicada a la predicción-extrapolación de datos microterritoriales*. Consejería de Economía e Innovación Tecnológica, Comunidad de Madrid.