UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2017

Modelos mixtos para datos composicionales: Una aplicación con resultados electorales en Colombia

Mixed models for Compositional Data: An application with electoral results in Colombia

Autor: Juan Manuel Liscano Fierro.[a]
juan.liscano@usantotomas.edu.co

Director: Andrés Felipe Ortiz Rico.[b]
andresortiz@usantotomas.edu.co

Resumen

El presente trabajo consiste en la aplicación de ciertas herramientas desarrolladas para el análisis de los datos composicionales. El propósito incluye la revisión de los aspectos teóricos; la geometría del símplex, la metodología del log-cociente y aspectos relacionados con las componentes nulas, además del desarrollo de un ejercicio práctico teniendo en cuenta las metodologías mencionadas junto con los modelos estadísticos, tales como regresión Dirichlet, un modelo lineal multivariado y finalmente el modelo mixto multivariado, que es el eje principal del ejercicio. Se ilustra la aplicación práctica de la teoría haciendo uso de la información disponible en cuanto a los procesos electorales llevados a cabo en Colombia y otras variables que definen la situación económica y política del país.

Los resultados de los datos analizados bajo el ajuste del modelo mixto responden de la mejor manera a los valores reales del plebiscito por la paz, identificando cómo las variables trabajadas influyen en el resultado de las votaciones. Sugiriendo que los departamentos con más problemas sociales están más a favor de la paz.

***Palabras clave*:** Datos composicionales, transformaciones, modelos estadísticos, procesos electorales, símplex.

Abstract

The present work consists in the application of certain tools developed for the analysis of the compositional data. The purpose includes the revision of the theoretical aspects; the geometry of the simplex, the log-ratio methodology and aspects related to null components, as well as the development of a practical exercise taking into account the mentioned methodologies along with the statistical models, such as Dirichlet regression, a multivariate linear model and finally the multivariate mixed model, which is the main axis of the exercise. It illustrates the practical application of the theory making use of the available information about the electoral processes carried out in Colombia and other variables that define the economic and political situation of the country.

The results of the data analyzed under the adjustment of the mixed model respond in the best way to the real values of the plebiscite, identifying how the variables worked influence the results of the voting. Suggesting that departments with more social problems are more in favor of peace.

***Keywords*:** Compositional data, logratio transformations, statistical models, electoral processes, simplex.

[a]Estudiante de estadística Universidad Santo Tomás Bogotá
[b]Profesor de estadística Universidad Santo Tomás Bogotá

AGRADECIMIENTOS:

A mi abuela, María Antonia Villarreal, mis tíos, Gloria Fierro y Saúl Neira, a mis papás, Mónica Fierro y Yesid Liscano, a todos gracias por el apoyo brindado a lo largo de mi vida. Por lo ánimos y la fuerza que me ofrecieron en cada etapa y paso que me trajo hasta este momento.

A mi director de tesis, Andrés Felipe Ortiz, por aceptarme para realizar la tesis de pregrado bajo su dirección. Por el apoyo, la disponibilidad y la paciencia que tuvo durante mi formación profesional y por toda la sabiduría que compartió conmigo durante este largo proceso.

1. Introducción

Los datos composicionales son datos multivariados no negativos que suman una misma constante definida, usualmente y por propósitos de conveniencia, es tomada como la unidad. Esta clase de datos se presentan como porcentajes, proporciones, concentraciones etc., y generalmente se encuentran en disciplinas tales como geología, arqueología, economía, ciencias políticas y ciencias forenses entre otras. Las restricciones de no negatividad y de suma constante que caracterizan este tipo de datos implican que las técnicas multivariantes habitualmente utilizadas no son adecuadas para su análisis y modelización. La cuestión clave es que la geometría del espacio muestral sobre el que se definen, el símplex, es diferente a la clásica geometría Euclídea del espacio real, y este problema ha sido fuente de preocupación desde que Pearson (1896) lo puso de manifiesto.

Aunque el tratamiento de estos datos es complicado, muchos autores han intentado afrontar los problemas del análisis estadístico de los datos composicionales. La solución no aparece hasta 1982 cuando Aitchison presenta, por primera vez, una forma de evitar la restricción de la suma constante, donde argumenta que las dificultades de la interpretación vienen motivadas por centrar la atención en las magnitudes absolutas de las partes de una composición; decía entonces que el centro de atención debía ser las magnitudes relativas, es decir los cocientes. Esta metodología permite la transformación de los datos composicionales al espacio real multivariante, donde es posible aplicar cualquier técnica estadística clásica.

Existen en la literatura numerosos trabajos con diferentes objetivos. Por ejemplo, Thomas & Aitchison (1998) estudian qué óxidos son más efectivos a la hora de discriminar entre dos tipos de calizas. Tolosana-Delgado et al. (2002) presentan un análisis discriminante de basaltos y rocas afines basándose en los elementos traza presentes en los mismos. Buccianti et al. (2002) utilizan el porcentaje de los componentes químicos de los gases en fumaloras de volcanes para estudiar las constantes de equilibrio en diferentes reacciones químicas. Weltje (2002) construye regiones de confianza en el símplex para rocas detríticas. Hijazi & Jernigan (2009) modelan datos composicionales a través de modelos dirichlet.

A lo largo de la historia colombiana se han realizado dos plebiscitos; uno para otorgarle el poder al frente nacional y el otro convocado para el 2 de octubre del 2016 para consultarle al pueblo si aprobaba o no los acuerdos logrados durante el proceso de paz. Este plebiscito fue quizá el más importante de la historia colombiana, ya que traía en sí la esperanza de poner fin a una guerra de sesenta años con las FARC, y es por eso que sus resultados y la naturaleza de los mismos se convierten en el eje principal del ejercicio, porque el hecho de analizar las tres alternativas posibles (sí, no, abstención) en conjunto y no por separado, es lo que hace del análisis de datos composicionales algo idóneo para el ejercicio.

Otra de las aplicaciones que puede tener el análisis de los datos composicionales, está relacionada con los resultados de las elecciones, como lo muestra Pawlowsky-Glahn & Buccianti (2011) o Márquez et al. (2011), es posible usar los datos composicionales bajo diferentes técnicas estadísticas clásicas para este tipo de ejercicios. En este documento haremos una aplicación con los datos del plebiscito por la paz realizado en Colombia.

El presente documento cuenta con el compendio de un marco teórico en la sección 2, en donde se explican las principales características de los datos composicionales, las definiciones a tomar en cuenta para su análisis y posibles distribuciones de probabilidad que se adecuan a los datos, además de las respectivas transformaciones que hacen posible el uso de las metodologías clásicas existentes. Posterior a esto, se presentan en la sección 2.5 los tres tipos de modelos ajustados, la inferencia sobre los mismos y los resultados obtenidos. Finalmente, se encuentran resumidos los hallazgos del ejercicio en las secciones 5, 6 y 7, junto con la metodología, resumida en la sección 4, implementada para llevar a cabo el ejercicio.

2. Marco Teórico

Los métodos de análisis composicional se basan en principios de invarianza por escala. Se utilizan logaritmos de cocientes entre proporciones (transformaciones logísticas), que permiten analizar estos datos como variables reales ordinarias. En términos matemáticos, su espacio muestral, llamado símplex $\mathbb{S}^d$, está dado por:

$$\mathbb{S}^d = \left\{ (x_1, ..., x_D)^T \, | x_i \geq 0, \sum_{i=1}^{D} x_i = 1 \right\}$$

Para el caso de $D = 3$, el símplex $\mathbb{S}^3$ se representa a través del diagrama ternario (triángulo equilátero de altura K, véase la figura 1). Existe una relación uno-a-uno entre las composiciones del vector de 3 partes y los puntos del diagrama ternario; es decir, cada elemento del vector corresponde a un solo punto del triángulo. Dado esto, un dato composicional $x = (x_1, x_2, x_3)'$ se corresponde con el punto que dista x_1, x_2 y x_3, respectivamente, de los lados opuestos a los vértices 1,2 y 3.

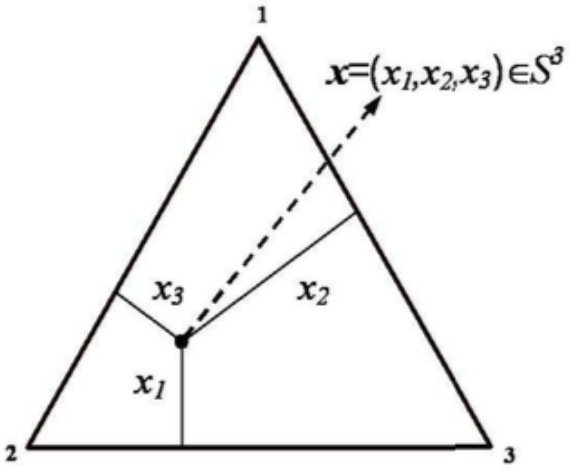

Figura 1: Representación de un dato composicional $(x_1, x_2, x_3)'$ en el simplex $\mathbb{S}^3$

2.1. Análisis de datos composicionales

Antes de indicar la problemática que ocasionan los análisis estadísticos de los datos composicionales sin tener en cuenta su especial naturaleza, es necesario introducir las siguientes definiciones: El operador clausura, la noción de subcomposición y la invarianza por permutación.

Para poder obtener un dato composicional $\mathbb{S}^D$ es necesario contar con un vector con componentes positivas y posteriormente dividir cada una de estas componentes por la suma de todas ellas.

- Operador clausura $\mathcal{C}$: Es una transformación que hace corresponder a cada vector $x = (x_1, x_2, \ldots, x_D)'$ de $\mathbb{R}_+^D$ su dato composicional asociado $\mathcal{C}(x) = \frac{kx}{(x_1+x_2+\cdots+x_D)}$ de $\mathbb{S}^D$, siendo k la constante de clausura.

  Tras aplicar esta transformación podemos comparar datos de muestras de diferentes tamaños. Por ejemplo: los vectores $\mathbf{A} = [12, 3, 4]$ y $\mathbf{B} = [2400, 600, 800]$ representan la misma composición, ya que la importancia relativa (los ratios) entre sus componentes es la misma. Por lo tanto, un análisis

de composición sensible debería proporcionar la misma respuesta independientemente del valor o incluso independientemente de si se aplicó la clausura, o bien si la suma de los vectores son valores diferentes. Decimos que el análisis será entonces invariante a la escala. Aitchison (1986) demostró que todas las funciones invariantes de escalamiento de una composición pueden expresarse como funciones de log-ratio $\ln\left(\frac{x_i}{x_j}\right)$.

- Noción de subcomposición: Si $\mathbb{S}$ es un subconjunto cualquiera de las partes $1, 2, \ldots, D$ de un dato composicional $x \in \mathbb{S}^D$, y $x_{\mathbb{S}}$ simboliza el subvector formado por las correspondientes partes de x, entonces $s = \mathcal{C}(x_{\mathbb{S}})$ recibe el nombre de subcomposición de las $\mathbb{S}$ partes de x. Si bien la formación de una subcomposición es una transformación de $\mathbb{S}^D$ a un símplex de dimensión inferior, esta tiende a conservar la magnitud relativa entre las partes.

- Invarianza por permutación: Las conclusiones que se obtienen de los análisis composicionales no están relacionadas con la ordenación de las partes. Es decir, si las partes de una composición se consideran ordenadas y/o clasificadas a través de categorías, al realizar el análisis composicional, esta información debida a la ordenación no debe tenerse en cuenta.

2.2. Marco geométrico

La característica fundamental de un vector de datos composicionales es, como se ha dicho, que la suma de sus componentes, siempre positivas, sea una constante. Así, un cambio en una componente conlleva un cambio en por lo menos una de las demás componentes.

Una de las dificultades más relevantes es la imposibilidad de interpretar correctamente las covarianzas y los coeficientes de correlación. En 1986 Pearson ya manifestó esta dificultad en "On a form of spurious correlation" , donde advertía a la comunidad científica de la existencia de una falsa correlación entre las partes de una composición que conducía a interpretaciones erróneas. En conclusión es que las correlaciones que involucran variables composicionales son espurias y no tiene sentido interpretarlas. Lo que implica replantearse el concepto de independencia. La cuestión clave es que la geometría del espacio muestral sobre el que se define un vector de proporciones es diferente de la clásica geometría Euclídea de R^D, por ello las técnicas multivariantes habitualmente utilizadas, y fundamentadas en esta geometría, no son directamente aplicables.

Es por esto que los conceptos propuestos por Aitchison (1986) han llevado a la llamada *Geometría de Aitchison del símplex* (Pawlowsky-Glahn & Egozcue 2001) a satisfacer los principios mencionados en la sección anterior.

Se consideran entonces las composiciones x, y en $\mathbb{S}^D$, cuyas componentes están denotadas por x_i, y_i, respectivamente. Entonces:

- La perturbación es la encargada de medir el cambio composicional en el símplex, siendo así la perturbación de x con y definida como la composición:

$$x \oplus y = \mathcal{C}(x_1 y_1, \ldots, x_D y_D)$$

- La potenciación de x con el número real α es la composición:

$$x \otimes y = \mathcal{C}(x_1^{\alpha}, \ldots, x_D^{\alpha})$$

Estas operaciones definidas en $\mathbb{S}^D$ cumplen los requisitos de las operaciones de un espacio vectorial. Pero el principal mérito de la perturbación es que, además de atender a los principios del análisis composicional, tiene interpretación en el campo sobre el cual se está trabajando.

Adicionalmente, es posible definir un producto interior en $\mathbb{S}^D$

$$\langle x, x^{'} \rangle = \left[\frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{x_i^{'}}{x_j^{'}} \right] \tag{1}$$

que confiere al símplex una estructura de espacio Euclídeo (Billheimer et al. 2001). Desde un punto de vista estadístico esta estructura no es irrelevante, ya que las medidas básicas de tendencia central, variabilidad y distancia deben ser coherentes con la naturaleza de los datos. El producto interior de la ecuación (1) induce una distancia en $\mathbb{S}^D$ definida como

$$d_a^2 \langle x, x^{'} \rangle = \sum_{i=1}^{D} \left(\ln \frac{x_i}{g(x)} - \ln \frac{x_i^{'}}{g(x^{'})} \right)^2 \tag{2}$$

donde $g(x) = (x_1, \ldots, x_D)^{1/D}$ es la media geométrica de la composición x. La distancia de la fórmula (2) se conoce como *distancia de Aitchison* y es totalmente compatible con las operaciones básicas en el símplex y con la particular naturaleza de los datos composicionales (Aitchison et al. 2000).

2.3. Transformaciones log-cociente

La invariancia por escala de las composiciones lleva de forma natural a utilizar cocientes entre las partes y, supuesta la escala relativa de los cocientes, a considerar sus logaritmos. Una vez mencionada la necesidad de centrar la atención en los cocientes entre las partes, surge la pregunta sobre qué tipo de transformaciones aplicar. Lo más importante de la metología propuesta por Aitchison (1986) es la transformación de una composición definida sobre $\mathbb{S}^D$ en un vector que involucre los cocientes entre las partes y que esté definido sobre el espacio real. Si esa transformación es biyectiva se establece una correspondencia uno a uno entre las composiciones del símplex y los correspondientes vectores transformados reales. Así, se tiene la posibilidad de resolver cualquier problema que afecte las composiciones usando las técnicas multivariantes habituales en espacios reales.

- Transformación logcociente aditiva (alr): Donde una composición pasa de $x \in \mathbb{S}^D$ a $y \in \mathbb{R}^{D-1}$ y se define como:

$$\begin{aligned} y = alr(x) &= \left[\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right] \\ &= \ln(x) \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & -1 \end{bmatrix} \end{aligned}$$

  y su inversa:

$$alr(y)^{-1} = \mathcal{C}\left[\exp([y; 0])\right]$$

  Esta transformación es biyectiva pero no es simétrica en las partes de x ya que la parte del denominador adquiere un protagonismo especial respecto al resto. Esta transformación tiene como fuerte la eliminación de las unidades, y está más vinculada a modelizaciones paramétricas ya que a partir de ella es posible definir una clase de distribuciones. Se observa que el cociente elimina las constantes de clausura o unidades que puedan multiplicar a las partes y el logaritmo se hace cargo de la escala relativa. Esta transformación adolece de la falta de simetría al elegir una de las partes, en este caso la última, como denominador común, violando el principio de invariancia por permutación de las partes.

- Transformación logcociente centrada (clr): Donde la composición pasa de $x \in \mathbb{S}^D$ a $z \in \mathbb{R}^D$ y se define como:

$$\begin{aligned} z = clr\,(x) &= \left[\ln \frac{x_1}{g\,(x)}, \ldots, \ln \frac{x_D}{g\,(x)}\right] \\ &= \frac{\ln\,(x)}{D} \cdot \begin{bmatrix} D-1 & -1 & \cdots & -1 \\ -1 & D-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & D-1 \end{bmatrix} \end{aligned}$$

  Donde $g\,(x) = \sqrt[D]{x_1 \cdots x_D}$ es la media geométrica de las D partes de x.

  Ahora, su inversa se define como:

$$clr\,(z)^{-1} = \mathcal{C}\left[\exp\,(z)\right]$$

  Esta trasformación es biyectiva, y simétrica entre las partes. Su imagen es el hiperplano de $\mathbb{R}^D$ que pasa por el origen y es ortogonal al vector de unidad, es decir, la suma de las componentes del vector transformado es igual a cero. Nos encontramos ante una nueva dificultad ya que la matriz de covarianzas del vector trasformado será singular. Esta transformación está vinculada en mayor parte a contextos no paramétricos, ya que a partir de ella es posible especificar la distancia de Aitchison en términos de la distancia Euclídea entre los vectores clr-transformados, esto es:

$$d_a\left(x, x^{'}\right) = d_e\left(clr\,(x), clr\left(x^{'}\right)\right)$$

  Por esta razón Aitchison aplica una estrategia. En las aplicaciones que exigen simetría en el tratamiento de sus componentes utiliza la transformación clr. Para la modelización de conjuntos de datos composicionales con distribuciones multivariantes, utiliza la transformación alr.

- Transformación logcociente isométrica (ilr): La caracterización del símplex como espacio euclídeo sugiere a Egozcue et al. (2003) proponer la transformación log-cociente isométrica, que salva los principales problemas de las dos anteriores. La transformación *ilr* de una composición se define como:

$$ilr\,(x) = [\langle x, e_1\rangle, \ldots, \langle x, e_{D-1}\rangle]$$

  donde $e_1, \ldots, e_{D-1}$ es una base ortonormal obtenida a partir del producto interior de la ecuación (1). Esto es, lo componentes del vector ilr transformado son las coordenadas de la composición x respecto a la base $e_1, \ldots, e_{D-1}$. Esta transformación, como su nombre lo indica, es isométrica, y los datos ilr transformados se representan en los habituales ejes ortogonales. En Egozcue et al. (2003) los autores obtienen una expresión explícita para la transformación dada una base ortonormal particular:

$$ilr\,(x) = y = [y_1, \ldots, y_{D-1}] \in \mathbb{R}^{D-1}, dondey_i = \frac{1}{\sqrt{i\,(i+1)}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i}\right)$$

  El principal problema aquí es determinar cuál es la base ortonormal más apropiada para un problema concreto, la que proporciona las expresiones que hacen más fácil la interpretación de los

resultados, ya que las coordenadas ilr no suelen ser fáciles de interpretar. Una posibilidad es utilizar las coordenadas y expresar los resultados en la base canónica de $\mathbb{R}^D$ sin abandonar el símplex. La función inversa de esta transformación se define así:

$$x = ilr^{-1}(y) = \bigoplus_{j=1}^{D-1} y \otimes e_j$$

2.4. Distribuciones de probabilidad

2.4.1. Distribución normal

Una composición aleatoria X tiene una distribución normal en el simplex, con vector de medias m y matriz de varianza Σ, que se denota como $N_{\mathbb{S}}(m, \Sigma)$ si al proyectarla en cualquier dirección arbitraria del simplex D con el producto escalar de Aitchison produce una variable aleatoria con una distribución normal univariante con vector de medias $\langle m, D\rangle$ y varianza clr $(D) \cdot \Sigma \cdot clr^t(D)$. Teniendo en cuenta lo anterior, si se tiene una base del simplex denotada por V la cual sea una matriz cuasi-ortonormal, entonces las coordenadas ilr (x) de la composición aleatoria tienen una distribución normal multivariante; es decir, su densidad conjunta con respecto a la medida de Aitchison $\lambda_{\mathbb{S}}$ es:

$$f(x; \mu_\nu, \Sigma_\nu) = \frac{1}{\sqrt{(2\pi)^{(D-1)} \cdot |\Sigma_\nu|}} exp\left[-\frac{1}{2}(ilr(x) - \mu_\nu)\Sigma_\nu^{-1} \cdot (ilr(x) - \mu_\nu)^t\right], \tag{3}$$

en donde μ_ν y Σ_ν son respectivamente el vector de medias y la matriz de varianza.

2.4.2. Distribución Aitchison

Una composición aleatoria X tiene una distribución Aitchison con D componentes, vector de localización θ y dispersión β_ν (matriz cuadrada de orden $(D-1)$) si su log-densidad se puede expresar como:

$$\log f(x; \theta, \beta_\nu) = -k(\theta, \beta_\nu) + (\theta - 1)\log(x)^t + ilr(x)\,\beta_\nu\, ilr(x)^t \tag{4}$$

en lo cual $k(\theta, \beta_\nu)$ es una constante que asegura que al integrar la densidad se obtendrá un resultado de 1, lo cual se tiene al cumplir las siguientes dos condiciones:

- β_ν es una forma cuadrática simétrica definida negativa y $\sum_{i=1}^{D} \theta_i \geq 0$
- β_ν es simétrica pero no definida positiva y si además $\theta_i > 0$ para todo $i = 1, ..., D$

Para los casos en los que se tenga una función que trabaje bajo la distribución de Aitchison y motivo por el cual los parámetros satisfagan unicamente la primera condición, fue propuesta una parametrización alternativa la cual brinda una mejor interpretatibilidad:

$$\begin{aligned} f(x; \alpha, m, \Sigma_\nu) = {} & exp\left[-k(\alpha, m, \Sigma_\nu)\right] \times exp\left[(\alpha - 1) \cdot \log(x) \cdot 1^t\right] \\ & \times exp\left[-\frac{1}{2} ilr(x \ominus m) \cdot \Sigma_\nu^{-1} \cdot ilr^t(x \ominus m)\right], \end{aligned} \tag{5}$$

donde:

- Σ_ν Es una matriz cuadrática $(D-1)$ definida positiva
- m una composición de $D - partes$

2.4.3. Distribución Dirichlet

Esta distribución se obtiene aplicando el operador de clausura descrito anteriormente a un vector D independiente, igualmente escalado (es decir, con el mismo parámetro λ) por variables que se distribuyen Gamma.

Por lo anterior si $x_i \sim \Gamma(\lambda, \alpha_i)$ luego $z = \mathcal{C}[x] \sim \mathcal{D}_i(\alpha)$, donde $x = [x_i]$ y $\alpha = [\alpha_i]$ son ambos vectores con D componentes positivas. Luego la notación de la función de densidad de esta distribución es la siguiente:

$$f(z;\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_D)} \frac{z_1^{\alpha_1}\ldots z_D^{\alpha_D}}{z_1\ldots z_D}, \tag{6}$$

donde $\alpha_1 = \alpha_1 + \ldots + \alpha_D$ y $\Gamma(\cdot)$. Adicionalmente se tiene que la media aritmética y la varianza de esta distribución se denotan de la siguiente manera:

1. Media: $E_{[z]} = \mathcal{C}[\alpha]$

2. Varianza: $Var_{\mathbb{R}}(z) = \frac{1}{\alpha_0 + 1}\left(diag(\overline{z}) - \overline{z}^t \cdot \overline{z}\right)$

Para mayores detalles acerca de las tres distribuciones mencionadas remitirse al Van den Boogaart & Tolosana-Delgado (2013)

2.4.4. Distribución Kent

Los datos composicionales se pueden transformar en datos direccionales mediante la transformación de raíz cuadrada ($y = \sqrt{\mathbf{u}} = (\sqrt{u_1}, \sqrt{u_2}), \cdots, \sqrt{u_p})$) y luego modelarse utilizando distribuciones definidas en la hiperesfera tal como la distribución de Kent. El enfoque actual para estimar los parámetros en el modelo de Kent para estos datos se basa en una suposición de gran concentración que supone que la mayoría de los datos transformados no se distribuye demasiado cerca de los límites del ortante positivo. La distribución de Kent para datos direccionales es un buen modelo candidato porque tiene una estructura de covarianza lo suficientemente general.

La distribución de 5 parámetros de Fisher-Bingham o la distribución de Kent, que lleva el nombre de Ronald Fisher, Christopher Bingham y John T. Kent, es una distribución de probabilidad en la esfera unitaria bidimensional S^2 en $\mathbb{R}^3$. Es el análogo en la esfera unitaria bidimensional de la distribución normal bivariada con una matriz de covarianza sin restricciones. La distribución pertenece al campo de las estadísticas direccionales.

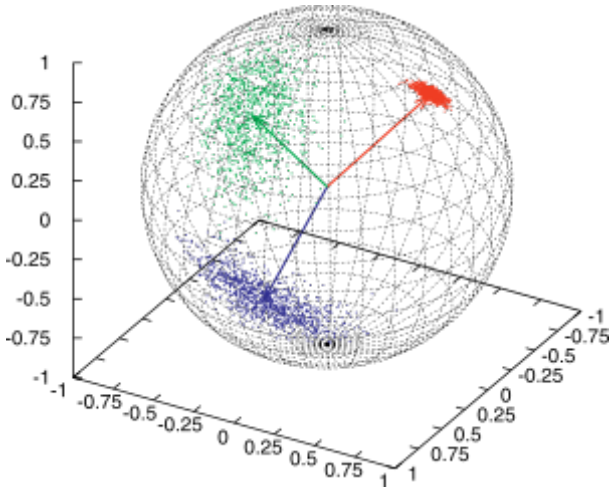

Figura 2: Representación gráfica de los datos direccionales.

La función de densidad de probabilidad $f(\mathbf{x})$, de la distribución de Kent está dada por:

$$f(\mathbf{x}) = \frac{1}{\mathrm{c}(\kappa,\beta)} \exp\{\kappa\boldsymbol{\gamma}_1 \cdot \mathbf{x} + \beta[(\boldsymbol{\gamma}_2 \cdot \mathbf{x})^2 - (\boldsymbol{\gamma}_3 \cdot \mathbf{x})^2]\}$$

Donde $\mathbf{x}$ es un vector unitario tridimensional y la constante normalizadora $\mathrm{c}(\kappa,\beta)$ es:

$$c(\kappa,\beta) = 2\pi \sum_{j=0}^{\infty} \frac{\Gamma(j+\frac{1}{2})}{\Gamma(j+1)} \beta^{2j} \left(\frac{1}{2}\kappa\right)^{-2j-\frac{1}{2}} I_{2j+\frac{1}{2}}(\kappa)$$

Donde $I_v(\kappa)$ es la función de Bessel modificada.

El parámetro κ (con $\kappa > 0$) determina la concentración o extensión de la distribución, mientras que β (con $0 \leq 2\beta < \kappa$) determina la elipticidad de los contornos de igual probabilidad. Cuanto más altos sean los parámetros κ y β más concentrada y elíptica será la distribución, respectivamente. γ_1, es la dirección media y los vectores γ_2, γ_3 son los ejes mayor y menor. Los dos últimos vectores determinan la orientación de los contornos de probabilidad iguales en la esfera, mientras que el primer vector determina el centro común de los contornos. La matriz 3x3 $(\gamma_1, \gamma_2, \gamma_3)$ debe ser ortogonal.

2.5. Modelos para datos composicionales

Actualmente existen diferentes formas para modelizar los datos composicionales, en la literatura se pueden encontrar diferentes transformaciones para los datos composicionales bajo las cuales la variable estudiada se ha de analizar como una variable ordinaria. Otra situación comúnmente presentada es el asumir una distribución adecuada para los datos de composición y aplicar ciertas regresiones sin alterar la variable.

2.5.1. Modelo Dirichlet

Bajo un modelo de efectos fijos suponemos que el tamaño del efecto real (entiéndase como un cambio en el resultado luego de una intervención) para todos los estudios es idéntico, y la única razón por la que el tamaño del efecto varía entre los estudios es error de muestreo (error en la estimación del tamaño del

efecto), bajo este modelo se tratan a las variables explicativas como si la información que contienen fuese no-aleatoria.

Dado lo anterior, se plantea un modelo cuya variable respuesta sigue una distribución Dirichlet, donde se trata de predecir los α para cada componente mediante un conjunto de variables. Debido a que α debe ser mayor que 0, podemos utilizar convenientemente un enlace log para esta parametrización. Así, para cada componente y_i hay un vector de coeficientes de regresión β junto con una matriz de diseño apropiada X.

$$log\left(\begin{bmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\\ \vdots\\ \alpha_C\end{bmatrix}\right) = \begin{bmatrix}X_1\beta_1\\ X_2\beta_2\\ X_3\beta_3\\ \vdots\\ X_C\beta_C\end{bmatrix}$$

Donde C es el número de categorías o proporciones de y

Debido a que todos los parámetros α son modelados individualmente, la heteroscedasticidad se contabiliza implícitamente.

La log-verosimilitud del modelo comúnmente parametrizado (l_c) se define en la ecuación 1. Donde C es el número de proporciones o categorías de la variable respuesta.

$$l_c\left(y/\alpha\right) = log\Gamma\left(\sum_{c=1}^{C}\alpha_c\right) - \sum_{c=1}^{C} log\Gamma\left(\alpha_c\right) + \sum_{c=1}^{C}\left(\alpha_c - 1\right) log\left(y_c\right) \tag{7}$$

Bajo la ecuación anterior, se espera que los datos causen ciertos problemas dado el rango en que oscilan, estos pueden venir del intervalo [0, 1] en lugar de (0, 1), ya que $log(y_c = 0) = -\infty$. Para remediar esto, es decir, si los valores de 0 y 1 están presentes en los datos, se obtiene una generalización de la transformación propuesta en Smithson & Verkuilen (2006), donde N es el número de observaciones.

$$y^* = \frac{y\left(N-1\right) + 1/C}{N} \tag{8}$$

Este ajuste tiene como función el comprimir los datos simétricamente alrededor de 0.5, desde un rango de 1 hasta $\left(N-1\right)/N$, con lo cual los valores extremos se afectan más que los cercanos a 0.5. Adicional, cuando $N \to \infty$ la compresión desaparece; los conjuntos de datos más grandes son menos afectados por la transformación. Ahora, si tenemos un conjunto de variables $y_c \sim D\left(\alpha_c\right)$ para $c \in \{1, \ldots, C \geq 2\}$, el modelo se puede reescribir de la siguiente manera:

$$g\left(\alpha_c\right) = \eta_c = X^{[c]}\beta^{[c]} \tag{9}$$

Donde $g\left(\cdot\right)$ es la función de enlace que será el $log\left(\cdot\right)$ para este modelo ya que $\alpha_c > 0$, y η se conoce como el predictor lineal. Como mencionamos anteriormente, hay predictores separados en cada componente c, y la matriz predictora se representa como $X^{[c]}$ donde el superíndice representa el componente estimado. El número de variables independientes conduce naturalmente a diferentes números de coeficientes de regresión en cada dimensión, lo que se explica por la escritura $\beta^{[c]}$. La matriz de predictores para un componente c, $X^{[c]}$ tiene $i = N$ filas y $j = J^{[c]}$ columnas mientras que los coeficientes de regresión $\beta^{[c]}$ son un vector columna con $j = J^{[c]}$ elementos.

Con el fin de acortar los términos, se reescribe la función de verosimilitud definiendo primero $A^{[c]} = X^{[c]}\beta^{[c]}$:

$$l_c\left(y_c|X^{[c]},\beta^{[c]}\right) = log\Gamma\left(\sum_{c=1}^{C} A^{[c]}\right) - \sum_{c=1}^{C} log\Gamma\left(A^{[c]}\right) + \sum_{c=1}^{C}\left(A^{[c]}-1\right) log\left(Y^{[c]}\right) \tag{10}$$

2.5.1.1 Ajuste del modelo

Para optimizar estos modelos multivariados que generalmente contienen un gran número de variables dependientes, se ajustan en tres etapas:

1. Valores iniciales y refinamiento de los mismos: La función interna ***get_starting_values()*** (Maier 2015, Maier 2014) calcula los valores iniciales para la parametrización común, en donde se asume que cada variable dependiente se distribuye beta (esto es admisible ya que las variables individuales están marginalmente distribuidas beta) y los parámetros α y β de la distribución son estimados utilizando el algoritmo ***maxBFGS()*** de la librería ***MaxLik*** (Henningsen & Toomet 2011)

2. Optimización Broyden-Fletcher-Goldfarb-Shanno (BFGS): Con los respectivos valores iniciales se lleva a cabo el algoritmo BFGS (Broyden 1970, Fletcher 1970, Goldfarb 1970, Shanno 1970) utilizando el gradiente analítico más que el numérico, con un criterio de convergencia de 10^{-5}. Este método ha demostrado ser rápido, robusto y fiable en la práctica (las estimaciones no cambian en el siguiente paso), sin embargo, los errores estándar que se extraen de la Hessiana no son confiables, por lo que este no es el paso final de la optimización. El algoritmo se realiza de la siguiente forma:

   A partir de una aproximación inicial $\mathbf{x}_0$ y una matriz aproximada Hessiana B_0, se repiten los siguientes pasos mientras $\mathbf{x}_k$ converge a la solución:

   a) Obtener una dirección $\mathbf{p}_k$ resolviendo $B_k\mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.

   b) Realice una línea de búsqueda para encontrar un tamaño de paso aceptable α_k en la dirección encontrada en el primer paso.

   c) Defina $\mathbf{s}_k = \alpha_k\mathbf{p}_k$ y actualice $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{s}_k$

   d) $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$.

   e) $B_{k+1} = B_k + \frac{\mathbf{y}_k\mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}}\mathbf{s}_k} - \frac{B_k\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}}B_k}{\mathbf{s}_k^{\mathrm{T}}B_k\mathbf{s}_k}$.

   $f(\mathbf{x})$ denota la función objetivo a ser minimizada. La convergencia puede comprobarse observando la norma del gradiente $|\nabla f(\mathbf{x}_k)|$. B_0 puede inicializarse como $B_0 = I$ tal que el primer paso será equivalente a un descenso de gradiente, pero los siguientes pasos son más y más refinados por B_k, la aproximación a la Hessiana.

   El primer paso del algoritmo se lleva a cabo utilizando la inversa de la matriz B_k, que se puede obtener de manera eficiente aplicando la fórmula de Sherman-Morrison al paso 5 del algoritmo, dando:

$$B_{k+1}^{-1} = \left(I - \frac{s_k y_k^T}{y_k^T s_k}\right) B_k^{-1} \left(I - \frac{y_k s_k^T}{y_k^T s_k}\right) + \frac{s_k s_k^T}{y_k^T s_k}.$$

   Esto se puede calcular eficientemente sin matrices temporales, asumiendo que B_k^{-1} es simétrica, y que $\mathbf{y}_k^{\mathrm{T}}B_k^{-1}\mathbf{y}_k$ y $\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k$ son escalares, usando una expansión tal como:

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k + \mathbf{y}_k^{\mathrm{T}}B_k^{-1}\mathbf{y}_k)(\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}})}{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k)^2} - \frac{B_k^{-1}\mathbf{y}_k\mathbf{s}_k^{\mathrm{T}} + \mathbf{s}_k\mathbf{y}_k^{\mathrm{T}}B_k^{-1}}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k}$$

.

3. Optimización de Newton-Raphson: Una vez se obtienen las estimaciones del paso dos, la estimación final se consigue a través del algoritmo NR, el cual se implementa en la función ***maxNR ()*** dentro de la librería ***maxLik*** (Henningsen & Toomet 2011), donde las derivadas analíticas de primer y segundo orden se suministran para acelerar la convergencia (el criterio es 10^{-10}) y para asegurar la estabilidad numérica. Los errores estándar se obtienen de la Hessiana $\left(H\left(\hat{\theta}\right)\right)$ así:

$$SE\left(\hat{\theta}\right) = \sqrt{diag\left(\Sigma\left(\hat{\theta}\right)\right)} = \sqrt{diag\left(-H\left(\hat{\theta}\right)^{-1}\right)} \tag{11}$$

2.5.1.2 Valores ajustados, predicción y residuales

Los valores ajustados para y_c se pueden obtener fácilmente dados los parámetros de los modelos $\hat{\beta}$ y los predictores X. Del mismo modo, las predicciones se pueden hacer para nuevos valores de x.

De los tipos residuales propuestos en Gueorguieva et al. (2008) se implementan los siguientes:

- Raw: Estos residuos son de poca importancia dada la asimetría y la heteroscedasticidad

$$\begin{aligned} r_c^{raw} &= y_c - E\left[y_c|\hat{\alpha}\right] \\ &= y_c - \frac{\hat{\alpha_c}}{\hat{\alpha_0}} \end{aligned} \tag{12}$$

- Pearson: Son una alternativa más significativa, a estos se les llaman estandarizados, y se definen de la siguiente forma:

$$\begin{aligned} r_c^{Pearson} &= \frac{y_c - E\left[y_c|\hat{\alpha}\right]}{\sqrt{VAR\left[y_c|\hat{\alpha}\right]}} \\ &= \frac{y_c - \frac{\hat{\alpha_c}}{\hat{\alpha_0}}}{\sqrt{\left[\hat{\alpha_c}\left(\hat{\alpha_0} - \hat{\alpha_c}\right)\right] / \left[\hat{\alpha_0}^2\left(\hat{\alpha_0} + 1\right)\right]}} \end{aligned} \tag{13}$$

- Residuos Composicionales: Son la suma de los residuos de Pearson al cuadrado.

$$r_c^{composite} = \sum_{c=1}^{C}\left(r_c^{Pearson}\right)^2 \tag{14}$$

2.5.1.3 Selección del modelo y pruebas

Con el fin de seleccionar el modelo que mejor se ajusta a los datos, se aplica una prueba de razón de verosimilitud. En las siguientes ecuaciones, se probará el modelo **b** que está anidado en **a**. La estadística de prueba resultante es asintóticamente distribuida con grados de libertad igual a la diferencia en los parámetros de los modelos $(n_a - n_b)$. En R, la forma de aplicar este test, es bajo la función ***anova ()*** (R Core Team 2016), la cual, a través del $valor-p$ nos indicará si el ajuste de un modelo es significativamente mejor que el de otro.

$$\begin{aligned} LRT &= -2log\left(\frac{L_b}{L_a}\right) \overset{as.}{\sim} \chi^2_{n_a - n_b} \\ LRT &= -2\left(l_b - l_a\right) \overset{as.}{\sim} \chi^2_{n_a - n_b} \end{aligned} \tag{15}$$

2.5.2. Modelo mixto

Los modelos mixtos son aquellos que contienen tanto efectos fijos como efectos aleatorios, están construidos bajo el razonamiento de muchos otros modelos estadísticos, en el cual se trata de describir la relación entre lo que se conoce como variable respuesta y una o varias variables explicativas, también conocidas como independientes o covariables. Estos modelos en resumen, amplían el procedimiento de los modelos generales, ofreciendo la oportunidad de modelar no solo las medias sino también las varianzas y covarianzas de los datos, y se representan de la siguiente forma:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

donde:

- $\boldsymbol{y}$ es un vector conocido, con media $E(\boldsymbol{y}) = X\boldsymbol{\beta}$;
- $\boldsymbol{\beta}$ es el vector asociado a los parámetros fijos;
- $\boldsymbol{\gamma}$ es el vector asociado a los parámetros aleatorios, con media $E(\boldsymbol{\gamma}) = \mathbf{0}$ y matriz de covarianza $\text{var}(\boldsymbol{\gamma}) = \Sigma$;
- $\boldsymbol{\epsilon}$ es un vector de errores aleatorios, con media $E(\boldsymbol{\epsilon}) = \mathbf{0}$ y varianza $\text{var}(\boldsymbol{\epsilon}) = \mathbf{R}$;
- X Es una matriz diseño conocida que representa el valor de las covariables explicativas asociadas a los parámetros fijos;
- Z Es una matriz diseño conocida que representa el valor de las covariables explicativas asociadas a los parámetros aleatorios;

Además, se supone que siendo i e i' dos unidades distintas, las componentes de error están no correlacionadas y no asociadas con los efectos aleatorios, $cov(e_i, e_i') = 0$, $cov(\gamma_i, \gamma_i') = 0$ y $cov(e_i, \gamma_i) = 0$.

La matriz de covarianzas Σ contiene tanto la covarianza entre los efectos aleatorios dentro de una respuesta dada como la covarianzas entre los efectos aleatorios de las distintas variables respuestas. La matriz $\mathbf{R}$ tiene una estructura específica para reflejar la correlación entre las mediciones repetidas. Si no se hacen supuestos sobre la forma de la estructura de covarianzas para cada fila y columna de ϵ_i, los parámetros de covarianzas de $\mathbf{R}$ serán numerosos para ser estimados correctamente. Además, si se sospecha la existencia de una estructura, utilizarla conduce a una estimación e inferencia más eficiente.

2.5.2.1 Ajuste del modelo

El paquete lme4 (Bates et al. 2015) para R proporciona funciones para ajustar y analizar modelos lineales mixtos, modelos mixtos lineales generalizados y modelos mixtos no lineales. En cada uno de estos nombres, el término "mixto" o, más plenamente, "efectos mixtos", denota un modelo que incorpora términos de efectos fijos y de efectos aleatorios en una expresión predictora lineal a partir de la cual se puede evaluar la media condicional de la respuesta.

Para la estimación de los efectos β y b bajo el supuesto de normalidad e independencia del error y del efecto aleatorio, b y ϵ, tienen una distribución conjunta normal multivariada:

$$f\left(b, \epsilon\right) = \frac{1}{(2\pi)^{(T+nk)/2}} \begin{bmatrix} \Sigma & 0 \\ 0 & \mathbf{R} \end{bmatrix}^{-n/2} \exp\left[-\frac{1}{2} \begin{bmatrix} \gamma \\ Y - X\beta - Z\gamma \end{bmatrix}' \begin{bmatrix} \Sigma & 0 \\ 0 & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \gamma \\ Y - X\beta - Z\gamma \end{bmatrix} \right] \tag{16}$$

Donde $T = \sum_{i=1}^{N} n_i$. El proceso de estimación de máxima verosimilitud se basa en la minimización del exponente de esta expresión. El proceso usual para llevarlo a cabo es el siguiente. Sea:

$$Q = \begin{bmatrix} \gamma \\ Y - X\beta - Z\gamma \end{bmatrix}^{'} \begin{bmatrix} \Sigma & 0 \\ 0 & \mathbf{R} \end{bmatrix}^{-1} \begin{bmatrix} \gamma \\ Y - X\beta - Z\gamma \end{bmatrix} \tag{17}$$

La matriz de varianzas en el término de la mitad, tiene submatrices en la diagonal invertibles, por lo tanto, la inversa existe, y Q se convierte en:

$$\begin{aligned} Q &= \left(\gamma\Sigma^{-1}\right)^{'} \gamma + \left(Y - X\beta - Z\gamma\right)^{'} R^{-1} \left(Y - X\beta - Z\gamma\right) \\ &= \left(\gamma\Sigma^{-1}\right)^{'} \gamma + \left(\left(Y - X\beta\right)^{'} - \left(Z\gamma\right)^{'}\right) R^{-1} \left(\left(Y - X\beta\right) - Z\gamma\right) \end{aligned}$$

Ahora se deriva Q con respecto a γ y β:

$$\frac{\delta Q}{\delta \gamma}, \frac{\delta Q}{\delta \beta}.$$

Entonces:

$$\frac{\delta Q}{\delta \beta} = \left(-2X\right)^{'} R^{-1} \left(Y - X\beta\right) + 2X^{'} R^{-1} Z\gamma$$

luego, igualando a cero,

$$X^{'} R^{-1} X\hat{\beta} + X^{'} R^{-1} Z\hat{\gamma} = X^{'} R^{-1} Y$$

Y de:

$$\frac{\delta Q}{\delta \gamma} = 2\Sigma^{-1}\gamma + 2Z^{'} R^{-1} Z\gamma - 2Z^{'} R^{-1} \left(Y - X\beta\right)$$

igualando a cero:

$$Z^{'} R^{-1} X\hat{\beta} + \left(Z^{'} R^{-1} Z + \Sigma^{-1}\right) \hat{\gamma} = Z^{'} R^{-1} Y$$

En forma matricial:

$$\begin{bmatrix} X^{'} R^{-1} X & X^{'} R^{-1} Z \\ Z^{'} R^{-1} X & Z^{'} R^{-1} Z + \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\gamma} \end{bmatrix} = \begin{bmatrix} X^{'} R^{-1} Y \\ Z^{'} R^{-1} Y \end{bmatrix}$$

Luego, los valores de $\hat{\beta}$ y de $\hat{\gamma}$:

$$\hat{\gamma} = BZ^{'} V^{-1} \left(Y - X\hat{\beta}\right)$$

$$\hat{\beta} = \left[X^{'} V^{-1} X\right]^{-1} X^{'} V^{-1} Y$$

Las estimaciones de máxima verosimilitud o la máxima verosimilitud restringida (REML) de los parámetros en modelos lineales de efectos mixtos se pueden determinar usando la función lmer en el paquete lme4 para R (Bates et al. 2015). Esta función se compone de cuatro módulos ampliamente independientes.

En el primer módulo, se analiza una fórmula de modelo mixto y se convierte en las entradas necesarias para especificar un modelo mixto lineal. El segundo módulo utiliza estas entradas para construir una función R que toma los parámetros de covarianza, θ, como argumentos y devuelve negativa dos veces la probabilidad de perfil de registro o el criterio de REML. El tercer módulo optimiza esta función objetivo para producir estimaciones de máxima verosimilitud (ML) o REML de θ. Finalmente, el cuarto módulo proporciona utilidades para interpretar el modelo optimizado.

3. Objetivos

3.1. General

- Ajustar un modelo mixto para datos composicionales a los datos del plebiscito por la paz de 2016 en Colombia utilizando el software *R*.

3.2. Específicos

- Ajustar un modelo lineal para datos composicionales a los datos del plebiscito por la paz de 2016 en Colombia en el software *R*.
- Comparar el ajuste de los modelos sobre el conjunto de datos analizado.
- Construir las rutinas computacionales necesarias para el ajuste de los modelos planteados en el software *R*.

4. Metodología

En vista a lo anterior y para cumplir con los objetivos antes propuestos, esta tesis ha sido desarrollada de la siguiente manera:

1. Levantamiento de la información necesaria: Se buscaron bases de datos de la Registraduría, el Departamento Nacional de Planeación (DNP), la Unidad para la Atención y Reparación Integral a las Víctimas (UARIV) y cualquier otra base de datos de libre acceso que disponía el gobierno.

2. Revisión Conceptual: Se indagaron diferentes fuentes que contenían información sobre metodologías que aseguraban el correcto tratamiento de los datos composicionales, se optó por buscar desarrollos actuales del tema. Como producto de ello se elaboró una compilación de las metodologías usadas en la actualidad que se espera sirva de base y permita establecer cuáles deben ser las herramientas a tener en cuenta a la hora de analizar datos composicionales.

3. Una vez realizada la revisión de la documentación existente en cuanto a los desarrollos teóricos, se avanzó en la implementación de las herramientas computacionales que hasta la actualidad han sido desarrolladas para facilitar el análisis de datos composicionales estableciendo así su aplicabilidad en un campo de interés particular como lo es el análisis de los resultados electorales en Colombia. Específicamente se trabajó con los paquetes "DirichletReg" (Maier 2015, Maier 2014), "robCompositions" (Templ et al. 2011) y "lme4" (Bates et al. 2015) los cuales proporcionan funciones para el análisis idóneo de los mismos.

4. Finalmente, a partir de los resultados, aspectos y logros alcanzados en este ejercicio, se trabajó en la consolidación de todo lo aprendido, teórica y prácticamente, para lo cual se procedió a elaborar un documento que contiene todos los conceptos y resultados computacionales producto del desarrollo del trabajo, en este caso, con aplicaciones en el campo de las estadísticas electorales más recientes que se han llevado a cabo en Colombia.

5. Resultados

El ejercicio de analizar los resultados del plebiscito por la paz en Colombia a través de los datos composicionales y modelos estadísticos tiene como soporte la revisión de información que trata sobre las tendencias en el país, dando así una idea generalizada de las condiciones sociales que caracterizan al mismo. Lo anterior, junto con resultados de procesos electorales similares, logra explicar de manera eficiente los resultados del plebiscito por la paz.

Las variables que se trabajaron a lo largo del ejercicio son las siguientes:

- Resultados del plebiscito por la paz (Sí,No,Resto votos)
- Resultados de la segunda vuelta de las elecciones presidenciales 2014 (Juan Manuel Santos, Oscar Iván Zuluaga, Resto votos)
- Índice de necesidades básicas insatisfechas (NBI o INBI): Método directo para identificar carencias críticas en una población y caracterizar la pobreza. Usualmente utiliza indicadores directamente relacionados con cuatro áreas de necesidades básicas de las personas (vivienda, servicios sanitarios, educación básica e ingreso mínimo).
- Índice de riesgo de victimización (IRV): Herramienta para el análisis de los diferentes escenarios de victimización en el marco del conflicto armado. Con este índice se busca monitorear las causas y los efectos de la victimización.

5.1. Análisis Descriptivo

Para el desarrollo computacional del ejercicio planteado, se tienen en cuenta las siguientes librerías:

- robCompositions (Templ et al. 2011)
- DirichletReg (Maier 2015, Maier 2014)
- lme4 (Bates et al. 2015)

5.1.1. Transformación ILR de cada variable composicional

Para las dos variables composicionales tratadas durante el ejercicio se les aplicó la transformación ILR. Estas transformaciones están representadas en las últimas dos columnas de la tabla 1 y 2.

5.1.1.1 Resultados plebiscito por la paz:

MUNICIPIO	plebiscito_Si	plebiscito_No	plebiscito_Abstencion	Ple_Si_ILR	Ple_No_ILR
MEDELLIN	0.1694	0.2878	0.5428	-0.6920	-0.4487
ABEJORRAL	0.0865	0.1402	0.7733	-1.0919	-1.2072
ABRIAQUI	0.1117	0.1739	0.7145	-0.9386	-0.9993
ALEJANDRIA	0.2146	0.1547	0.6307	-0.3066	-0.9938
AMAGA	0.0955	0.1893	0.7153	-1.1017	-0.9400
AMALFI	0.0786	0.1379	0.7835	-1.1688	-1.2282

Tabla 1: Resultados del plebiscito por la paz y su respectiva transformación ILR.

5.1.1.2 Resultados de la segunda vuelta de las elecciones presidenciales 2014:

MUNICIPIO	JMS_2	OIZ_2	Otros_Votos	JMS_ILR	OIZ_ILR
MEDELLIN	0.2821	0.6063	0.1116	0.0665	1.1969
ABEJORRAL	0.1822	0.7641	0.0538	-0.0871	1.8767
ABRIAQUI	0.3520	0.5872	0.0608	0.5078	1.6034
ALEJANDRIA	0.5076	0.4346	0.0578	0.9502	1.4264
AMAGA	0.3817	0.5135	0.1047	0.4070	1.1244
AMALFI	0.1799	0.7161	0.1039	-0.3400	1.3647

Tabla 2: Resultados de la segunda vuelta de las elecciones presidenciales 2014 y su respectiva transformación ILR.

Este primer resultado muestra el escalamiento de cada una de las variables de tipo composicional, en las cuales se trabaja con el porcentaje de votos para cada una de las composiciones con el fin de eliminar el efecto del tamaño del municipio.

5.1.2. Gráficos Descriptivos

El utilizar gráficos en los análisis descriptivos es una actividad bastante provechosa ya que en la mayoría de los casos se logra obtener una visión clara acerca del comportamiento de la información con la que se está trabajando. Para el caso de estudio, la información está tomada a nivel municipal, esto facilita la construcción de mapas para el análisis descriptivo de la información como los que se muestran en las gráficas 3 y 4.

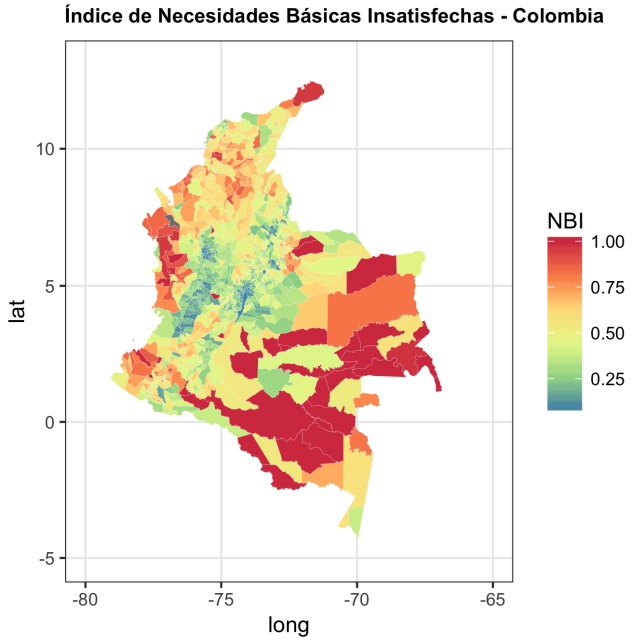

Figura 3: Distribución NBI 2011

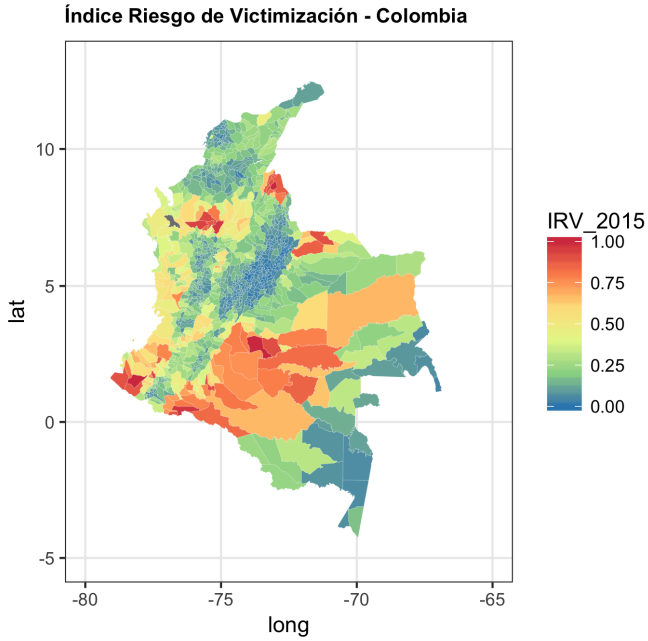

Figura 4: Distribución IRV 2015

- El índice de necesidades básicas insatisfechas (NBI) presenta que las zonas con mayor puntaje, es decir, las más "pobres", son las zonas de la periferia del país, ya que son zonas que se caracterizan por viviendas inadecuadas; con hacinamiento crítico, servicios inadecuados, alta dependencia

económica y niños en edad escolar que no asisten a la escuela. Hacia el centro del país vemos que el valor del índice se va reduciendo indicando mejoría en las condiciones económicas.

- El IRV nos muestra que las áreas más propensas a sufrir hechos victimizantes son los puntos de encuentro de las regiones Andes, Orinoquía y Amazonía, esta última en menor proporción, zonas que son consideradas como "rojas" en cuanto al conflicto armado. Dentro de estas regiones "rojas", podemos encontrar departamentos como son Antioquia, Cauca, Caquetá, Nariño, Valle del Cauca, Norte de Santander, Arauca, Putumayo y Meta.

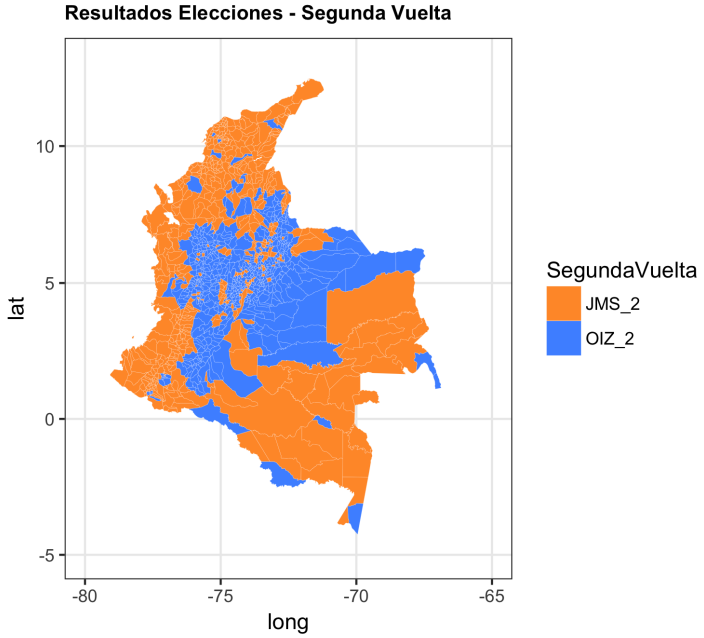

Figura 5: Candidatos segunda vuelta

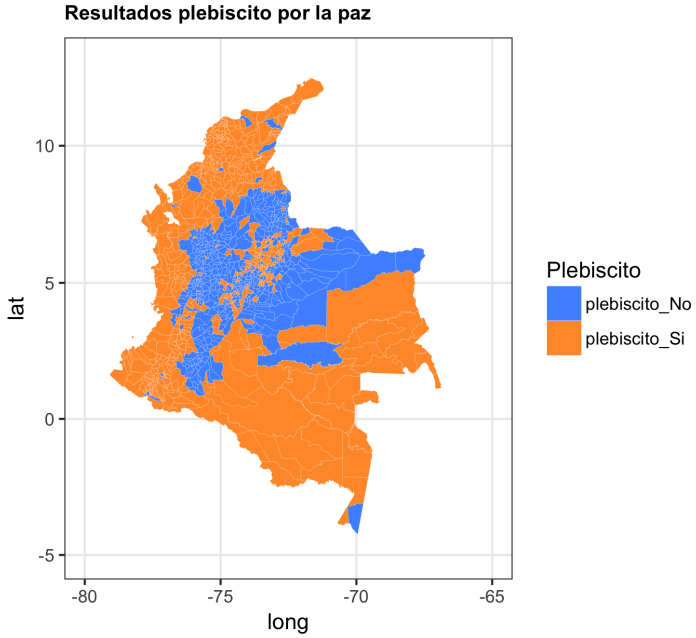

Figura 6: Resultados plebiscito por la paz

Para este par de gráficos se observa un patrón bastante importante; de las elecciones presidenciales para la segunda vuelta, los municipios en los que Juan Manuel Santos obtuvo el mayor número de votos, resultaron siendo los mismos municipios (al menos en un gran porcentaje) en los que ganó el "Sí" a la paz. Patrón que se invierte para los municipios en los que Óscar Iván Zuluaga ganó.

5.2. Modelos para los datos

5.2.1. Regresión Dirichlet

Teniendo en cuenta la información con la que se trabajó, el modelo cuya variable respuesta (Plebiscito por la paz) sigue una distribución dirichlet, se plantea de la siguiente forma:

$$log\left(\begin{bmatrix}\alpha_{Si}\\ \alpha_{No}\\ \alpha_{Abs}\end{bmatrix}\right) = \begin{bmatrix}\beta_{0,Si}+NBI_{Si}\beta_{nbi,si}+IRV_{Si}\beta_{irv,si}+JMS_{Si}\beta_{jms,si}+OIZ_{Si}\beta_{oiz,si}+OtrosVotos_{Si}\beta_{ov,si}\\ \beta_{0,No}+NBI_{No}\beta_{nbi,no}+IRV_{No}\beta_{irv,no}+JMS_{No}\beta_{jms,no}+OIZ_{No}\beta_{oiz,no}+OtrosVotos_{No}\beta_{ov,no}\\ \beta_{0,Abs}+NBI_{Abs}\beta_{nbi,abs}+IRV_{Abs}\beta_{irv,abs}+JMS_{Abs}\beta_{jms,abs}+OIZ_{Abs}\beta_{oiz,abs}+OtrosVotos_{Abs}\beta_{ov,abs}\end{bmatrix} \tag{18}$$

A continuación se muestran varios gráficos que resumen el ajuste del modelo mencionado.

```
---------------------------------------------------------------
Beta-Coefficients for variable no. 1: plebiscito_Si
                 Estimate   Std. Error z value             Pr(>|z|)
(Intercept)    2.654090579  0.081403615  32.604 < 0.0000000000000002 ***
NBI           -1.562041083  0.152998722 -10.210 < 0.0000000000000002 ***
IRV_2015      -0.044902108  0.169300650  -0.265              0.79084
JMS_2          0.000006425  0.000002478   2.593              0.00951 **
OIZ_2         -0.000024529  0.000011687  -2.099              0.03582 *
Otros_Votos_2  0.000064821  0.000052553   1.233              0.21742
---------------------------------------------------------------
Beta-Coefficients for variable no. 2: plebiscito_No
                 Estimate   Std. Error z value             Pr(>|z|)
(Intercept)    3.344383547  0.076386067  43.783 < 0.0000000000000002 ***
NBI           -3.181340675  0.148839996 -21.374 < 0.0000000000000002 ***
IRV_2015      -0.464276285  0.154621809  -3.003              0.00268 **
JMS_2         -0.000004975  0.000003301  -1.507              0.13180
OIZ_2          0.000014324  0.000008369   1.712              0.08698 .
Otros_Votos_2 -0.000079725  0.000042661  -1.869              0.06165 .
---------------------------------------------------------------
Beta-Coefficients for variable no. 3: plebiscito_Abstencion
                  Estimate    Std. Error z value             Pr(>|z|)
(Intercept)    3.9373008582  0.0772386183  50.976 <0.0000000000000002 ***
NBI           -1.5754428200  0.1525539184 -10.327 <0.0000000000000002 ***
IRV_2015      -0.0452362237  0.1663242208  -0.272              0.7856
JMS_2          0.0000054130  0.0000025064   2.160              0.0308 *
OIZ_2         -0.0000005518  0.0000092540  -0.060              0.9525
Otros_Votos_2 -0.0000553764  0.0000452537  -1.224              0.2211
---------------------------------------------------------------
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: 3241 on 18 df (600 BFGS + 3 NR Iterations)
AIC: -6446, BIC: -6356
Number of Observations: 1122
Link: Log
Parametrization: common
```

Figura 7: Estimación parámetros para la regresión dirichlet

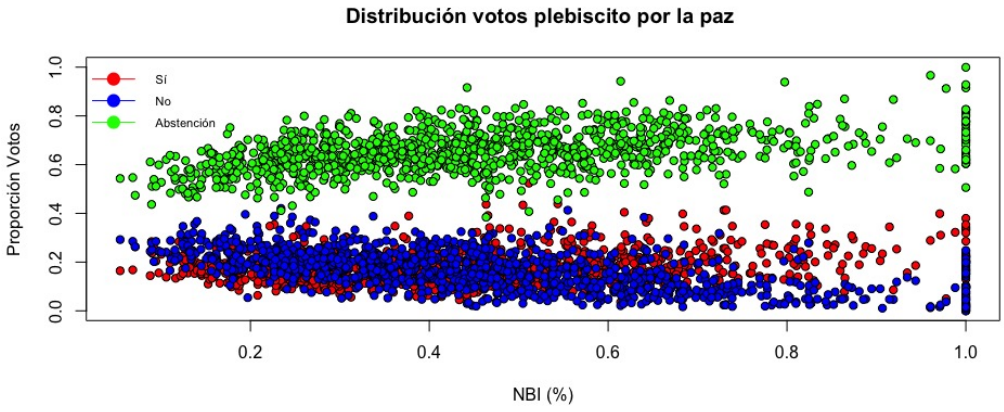

Figura 8: Distribución votos con respecto al NBI

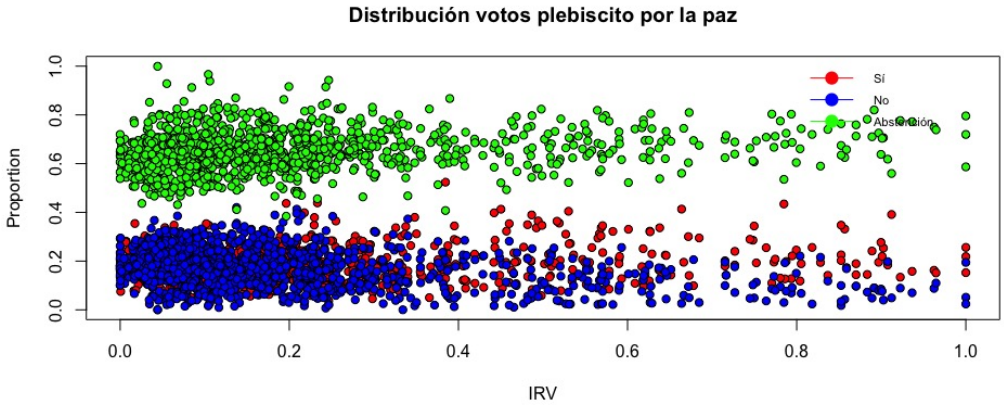

Figura 9: Distribución votos con respecto al IRV

De los parámetros estimados para NBI presentes en la figura 7 y las proporciones distribuidas según qué variables encontradas en las figuras 8 y 9 podemos decir que a medida que aumenta el valor de este índice, disminuye la proporción de votantes en cada una de las composiciones, pero en la que mayor fuerza tiene es en los votantes por el "No". Es decir, que entre más alto sea el NBI menor será la proporción de votantes por el "No". Una interpretación similar ha de añadirse al IRV, solo que en este caso, la variable no presenta disminuciones importantes en las proporciones de votantes.

Con respecto a los betas asociados a los resultados de las elecciones presidenciales para la segunda vuelta, vemos que el beta correspondiente a Juan Manuel Santos es significativo en la proporción de votantes hacia el "Sí" y la abstención, y que el beta correspondiente a Óscar Iván Zuluaga es significativo en la proporción de votantes por el "No".

A continuación se muestran varios gráficos en los cuales se puede observar la relación que se da entre cada uno de los resultados del plebiscito por la paz reales frente a los valores ajustados por el modelo.

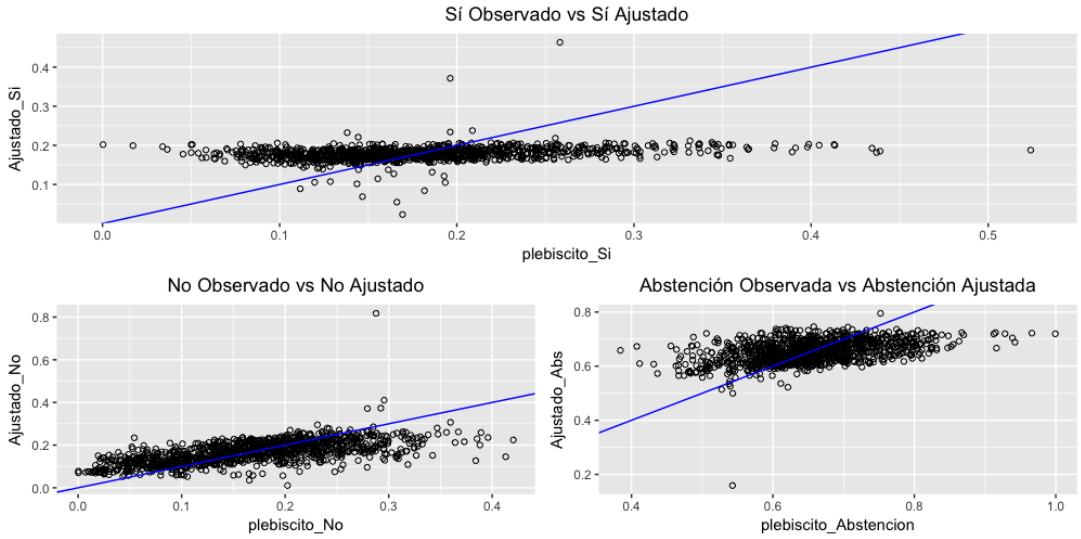

Figura 10: Validación ajuste del modelo

La relación de los valores ajustados con los resultados reales del plebiscito, presenta comportamientos no esperados, tal como lo enseña el gráfico para la composición del sí, en el que se aprecia una mala

representación de los valores reales. El ajuste del modelo presenta una mejor explicación de la respuesta "No"; en donde los valores ajustados muestran más similitud con los reales. Una vez se grafica la curva ajustada del modelo, se obtiene lo siguiente:

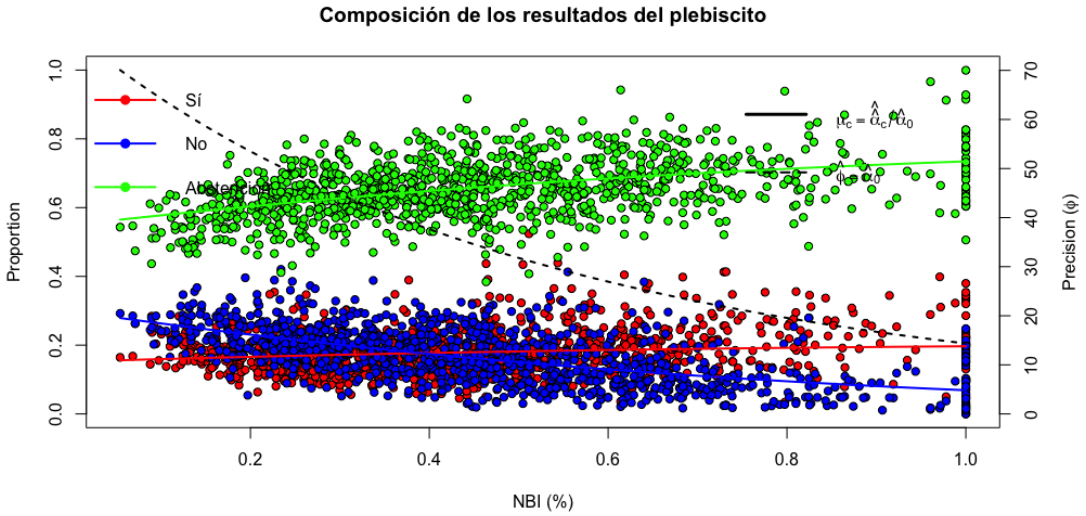

Figura 11: Validación ajuste del modelo

Ahora, veremos la comparación geográfica de los resultados del plebiscito por la paz vs los ajustados a nivel departamental:

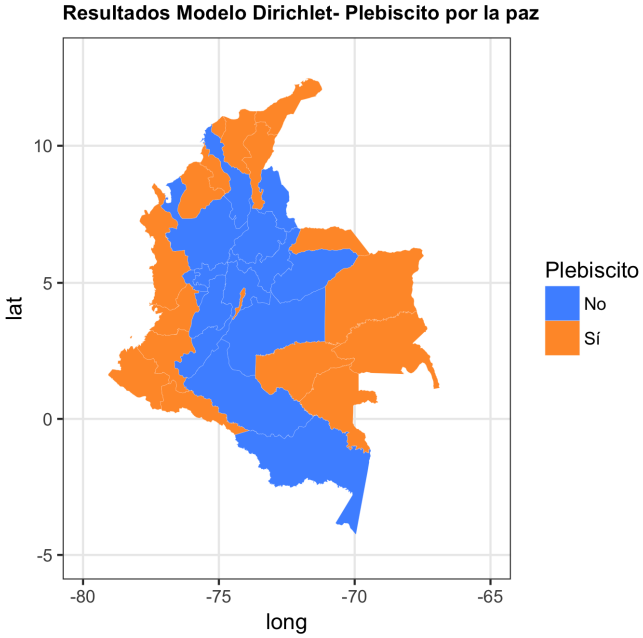

Figura 12: Resultados Regresión Dirichlet

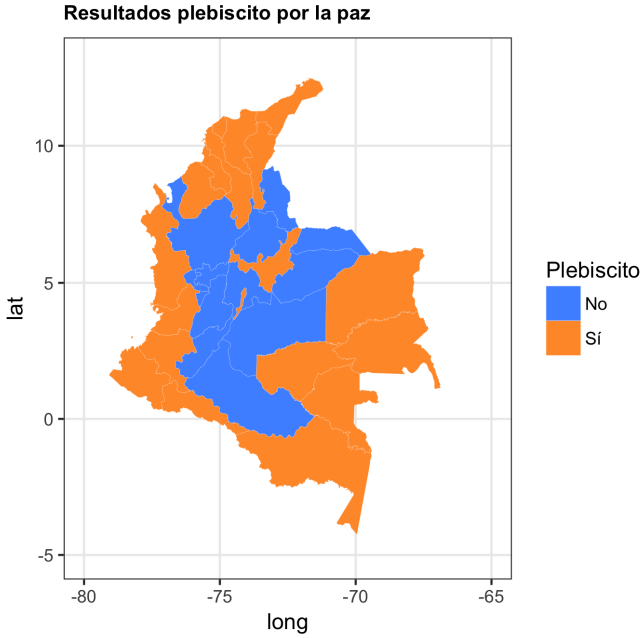

Figura 13: Resultados plebiscito por la paz

A través de este resultado, vemos que el ajuste del modelo logra explicar en su totalidad los resultados del plebiscito, perdiendo confiabilidad en el costado Sur, hacia el Amazonas.

Adicionalmente es posible plantear una conclusión de carácter general en la cual se establece una comparación de la herramienta, a través de la aplicación de los resultados a total país del modelo ajustado versus los datos reales:

Resultados a nivel nacional		
	Sí	No
Datos reales	49,76 %	50,24 %
Modelo	55,09 %	44,91 %

Tabla 3: Comparación porcentajes reales y ajustados por el modelo

Como conclusión del modelo, es posible señalar que, a pesar de que los resultados a nivel de departamento en los mapas, no es del todo errónea y que las variables incluidas son significativas, este modelo no logra reproducir de una manera acertada los resultados a nivel nacional del plebiscito por la paz.

5.2.2. Modelo de regresión lineal multivariada

Extendemos ahora el modelo de regresión a la situación en la que hemos medido m veces la respuesta $Y_1, Y_2, \ldots, Y_p$ y el mismo conjunto de r predictores $z_1, z_2, \ldots, z_r$ en cada muestra para obtener una estructura de la siguiente forma:

$$\underset{(n\times m)}{Y} = \underset{(n\times r+1)(r+1\times m)}{X\quad B} + \underset{(n\times m)}{E} \tag{19}$$

Donde Y es la matriz de n observaciones sobre m variables respuesta. X es la matriz del modelo que contiene las variables explicativas y una columna adicional que representa la constante. B es la matriz de los coeficientes de regresión, y E es la matriz de los errores, en la cual cada i-ésima fila sigue una distribución Normal m-variada con media 0 y matriz de varianzas Σ . Este modelo puede ser ajustado con la función lm (R Core Team 2016) en R, donde el lado izquierdo del modelo comprende una matriz de variables respuesta, y el lado derecho se especifica exactamente como para un modelo lineal univariante (es decir, con una sola respuesta variable).

Dado lo anterior, para poder aplicar el modelo mencionado, se aplica entonces una transformación ILR a la variable composicional respuesta "Plebiscito por la paz" y a la variable composicional explicativa "Resultados elecciones presidenciales segunda vuelta", ya que de esa manera es posible ajustar un modelo lineal convencional a cada coordenada ilr. Obteniendo el siguiente modelo con su respectiva estimación:

$$ilr\left(Plebiscito_i\right) = \begin{bmatrix} \beta_{0,si} \\ \beta_{0,no} \end{bmatrix} + \begin{bmatrix} \beta_{1,si}\times NBI \\ \beta_{1,no}\times NBI \end{bmatrix} + \begin{bmatrix} \beta_{2,si}\times IRV \\ \beta_{2,no}\times IRV \end{bmatrix} + ilr\left(Resultados2daVuelta\right)\begin{bmatrix} \beta_{3,si} \\ \beta_{3,no} \end{bmatrix} \tag{20}$$

```
Response plebiscito_Si :

Call:
lm(formula = plebiscito_Si ~ NBI + IRV_2015 + JMS_2 + OIZ_2,
    data = base_modelo)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6548 -0.1827  0.0033  0.1732  1.6414

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6148     0.0445  -13.82  < 2e-16 ***
NBI           0.3647     0.0577    6.32  3.8e-10 ***
IRV_2015     -0.0677     0.0546   -1.24     0.22
JMS_2         0.3702     0.0151   24.52  < 2e-16 ***
OIZ_2        -0.2291     0.0254   -9.01  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.315 on 1117 degrees of freedom
Multiple R-squared:  0.546,     Adjusted R-squared:  0.545
F-statistic:  337 on 4 and 1117 DF,  p-value: <2e-16
```

Figura 14: Coeficientes para el Sí

```
Response plebiscito_No :

Call:
lm(formula = plebiscito_No ~ NBI + IRV_2015 + JMS_2 + OIZ_2,
    data = base_modelo)

Residuals:
   Min     1Q Median     3Q    Max
-3.493 -0.171  0.025  0.201  1.287

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2135     0.0463   -4.61  4.5e-06 ***
NBI          -0.8409     0.0601  -13.99  < 2e-16 ***
IRV_2015     -0.1681     0.0569   -2.96   0.0032 **
JMS_2        -0.3988     0.0157  -25.37  < 2e-16 ***
OIZ_2        -0.0472     0.0265   -1.79   0.0744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.328 on 1117 degrees of freedom
Multiple R-squared:  0.626,     Adjusted R-squared:  0.624
F-statistic:  467 on 4 and 1117 DF,  p-value: <2e-16
```

Figura 15: Coeficientes para el No

Un aspecto importante que se debe tener en cuenta para efectos de análisis es que dadas las transformaciones que se le realizan a los datos, resulta complejo interpretar los coeficientes del modelo como suele hacerse con los métodos clásicos. Teniendo en cuenta esto, a continuación se muestran varios gráficos en los cuales se puede observar la relación que se da entre cada uno de los elementos de la composición del plebiscito observado frente a los valores ajustados por el modelo.

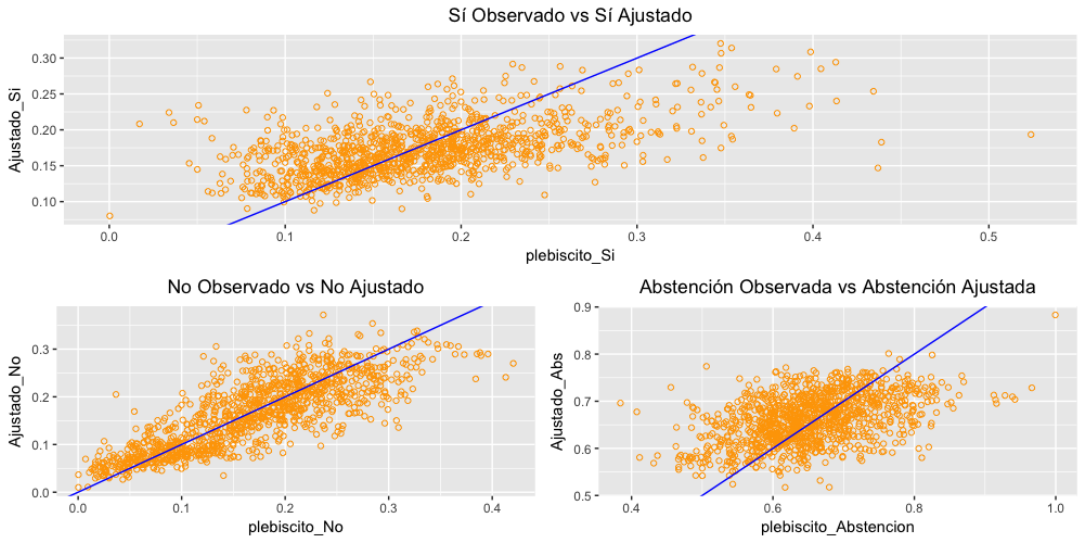

Figura 16: Validación ajuste del modelo lineal multivariado

Teniendo en cuenta los gráficos, vemos que la relación entre lo real y lo ajustado por el modelo mejora considerablemente con respecto al ejercicio anterior. Para validar estas estimaciones, haremos las comparaciones geográficas de los resultados reales del plebiscito y los ajustados por el modelo.

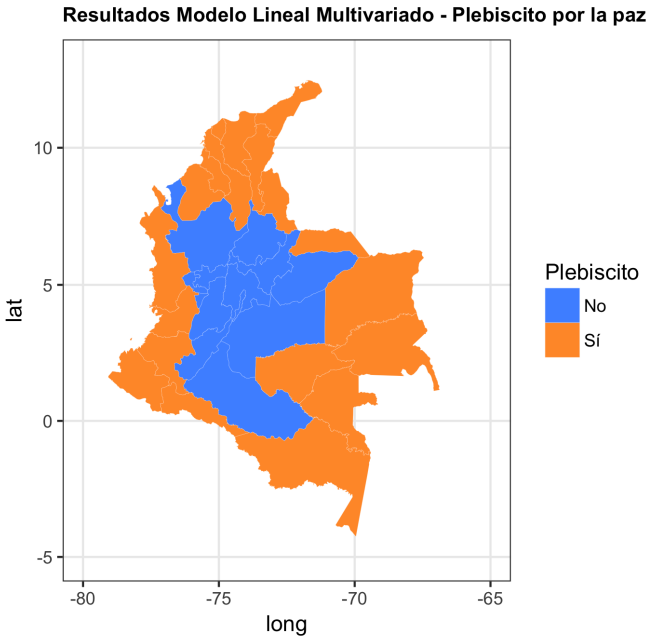

Figura 17: Resultados Modelo Lineal Multivariado

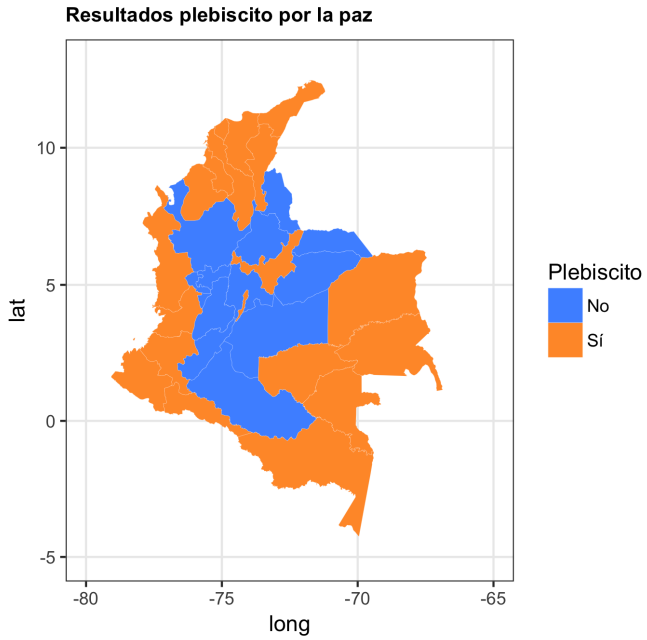

Figura 18: Resultados plebiscito por la paz

Se aprecia que este modelo presenta una mejor estimación del resultado ganador en cada uno de los departamentos salvo para Bogotá, Santander y Boyacá.

Para terminar de comparar los resultados del modelo, se realiza nuevamente la comparación de los resultados a nivel nacional.

Resultados a nivel nacional		
	Sí	No
Datos reales	49,76 %	50,24 %
Modelo	49,18 %	50,82 %

Tabla 4: Resultados Modelo Lineal Multivariado

Se aprecia que el resultado nacional estimado por el modelo se acerca mucho más a la realidad que el modelo anterior. Esto demuestra una mejora significativa en el ajuste.

5.2.3. Modelo mixto multivariado o de efectos aleatorios

Para poder realizar el ajuste del modelo mixto clásico, es necesario usar nuevamente la transformación ILR. Este modelo está compuesto de la siguiente forma:

$$ilr\left(Ple_i\right) = \begin{bmatrix} \beta_{0,si} \\ \beta_{0,no} \end{bmatrix} + \begin{bmatrix} \beta_{1,si} \times NBI \\ \beta_{1,no} \times NBI \end{bmatrix} + \begin{bmatrix} \beta_{2,si} \times IRV \\ \beta_{2,no} \times IRV \end{bmatrix} + ilr\left(2daVuelta\right) \begin{bmatrix} \beta_{3,si} \\ \beta_{3,no} \end{bmatrix} + \begin{bmatrix} \gamma_{1,si} \times DPTO \\ \gamma_{1,no} \times DPTO \end{bmatrix} \quad (21)$$

de la cual se logran obtener los siguientes $\hat{\beta}$ y $\hat{\gamma}$:

```
Scaled residuals:
   Min     1Q Median     3Q    Max
-9.456 -0.503  0.000  0.547  8.150

Random effects:
 Groups       Name                  Variance Std.Dev. Corr
 DEPARTAMENTO variableplebiscito_Si 0.0397   0.199
              variableplebiscito_No 0.0763   0.276    -0.07
 Residual                           0.0707   0.266
Number of obs: 2244, groups:  DEPARTAMENTO, 33

Fixed effects:
                               Estimate Std. Error t value
variableplebiscito_Si           -0.7055     0.0600  -11.76
variableplebiscito_No           -0.2482     0.0691   -3.59
variableplebiscito_Si:NBI        0.1159     0.0572    2.03
variableplebiscito_No:NBI       -0.6254     0.0578  -10.83
variableplebiscito_Si:IRV_2015   0.0664     0.0590    1.13
variableplebiscito_No:IRV_2015  -0.2602     0.0599   -4.34
variableplebiscito_Si:JMS_2      0.4414     0.0174   25.34
variableplebiscito_No:JMS_2     -0.4092     0.0177  -23.12
variableplebiscito_Si:OIZ_2     -0.1728     0.0240   -7.20
variableplebiscito_No:OIZ_2     -0.0933     0.0241   -3.88

Correlation of Fixed Effects:
            vrbl_S vrbl_N v_S:NB v_N:NB v_S:IR v_N:IR v_S:JM v_N:JM v_S:OI
vrblplbsc_N -0.028
vrblp_S:NBI -0.090  0.001
vrblp_N:NBI  0.001 -0.087 -0.002
v_S:IRV_201 -0.189  0.000 -0.373  0.001
v_N:IRV_201  0.000 -0.157  0.001 -0.377 -0.003
vrb_S:JMS_2 -0.371  0.001 -0.231 -0.001 -0.120  0.002
vrb_N:JMS_2  0.002 -0.331 -0.001 -0.213  0.002 -0.133 -0.003
vrb_S:OIZ_2 -0.640  0.001 -0.338  0.000  0.232  0.000  0.305 -0.001
vrb_N:OIZ_2  0.001 -0.560  0.000 -0.330  0.000  0.224 -0.001  0.305 -0.001
```

Figura 19: Ajuste del modelo mixto

Sabemos ya del caso anterior que dada la transformación aplicada a los datos, no es factible realizar una interpretación directa. Una conclusión que se puede obtener a través de los signos de los coeficientes es, por ejemplo, el signo negativo de **variableplebiscito_No:NBI** (que representa el efecto del NBI en la proporción de votos para el No) representa una disminución de los votos a medida que aumenta el NBI, cosa que ha visto a través de los modelos anteriores.

Para tener una idea adicional del ajuste del modelo, se hace uso de la función **r.squaredLR()** de la librería **MuMIn** (Barton 2016), la cual nos arroja el cálculo del coeficiente de determinación basado en el test de log-verosimilitud, siendo este valor de $R^2 = 0.93439$, una razón para creer que la variable se encuentra bien explicada por la estructura del modelo planteada.

A continuación se muestran gráficos que representan los resultados obtenidos por el modelo mixto:

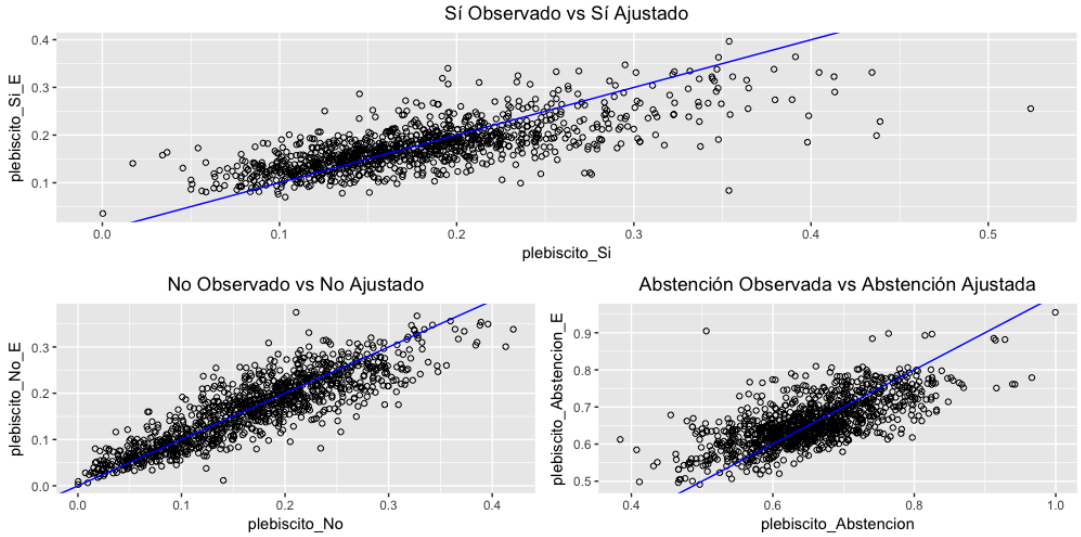

Figura 20: Ajuste modelo mixto multivariado

Del gráfico anterior vemos que este modelo presenta una relación mucho más significativa (gráficamente) que los modelos anteriores, ya que los valores ajustados por el modelo y los resultados reales muestran una relación más a fin con la recta $y = x$. Con el fin de continuar validando estas conclusiones, se presentan los mapas de los valores ajustados contra los reales.

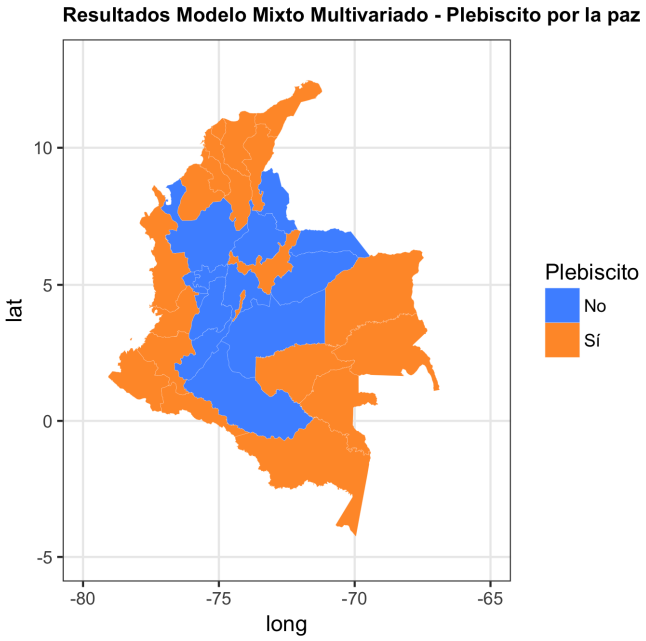

Figura 21: Resultados Modelo Mixto

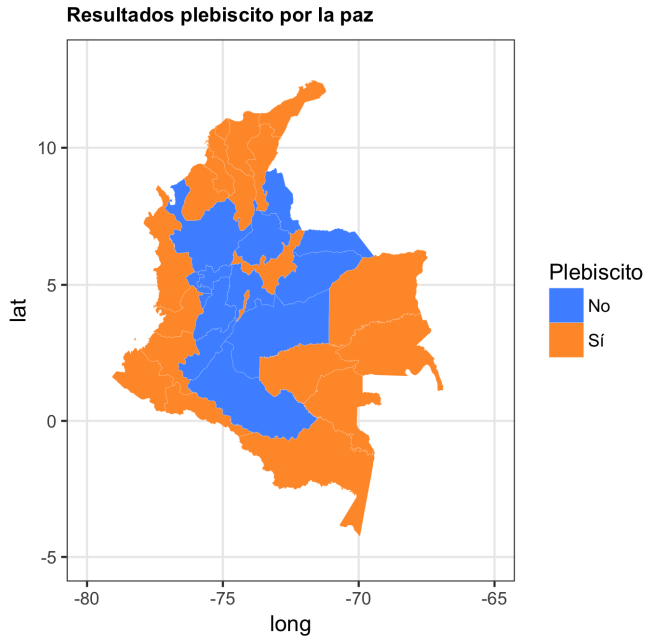

Figura 22: Resultados plebiscito por la paz

Se aprecia de los mapas con los resultados que este modelo presenta una estimación totalmente acertada del resultado ganador en el 100 % de los departamentos. Esto en total sintonía con el R^2 que indicaba una alta explicación de la variable respuesta a través de las explicativas. Dado este resultado, se revisa el comportamiento de la variable respuesta para cada una de las variables explicativas.

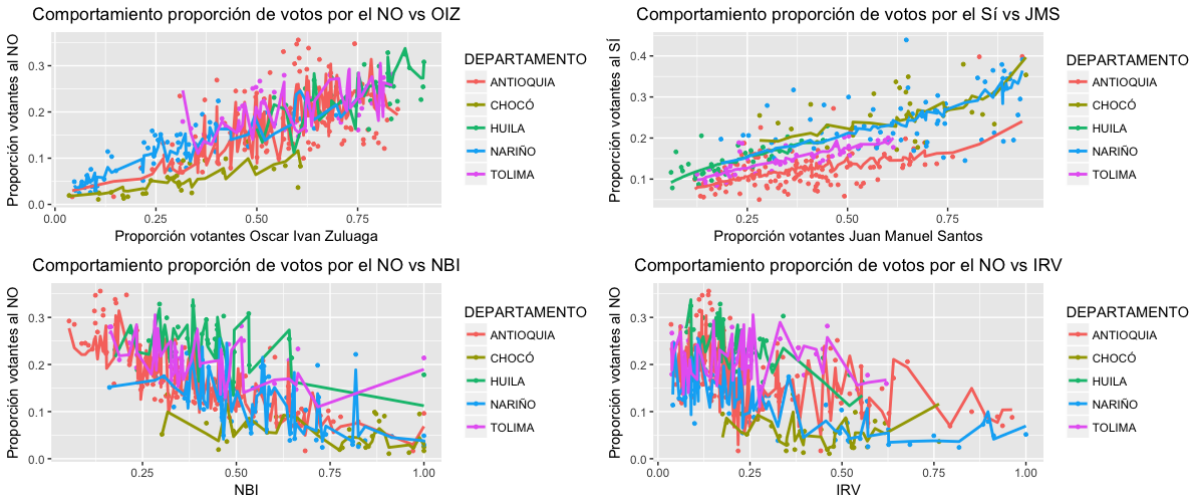

Figura 23: Consolidado Gráficos

Para terminar de analizar los resultados del modelo y dejar en claro la eficiencia de este, se realiza nuevamente la comparación de los resultados a nivel nacional.

Resultados a nivel nacional		
	Sí	No
Datos reales	49,76 %	50,24 %
Modelo	49,5 %	50,5 %

Tabla 5: Resulados Modelo Mixto Multivariado

A través de la tabla se ve una mejora considerable en la estimación de los resultados, esto sumado a la réplica del resultado ganador en cada departamento como se muestra en la figura 21, entendiendo según el modelo que el hecho de considerar una estructura de correlación entre los municipios de un mismo departamento mejora la estimación de los resultados.

6. Conclusiones

Como aporte de los resultados obtenidos del análisis composicional en el campo político de Colombia, se presentan un conjunto de conclusiones, limitaciones y propuestas de trabajo sobre esta temática.

1. La presente tesis tuvo como objetivo principal mostrar a través de las herramientas y técnicas adecuadas para el análisis composicional, el ajuste de un modelo mixto multivariado que lograra explicar bajo variables que representan la realidad social y económica de Colombia, los resultados del plebiscito por la paz del 2016.

   La implementación de este modelo comprueba que el uso de estas metodologías especiales para los datos composicionales, garantizan una excelente representación de los resultados finales reales del proceso de votación, lo anterior se ilustra en la réplica de los resultados a nivel departamental según lo mostraban los mapas comparativos entre lo que realmente sucedió y lo que nos arrojaba el modelo. La trascendencia de este modelo radica en el hecho de tener en cuenta la dependencia de los departamentos dentro de una variable aleatoria, es decir, asumir una estructura de correlación, y sobre el cual se identifica que los departamentos con más problemas sociales y económicos eran los que más estaban a favor del proceso de paz.

   Todo esto motivando entonces a realizar estudios más complejos y profundos que fortalezcan lo desarrollado hasta el momento en la teoría, en búsqueda de aportar conocimientos sobre las diversas áreas de aplicación en las que se encuentran los datos composicionales.

2. Es importante resaltar que los resultados del ejercicio desarrollado en esta oportunidad, cumplen adecuadamente con los objetivos y metodología propuestos al inicio del trabajo. Esto para aspectos tanto teóricos como prácticos.

7. Limitaciones

En el desarrollo del ejercicio se presentaron las siguientes limitaciones:

- A raíz de las transformaciones aplicadas a los datos, los coeficientes estimados de los modelos carecían de interpretación. Esto llevándonos a que la interpretación de los mismos no sea recomendable dado que se puede incurrir en conclusiones erróneas.

8. Trabajos futuros

- A partir de las conclusiones expuestas y las limitaciones que subyacen en el trabajo, una primera propuesta sería la necesidad ajustar el modelo mixto sin transformar la variable, esto modificando la verosimilitud del modelo ligándola a una distribución asociada directamente a los datos composicionales, tal como lo es la distribución Dirichlet.
- A raíz del modelo mixto y de asumir una estructura de correlación para los municipios de un mismo departamento, se planea trabajar un modelo que capture la correlación espacial de los mismos datos.

Referencias

Aitchison, J. (1986), 'The statistical analysis of compositional data'.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. & Pawlowsky-Glahn, V. (2000), 'Logratio analysis and compositional distance', *Mathematical Geology* **32**(3), 271–275.

Barton, K. (2016), *MuMIn: Multi-Model Inference*. R package version 1.15.6.
*https://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software* **67**(1), 1–48.

Billheimer, D., Guttorp, P. & Fagan, W. F. (2001), 'Statistical interpretation of species composition', *Journal of the American statistical Association* **96**(456), 1205–1214.

Broyden, C. G. (1970), 'The convergence of a class of double-rank minimization algorithms 1. general considerations', *IMA Journal of Applied Mathematics* **6**(1), 76–90.

Buccianti, A., Montegrossi, G., Tassi, F. & Vaselli, O. (2002), 'Log-contrast analysis of volcanic fluid composition: a way to check equilibrium conditions', *Terra Nostra, special issue: IAMG* **2**, 405–410.

Christ, A. (2009), 'Mixed effects models and extensions in ecology with r'.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. & Barcelo-Vidal, C. (2003), 'Isometric logratio transformations for compositional data analysis', *Mathematical Geology* **35**(3), 279–300.

Fletcher, R. (1970), 'A new approach to variable metric algorithms', *The computer journal* **13**(3), 317–322.

Goldfarb, D. (1970), 'A family of variable-metric methods derived by variational means', *Mathematics of computation* **24**(109), 23–26.

Gueorguieva, R., Rosenheck, R. & Zelterman, D. (2008), 'Dirichlet component regression and its applications to psychiatric data', *Computational statistics & data analysis* **52**(12), 5344–5355.

Henningsen, A. & Toomet, O. (2011), 'maxlik: A package for maximum likelihood estimation in R', *Computational Statistics* **26**(3), 443–458.
*http://dx.doi.org/10.1007/s00180-010-0217-1

Hijazi, R. H. & Jernigan, R. W. (2009), 'Modelling compositional data using dirichlet regression models', *Journal of Applied Probability & Statistics* **4**(1), 77–91.

Maier, M. J. (2014), Dirichletreg: Dirichlet regression for compositional data in r, Research Report Series/ Department of Statistics and Mathematics 125, WU Vienna University of Economics and Business, Vienna.
*http://epub.wu.ac.at/4077/

Maier, M. J. (2015), *DirichletReg: Dirichlet Regression in R*. R package version 0.6-3.
*http://dirichletreg.r-forge.r-project.org/

Márquez, J., Aparicio, F. et al. (2011), 'Modelos estadísticos para sistemas electorales multipartidistas en stata', *Stata Press book chapters* .

Pawlowsky-Glahn, V. & Buccianti, A. (2011), *Compositional data analysis: Theory and applications*, John Wiley & Sons.

Pawlowsky-Glahn, V. & Egozcue, J. J. (2001), 'Geometric approach to statistical analysis on the simplex', *Stochastic Environmental Research and Risk Assessment* **15**(5), 384–398.

Pearson, K. (1896), 'Mathematical contributions to the theory of evolution.–on a form of spurious correlation which may arise when indices are used in the measurement of organs', *Proceedings of the royal society of london* **60**(359-367), 489–498.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Shanno, D. F. (1970), 'Conditioning of quasi-newton methods for function minimization', *Mathematics of computation* **24**(111), 647–656.

Smithson, M. & Verkuilen, J. (2006), 'A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables.', *Psychological methods* **11**(1), 54.

Templ, M., Hron, K. & Filzmoser, P. (2011), *robCompositions: an R-package for robust statistical analysis of compositional data*, John Wiley and Sons.

Thomas, C. & Aitchison, J. (1998), The use of logratios in subcompositional analysis and geochemical discrimination of metamorphosed limestones from the northeast and central scottish highlands, *in* 'Proceedings of IAMG98, The Fourth Annual Conference of the International Association for Mathematical Geology, De Frede, Naples', pp. 549–554.

Tolosana-Delgado, R., Palomera-Román, R., Gimeno-Torrente, D., Pawlowsky-Glahn, V. & Thió-Henestrosa, S. (2002), 'A first approach to the classification of basalts using trace elements', *See Bayer, Burger, and Skala* pp. 435–440.

Van den Boogaart, K. G. & Tolosana-Delgado, R. (2013), *Analyzing compositional data with R*, Vol. 122, Springer.

Weltje, G. J. (2002), 'Quantitative analysis of detrital modes: statistically rigorous confidence regions in ternary diagrams and their use in sedimentary petrology', *Earth-Science Reviews* **57**(3), 211–253.

9. Anexo: Código en R

```
library(readxl)
library(dplyr)
library(robCompositions)
library("DirichletReg")

base[,c("plebiscito_Si","plebiscito_No","plebiscito_Abstencion")] <- DR_data(base[c("plebiscito_Si

#########################
####    Modelo LME4   ####
#########################

library(reshape2)
library(robCompositions)
library(lme4)

base <- base %>% mutate(Otros_Votos = Votos_Blanco_2+Votos_Nulos_2+Votos_No_Marcados_2) %>%
select(-Votos_Blanco_2,-Votos_Nulos_2,-Votos_No_Marcados_2)

base[,c("JMS_2","OIZ_2","Otros_Votos")] <- DR_data(base[c("JMS_2","OIZ_2","Otros_Votos")])
base_modelo <- base

base_modelo[,c("plebiscito_Si", "plebiscito_No")] <- pivotCoord(as.data.frame(base_modelo[,c("pleb
base_modelo <- base_modelo %>% select(-plebiscito_Abstencion)

base_modelo[,c("JMS_2","OIZ_2")] <- pivotCoord(as.data.frame(base_modelo[,c("JMS_2","OIZ_2","Otros
base_modelo <- base_modelo %>% select(-Otros_Votos)

base_modelo

m_base <- melt(base_modelo, measure.vars = c("plebiscito_Si", "plebiscito_No"))
m_base$DEPARTAMENTO <- as.factor(m_base$DEPARTAMENTO)
head(m_base)
options(digits = 5)
m2 <- lmer(value ~ -1 + variable + variable:NBI + variable:IRV_2015 + variable:JMS_2 + variable:OI
data = m_base)


#coeficientes de homogeneidad o correlacion intraclase
summary(m2)
anova(m2)

library(MuMIn)
r.squaredLR(m2) # Calculate a coefficient of determination based on the likelihood-ratio test (R_L

m_base_2 <- m_base
m_base_2$Predicciones <- predict(m2)


base_2 <- dcast(m_base, NBI + IRV_2015 + JMS_2 + OIZ_2 + DEPARTAMENTO ~ variable, value.var='value'
base_2 <- base_2 %>% select(one_of(colnames(base_modelo)))
all.equal(base_2,base_modelo)

base_predicciones <- dcast(m_base_2, NBI + IRV_2015 + JMS_2 + OIZ_2 + DEPARTAMENTO ~ variable, valu
base_predicciones <- base_predicciones %>% select(one_of(colnames(base_modelo)))
```

```
base_predicciones_final <- cbind(base_predicciones,pivotCoordInv(base_predicciones[,c("plebiscito_

base_predicciones_final <- base_predicciones_final %>% rename(plebiscito_Si_E = X1,
plebiscito_No_E = X2,
plebiscito_Abstencion_E = X3)

ilr_tinv <- pivotCoordInv(base_predicciones[,c("plebiscito_Si", "plebiscito_No")]) #igual a la lin



base_predicciones_final <- cbind(base_predicciones,ilr_tinv)


base_comparaciones <- cbind(base,base_predicciones_final[,8:10])
base_comparaciones <- cbind(base,ilr_tinv)

############

library(ggplot2)
library(gridExtra)
scatter_1 <- ggplot(base_comparaciones, aes(x=plebiscito_Si, y=plebiscito_Si_E)) +
geom_point(shape=1) +    # Use hollow circles
geom_smooth() + labs(title = "Sí Observado vs Sí Ajustado") +  theme(plot.title = element_text(hju

scatter_2 <- ggplot(base_comparaciones, aes(x=plebiscito_No, y=plebiscito_No_E)) +
geom_point(shape=1) +    # Use hollow circles
geom_smooth() + labs(title = "No Observado vs No Ajustado") + theme(plot.title = element_text(hjus

scatter_3 <- ggplot(base_comparaciones, aes(x=plebiscito_Abstencion, y=plebiscito_Abstencion_E)) +
geom_point(shape=1) +    # Use hollow circles
geom_smooth() + labs(title = "Abstención Observada vs Abstención Ajustada") +  theme(plot.title =

grid.arrange(scatter_1, scatter_2, scatter_3,ncol=2,layout_matrix = rbind(c(1,1), c(2,3)))

######################## Mapas ####################
library(sp)
library(RColorBrewer)
library(ggplot2)
library(maptools)
library(scales)
ohsCol2 <- readShapeSpatial("Muni.shp") #Municipios
Daa_2 <- as(ohsCol2,"data.frame")
library(dplyr)
Daa_2 %>% group_by(MPIOS) %>% summarise(cuenta=n()) %>% filter(cuenta>1)

ohsColI2 <- fortify(ohsCol2,region = "DPTO") #Cambia el ID de los registros
# ohsColI2 <- ohsColI2[ohsColI2$id==0,]
base_w <- read_excel("~/Dropbox/Tesis/BD/Base_Consolidada.xlsx",sheet = "Consolidado_Final")
grupo2 <- data.frame(ID_MUN = base_w$ID_MUN, ilr_tinv)


grupo2$DPTO <- substr(grupo2$ID_MUN,1,2)

total_votos <- base_w %>% mutate(Votos_Ple = plebiscito_Si+plebiscito_No+plebiscito_Abstencion) %
head(total_votos)

grupo2 <- left_join(grupo2,total_votos)
head(grupo2)
```

```
grupo2 <- grupo2 %>% mutate(plebiscito_Si = plebiscito_Si_E*Votos_Ple,
plebiscito_No = plebiscito_No_E*Votos_Ple)


sum(base_w$plebiscito_Si)/sum(base_w$plebiscito_Si+base_w$plebiscito_No)
sum(base_w$plebiscito_No)/sum(base_w$plebiscito_Si+base_w$plebiscito_No)

sum(grupo2$plebiscito_Si)/sum(grupo2$plebiscito_Si+grupo2$plebiscito_No)
sum(grupo2$plebiscito_No)/sum(grupo2$plebiscito_Si+grupo2$plebiscito_No)



grupo2 <- grupo2 %>% group_by(DPTO) %>% summarise(plebiscito_Si = sum(plebiscito_Si),
plebiscito_No = sum(plebiscito_No)) %>% ungroup()


variables_Plebiscito <- c("plebiscito_Si","plebiscito_No")
grupo2$Plebiscito <- colnames(grupo2[variables_Plebiscito])[apply(grupo2[variables_Plebiscito],1,v


grupo2 <- grupo2 %>% select(DPTO, Plebiscito)
colnames(grupo2)

ohsColI2 <- left_join(ohsColI2, grupo2, by = c("id" = "DPTO"))
colnames(ohsColI2)


######## Plebiscito
ohsColI2$Plebiscito <- as.factor(ohsColI2$Plebiscito)

mapCol_Plebiscito_DPTO <- ggplot() +
geom_polygon(data=ohsColI2, aes(x=long, y=lat, group = group,
fill = Plebiscito), colour ="white", size = 0.03) +
labs(title = "Resultados Modelo Mixto Multivariado - Plebiscito por la paz") +
scale_x_continuous(limits=c(-80,-65))+
scale_y_continuous(limits=c(-5,13)) + coord_fixed(1) + coord_map() +
theme_bw() +
theme(plot.title = element_text(size = 10, face = "bold"))
mapCol_Plebiscito_DPTO
```