UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

**Facultad de Estadística**
Trabajo de Grado
Enero 2017

# PLEBISCITO POR LA PAZ EN COLOMBIA: ANÁLISIS ESTADÍSTICO A PARTIR DE DATOS COMPOSICIONALES

## PLEBISCITE FOR PEACE IN COLOMBIA: STATISTICAL ANALYSIS USING COMPOSITIONAL DATA

Autor: Catalina Plata Rincón .[a]
catalina.plata@usantotomas.edu.co

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

**Resumen**

Este documento presenta algunas de las herramientas actualmente disponibles entorno al análisis y manejo de datos composicionales, con el fin de evidenciar los alcances que pueden tener estas herramientas cuando se utilizan de forma adecuada; esto mediante la revisión de la teoría que hasta el momento se encuentra disponible y desarrollando un ejercicio práctico (computacional) que hace uso de datos disponibles en cuanto a procesos electorales que se han llevado a cabo en Colombia en los últimos años.

***Palabras clave*:** Datos composicionales, transformaciones, plebiscito en Colombia.

**Abstract**

This document seeks to capture some of the tools currently available in the analysis and management of compositional data, in order to provide evidence of the great scope that can be obtained if these tools are used properly; This is done by reviewing the theory that is available till the date and also by developing a practical (computational) exercise that makes use of available data on electoral processes that have happened in Colombia in the last years .

***Keywords*:** Compositional data, logratio tranformations, plebiscite in Colombia.

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

**A mi abuelo:** Quien fue mi gran apoyo durante toda mi carrera universitaria y quien siempre ha sido un ejemplo de vida.

**A mi papás:** Quienes me permitieron tener una formación profesional y siempre creyeron en mi.

**A mi director de tesis:** Por su apoyo y enseñanzas durante mi proceso de formación en la universidad.

**A mis amigos:** Con quienes nos apoyamos mutuamente en nuestra formación profesional.

# 1. Introducción

Colombia, centenario en desarrollo de guerras internas de diversa naturaleza, ha propiciado de múltiples maneras la búsqueda de la paz y la convivencia, mediante diálogos con los contradictores y en algunos casos con soluciones parciales a las peticiones de los alzados en armas.

Se han logrado acuerdos y procesos de desmovilización e integración al sistema, de diversos grupos y, lo más reciente en esa dirección son sin duda los esfuerzos para terminar con la guerra de más de 50 años con las FARC. Las conversaciones en este caso lograron plasmarse en un documento de acuerdo, cuyo contenido por decisión de las partes, Gobierno y Farc sería avalado, por el voto popular, mediante un plebiscito.

Una vez divulgado los textos del acuerdo y establecida la fecha del plebiscito, los grupos interesados en su aprobación, así como los interesados en su rechazo adelantaron acciones de la más variada naturaleza para lograr sus objetivos. Por su parte, las empresas encuestadoras del país en seguimiento a este proceso efectuaron encuestas y estudios los cuales en todos los casos apuntaban a señalar que en el plebiscito del 2 de octubre se alcanzaría un SI a los acuerdos.

Este no fue el caso y lo que ocurrió es que hubo una mayoría de votantes colombianos que dijeron NO a los acuerdos firmados y con ello se puso en evidencia que los estudios de las Empresas encuestadoras estaban alejados de la realidad.

Considerando entonces la falta de coherencia entre la realidad que se plasmó en el plebiscito, con lo que lograron pre-establecer, mediante procedimientos estadísticos clásicos, las firmas encuestadoras, surge la motivación para este trabajo de grado, el cual pretende analizar la información en forma integral, mediante la perspectiva de considerar el conjunto de datos composicionales. Lo que se plantea, en este caso, es analizar los resultados del ejercicio de votación del plebiscito, basados en las cifras de las tres alternativas posibles (abstención, SI, NO), en lugar de analizar solo uno de los posibles resultados, sacando conclusiones basadas en la consideración en forma separada de las cifras que aportan cada uno de las posibilidades, lo cual llevó indudablemente a un visión sesgada de la realidad.

# 2. Marco teórico

Los datos composicionales hacen referencia a cualquier vector $x$ para el cual sus componentes pueden estar expresados como partes de un total, ya sean: porcentajes, partes por millón, proporciones, entre otras, y las cuales están sujetas a la restricción que indica que la suma de sus componentes sea la unidad, o para un caso más general, una constante definida.

## 2.1. Principios del análisis composicional

Para comprender de manera correcta los temas se desarrollan a lo largo del documento, es importante hacer claridad en torno los siguientes términos:

- **Correlación espuria:** Este concepto surge en 1987 gracias a que Pearson expuso las dificultades existentes al analizar de manera idónea las covarianzas o coeficientes de correlación cuando se trabaja con datos composicionales, principalmente por el hecho de tener correlaciones "falsas" entre las partes de una composición. Puntualmente, al calcular una matriz de coeficientes de correlación sobre un conjunto de datos composicionales, ya que es inevitable que si aumenta una componente las restantes deben disminuir para efecto de cumplir con la restricción de suma constante, por lo anterior se observan falsas correlaciones negativas (correlaciones espurias), sin importar los datos que se estén analizando.

Es necesario tener presente que, si bien los **datos composicionales** (DaCo) hacen referencia a las partes de un todo, los cocientes entre componentes son suficientes para resumir la información total de los datos; por lo anterior surgen los principios que se explican a continuación los cuales deben ser tenidos en cuenta a la hora de trabajar con este tipo de datos:

- **Invarianza a la escala:** Este concepto afirma que la información contenida en un conjunto de datos escalado (como ocurre cuando se analizan porcentajes) es la misma que contiene el conjunto de datos sin escalar. Por lo anterior y para facilitar la interpretación de los análisis que se quieran llevar a cabo, suele ser más común trabajar con los datos ya escalados. La forma más usual para realizar este procedimiento es normalizando el vector, de tal manera que los componentes sumen una constante dada k= 1, 100, 1000, $10^6$ o cualquier otra constante positiva.

  Este procedimiento se conoce como la operación de cierre y se muestra a continuación:

$$\mathcal{C}_x = \left( \frac{kx_1}{\sum_{i=1}^{D} x_i}, \frac{kx_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{kx_D}{\sum_{i=1}^{D} x_i} \right), \quad \text{en donde } x = (x_1, x_2, ..., x_D) \tag{1}$$

  Los componentes del vector cerrado se llaman partes relativas a un total $k$ y así mismo el conjunto de vectores con $D$ componentes positivas (llamados composiciones) que suman a la constante $k$, forman el *simplex de D-partes*, denotado por $S^D$.

- **Coherencia subcomposicional:** Este concepto radica en entender que el análisis relativo a un subconjunto de las partes no tiene que depender de las otras partes no involucradas.

  De acuerdo a este principio, se debe garantizar que el estudio de una subcomposición, no genere resultados contradictorios a los obtenidos a partir de la composición completa.

- **Invarianza de permutación:**

  Las conclusiones de un análisis composicional no debe depender del orden de las partes o componentes.

## 2.2. Algebra para datos composicionales

### 2.2.1. Permutación y potencia:

Como lo mencionan Pawlowsky & Buccianti (2011), una perturbación $p = (p_1, p_2, ..., p_d)$ es un operador de escala diferencial, análogo a la adición entre números reales que cuando se aplica a una composición $x = x_1, ..., x_D$ produce la composición:

$$X = p \oplus x = \mathcal{C}(p_1 x_1, ..., p_D x_D), \tag{2}$$

en donde $\mathcal{C}$ es el operador de cierre descrito en (1) que escala los elementos para asegurar que se mantenga en la unidad de simplex $S^D$.

El espacio $S^D$ dotado con la operación de perturbación forma un grupo desde el punto de vista del álgebra abstracta con una perturbación identidad $e = (1/D, ..., 1/D)$. Adicional a esto vale la pena mencionar que cualquier elemento de $S^D$ (composición) puede ser expresado como la suma ($\oplus$) de dos composiciones.

Así como con la permutación, se puede definir la operación de "potencia" entre una constante $a$ y una composición $x$ como el análogo a la multiplicación por escalares en álgebra como se muestra a continuación:

$$X = a \odot x = \mathcal{C}(x_1^a, ..., x_D^a) \tag{3}$$

Basados en los conceptos anteriores, se define la métrica simplicial o distancia de Aitchison de la siguiente manera:

$$d_a(x, y) = \left\{ \sum_{i=1}^{D} \left[ \log \frac{x_i}{g_m(x)} - \log \frac{y_i}{g_m(y)} \right]^2 \right\}^{1/2}, \tag{4}$$

donde $g_m$ es la media geométrica de los componentes. Estos conceptos son de vital importancia ya que son la base del desarrollo teórico entorno a los DaCo.

### 2.2.2. Ejemplificación:

Dada la importancia de estos conceptos y con el fin de proporcionar mayor claridad en cuanto a los conceptos ya vistos (operación de cierre, perturbación y potencia), a continuación se muestra una breve explicación de ellos:

En el primer cuadro se pueden observar dos vectores $x$ y $y$ los cuales cada uno tienen asociados su marginal respectiva, estos dos vectores servirán para mostrar como calcular las operaciones ya mencionadas:

**Algebra de datos composicionales: operaciones básicas**

| | DATOS ORIGINALES | | | | | |
|---|---|---|---|---|---|---|
| INDIVIDUO | *var 1* | *var 2* | *var 3* | *var 4* | *var 5* | *Marginal* |
| x | 69207 | 271400 | 109994 | 132519 | 43931 | 627051 |
| y | 576 | 3029 | 318 | 501 | 86 | 4510 |

Tabla 1: Vectores originales ejemplo algebra para CoDa

A continuación se muestra la operación de cierre:

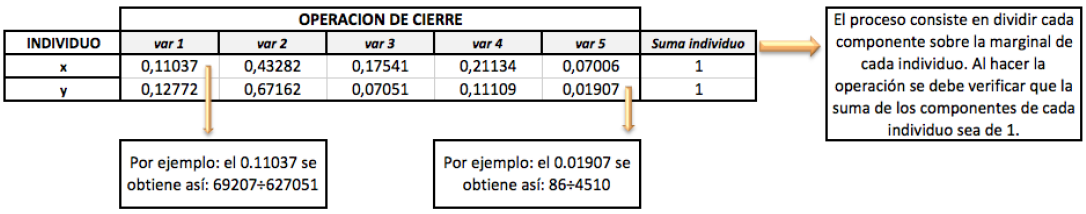



Tabla 2: Ejemplo cierre algebra para CoDa

Como bien se observa en la tabla anterior, aunque la operación de cierre es un procedimiento fundamental cuando se trabaja con DaCo, su aplicación es a su vez bastante directa. En el siguiente cuadro se puede apreciar el ejemplo para el caso de la operación de permutación y potencia:

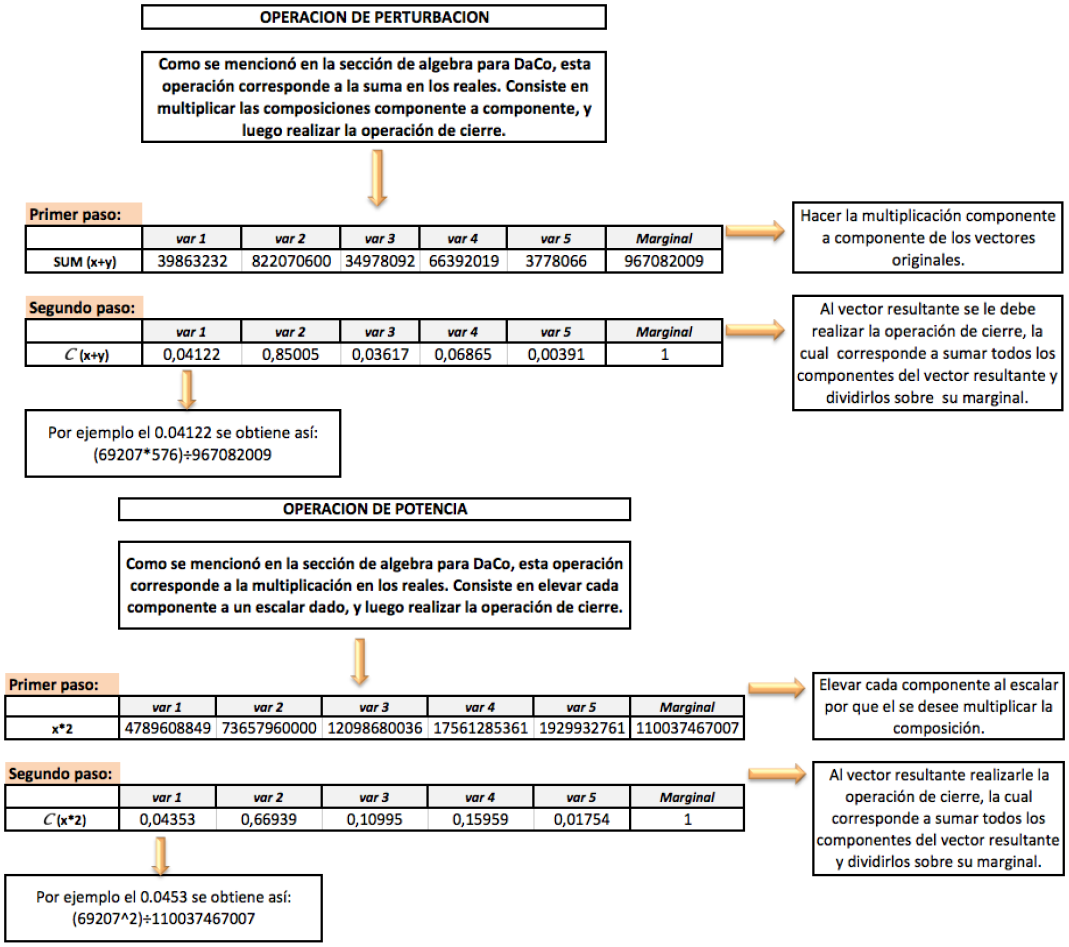



Tabla 3: Ejemplo permutación y potencia algebra para CoDa

Para estas dos operaciones hay que tener muy claro que la operación de cierre se realiza una vez se hayan completado los demás pasos.

## 2.3. Geometría del simplex de D-partes

Aitchison (1986) propuso representaciones adecuadas y completas para las composiciones mediante el uso de un grupo de logaritmos de cocientes, de tal forma que la información contenida en la composición original fuera la misma información contenida en el conjunto de logaritmos de cocientes.

### 2.3.1. Transformación ALR

Un primer acercamiento a estas representaciones fue la transformación del log-ratio aditivo (ALR), en la cual si $x$ es una composición del simplex, $S^D$, se tiene entonces:

$$alr\left(x\right) = \ln\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots \frac{x_{D-1}}{x_D}\right), \tag{5}$$

donde el logaritmo natural $ln$ se aplica componente a componente, lo cual lleva a que el i-ésimo componente de alr este dado por:

$$alr_i\left(x\right) = \ln\left(\frac{x_i}{x_D}\right) \tag{6}$$

Esta transformación se puede invertir lo cual permite reducir las operaciones de perturbación y potencia a las operaciones de suma y producto en números reales de la siguiente forma:

$$alr\left(\left(\alpha \odot x\right) \oplus \left(\beta \odot y\right)\right) = \alpha \cdot alr\left(x\right) + \beta \cdot alr\left(y\right), \tag{7}$$

para cualquier composición $x$, $y$ y cualquier constante real $\alpha$ y $\beta$.

Sin embargo, esta transformación presenta una desventaja ya que no es invariante a las permutaciones de componentes (la cual es una de las 3 propiedades mencionadas en la sección 2.2), debido a que un cambio en el orden de los componentes produce a su vez un cambio en el denominador de cada cociente.

### 2.3.2. Transformación CLR

Teniendo en cuenta lo anterior Aitchison (1986) introdujo la transformación log-ratio centrada CLR, la cual representa una composición con $D$ componentes, y se define como se muestra a continuación:

$$v = clr\left(x\right) = \ln\left[\frac{x_1}{g_m\left(x\right)}, \frac{x_2}{g_m\left(x\right)}, \ldots, \frac{x_D}{g_m\left(x\right)}\right], \qquad g_m\left(x\right) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}, \tag{8}$$

donde los $D$ coeficientes de $clr\left(x\right) = \ln\left(x_i/g_m\left(x\right)\right)$ son log-ratios de contraste. En esta transformación $clr(x)$, la composición $x$ se recupera con la transformación $clrinversa$:

$$x = clr^{-1}\left(v\right) = \mathcal{C}\exp\left(v\right), \tag{9}$$

donde la función exponencial se aplica componente a componente a $v = clr(x)$. De forma similar a lo visto para la transformación $alr$, la perturbación, y la potencia en $S^D$ corresponden a la suma y al producto en el espacio $R^D$ real D-dimensional:

$$clr\left((\alpha \odot x) \oplus (\beta \odot y)\right) = \alpha \cdot clr\left(x\right) + \beta \cdot clr\left(y\right) \tag{10}$$

Un aspecto negativo de esta transformación es que los componentes cambian cuando se trabaja con una subcomposición, debido a que al tomar composiciones distintas así mismo se presenta un cambio en el denominador de cada cociente.

Como lo afirman Pawlowsky & Buccianti (2011) la representación *clr* de composiciones se puede usar con el fin de definir una estructura métrica en el simplex definiendo las operaciones de producto interno, norma y distancia en $S^D$ como se muestra a continuación:

- Producto interno: $\langle x, y\rangle_a = \langle clr\left(x\right), clr\left(y\right)\rangle$
- La norma: $\|x\|_a = \|clr(x)\|$
- La distancia: $d_a\left(x, y\right) = d\left(clr\left(x\right), clr\left(y\right)\right),$

donde $\langle\cdot\rangle$, $\|\cdot\|$, $d(\cdot,\cdot)$, denotan el producto interno euclidiano, la norma y la distancia ordinarios. Teniendo en cuenta lo anterior se define así mismo la distancia de Aitchison como:

$$d_a\left(x, y\right) = \sqrt{\sum_{i=1}^{D}\left[clr_i\left(x\right) - clr_i\left(y\right)\right]^2} \tag{11}$$

Estos 3 conceptos (producto interno, norma y distancia) respetan los principios descritos en la sección 2.1, luego son libres de restricciones y junto con la perturbación y la potencia proveen una estructura en el simplex, llamada **Geometría simplicial de Aitchison**.

Lo anterior permite e incentiva la busqueda de metodologías que den posibilidad de explotar y utilizar todas las propiedades de los espacios euclidianos como las bases ortonormales o las proyecciones ortogonales en el análisis de DaCo. Por lo tanto es necesario entonces definir y construir bases ortonormales y sus correspondientes coordenadas. (Pawlowsky & Buccianti 2011)

### 2.3.3. Transformación ILR

Teniendo en cuenta que la definición de una base ortonormal para $S^D$ es un conjunto de composiciones $e_1, e_2, ..., e_{D-1}$ tal que $\langle e_i, e_j\rangle_a = 0$ para $i \neq j$ y $\|e_i\|_a = 1$; entonces para una base fija las coordenadas de una composición se obtienen utilizando la función:

$$x^* = ilr\left(x\right) = \left(\langle x, e_1\rangle_a, \langle x, e_2\rangle_a, ..., \langle x, e_{D-1}\rangle_a\right).$$

Así mismo la anterior función tiene como inversa: $x = ilr^{-1}\left(x^*\right) = \oplus_{j=1}^{D-1} x_j^* \odot e_j$.

Egozcue et al. (2003) definió la construcción de coordenadas ortonormales como **isometric log-ratio transformation** (ilr), debido a que las coordenadas $x_j^* = ilr_j\left(x\right)$ son log-ratios de contraste y son isométricas, por lo tanto se tiene:

$$\begin{aligned} ilr\left((\alpha \odot x) \oplus (\beta \odot y)\right) &= \alpha \cdot ilr\left(x\right) + \beta \cdot ilr\left(y\right) \\ \langle x, y\rangle_a &= \langle ilr\left(x\right), ilr\left(y\right)\rangle \\ \|x\|_a &= \|ilr(x)\| \\ d_a\left(x, y\right) &= d\left(ilr\left(x\right), ilr\left(y\right)\right) \end{aligned} \tag{12}$$

Para este caso el producto interno, la norma y la distancia entre los vectores de coordenadas *ilr* corresponden a un espacio real de dimensión $D-1$, el cual es isomórfico a $S^D$. (Entendiendo isomórfico como el hecho de que ambos espacios tienen la misma estructura.)

Por último como lo afirman van den Boogaart & Delgado (2013) a modo de comentarios a tener en cuenta a la hora de decidir que transformación usar se tiene lo siguiente: 1. La transformación alr no se debe usar en caso de que las distancias, ángulos y formas están involucrados, ya que los deforma, 2: la transformación clr produce matrices de covarianza singulares, y esto puede llegar a ser una fuente de problemas si el método estadístico utilizado necesita ser invertido, 3: la transformación ilr tiene el problema de que cada coordenada puede implicar muchas partes, lo que hace practicamente imposible realizar alguna interpretación.

## 2.4. Distribuciones de probabilidad

Con fines ilustrativos a continuación se podrán encontrar las distribuciones más representativas en el campo de los datos composicionales.

### 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 $\Sigma$, 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:

$$f(x;\mu_v,\Sigma_v)=\frac{1}{\sqrt{(2\pi)^{(D-1)}\cdot|\Sigma_v|}}exp\left[-\frac{1}{2}\left(ilr(x)-\mu_v\right)\Sigma_v^{-1}\cdot\left(ilr(x)-\mu_v\right)^t\right], \tag{13}$$

en donde $\mu_v$ y $\Sigma_v$ son respctivamente 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 $\theta$ y dispersión $\beta_v$ (matriz cuadrada de orden $(D-1)$) si su log-densidad se puede expresar como:

$$\log f(x;\theta,\beta_v)=-k(\theta,\beta_v)+(\theta-1)\log(x)^t+ilr(x)\,\beta_v\,ilr(x)^t \tag{14}$$

en lo cual $k(\theta,\beta_v)$ 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:

- $\beta_v$ es una forma cuadrática simétrica definida negativa y $\sum_{i=1}^{D}\theta_i\geq 0$
- $\beta_v$ 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:

$$\begin{aligned} f\left(x;\alpha,m,\Sigma_v\right) = exp\left[-k\left(\alpha,m,\Sigma_v\right)\right] \times exp\left[\left(\alpha-1\right)\cdot\log\left(x\right)\cdot 1^t\right] \\ \times exp\left[-\frac{1}{2} ilr\left(x\ominus m\right)\cdot\Sigma_v^{-1}\cdot ilr^t\left(x\ominus m\right)\right], \end{aligned} \tag{15}$$

donde:

- $\Sigma_v$ 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 cierre descrito en (1) a un vector $D$ independiente, igualmente escalado (es decir, con el mismo parámetro $\lambda$) por variables que se distribuyen Gamma.

Por lo anterior si $x_i \sim \Gamma\left(\lambda,\alpha_i\right)$ luego $z = \mathcal{C}\left[x\right] \sim \mathscr{D}i\left(\alpha\right)$, 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:

$$f\left(z;\alpha\right) = \frac{\Gamma\left(\alpha_0\right)}{\Gamma\left(\alpha_1\right)\ldots\Gamma\left(\alpha_D\right)}\frac{z_1^{\alpha_1}\ldots z_D^{\alpha_D}}{z_1\ldots z_D}, \tag{16}$$

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

Para mayores detalles acerca de las tres distribuciones mencionadas remitirse a (van den Boogaart & Delgado 2013).

## 2.5. Modelos para datos composicionales

En la literatura actual se pueden encontrar plasmadas distintas situaciones las cuales podrían presentarse a la hora de querer ajustar un modelo cuando se trabaja con DaCo, algunas de ellas son:

- Cuando la parte independiente es composicional
- Cuando la parte dependiente es composicional
- Cuando la composición se encuentra tanto en la parte dependiente como en la independiente

(∗) Para ahondar en las tres opciones mencionadas referirse a (van den Boogaart & Delgado 2013).

Luego al igual que cuando se trabaja con datos "clásicos" también se pueden ajustar modelos lineales a datos composicionales en donde la variable respuesta sea una composición y en las variables explicativas se tengan tanto composiciones como variables reales, la notación es la siguiente (este caso se explica más detalladamente dado que es el modelo que se usará más adelante en la parte computacional):

$$ilr\left(Y\right) = a + ilr\left(x\right) \cdot B + a + X \cdot b + \varepsilon_i, \tag{17}$$

- en donde $a$ y $b$: Son constantes composicionales.
- $Y$ es una composición como variable respuesta.
- $x$ es una composición como variable explicativa.
- $X$ es una variable real explicativa.
- $B$ es una matriz cuadrada que representa una transformación lineal entre composiciones en el espacio ilr.

La implementación computacional se presenta más adelante en la sección 5.

## 2.6. Estadística multivariada

En esta sección se mostrará como funcionan el anális de cluster y el análisis discriminante para DaCo.

### 2.6.1. Análisis de cluster

Del análisis de cluster se sabe que es una técnica que busca de alguna forma agrupar o clasificar los individuos originales en grupos, por esta razón se pasa de centrar la atención en cada uno de los individuos a los grupos que se busca definir, luego mediante una serie de variables medidas para cada uno de los individuos y con respecto a estas variables se mide la similitud entre ellos.

Con el fin de hallar estos grupos se deben medir distancias entre los individuos y como lo dicen van den Boogaart & Delgado (2013) la mayoría de las medidas de distancia para conjuntos de datos multivariados pueden ser generalizadas a las composiciones si se aplican al conjunto de datos ya transformado; adicionalmente al igual que con los métodos clásicos la distancia más usada es la euclideana y se tiene como la raíz cuadrada de la suma de las diferencias al cuadrado entre las coordenadas de los dos individuos que se comparan.

Lo anterior resulta ser equivalente en DaCo a la distancia de Aitchison ya mencionada en (4):

$$d_a\left(x,y\right) = \left\{\sum_{i=1}^{D}\left[\log\frac{x_i}{g_m\left(x\right)} - \log\frac{y_i}{g_m\left(y\right)}\right]^2\right\}^{1/2}$$

La implementación computacional se ve más adelante en la sección 5.

### 2.6.2. Análisis discriminante

Como se mencionó anteriormente una de los beneficios que brinda el trabajar con coordenadas es que esto permite aplicar cualquier método estadístico clásico al conjunto de datos de coordenadas ortonormales

que se tenga; por lo tanto desde que se tenga un conjunto transformado mediante *ilr* esto permite usar las técnicas ya conocidas como LDA o QDA pero ahora a un conjunto de DaCo. Como bien se sabe en el análisis discriminante se usan normas de discriminación para hacer la debida asignación de las observaciones a un grupo en particular. Se muestra a continuación de manera breve la regla bayesiana para el caso de los DaCo. En el método bayesiano se tienen las siguientes dos expresiones:

**Análisis discriminante cuadrático:**

$$\text{QDA} = d_j^Q(z) = -\frac{1}{2}\ln\left[det\left(\Sigma_j\right)\right] - \frac{1}{2}\left(z-u_j\right)^t\Sigma_j^{-1}\left(z-u_j\right) + \ln\left(e_j\right) \tag{18}$$

en donde $u_j$ es el vector de medias, $\Sigma_j$ es la matriz de covarianza para cada uno de los grupos $j = 1, ..., g$. Y para el caso en el que se tome $\Sigma_j = \ldots = \Sigma_g$ se tiene:

**Análisis discriminante lineal:**

$$\text{LDA} = d_j^L = u_j^t\Sigma^{-1}z - \frac{1}{2}u_j^t\Sigma^{-1}u_j + \ln\left(p_j\right) \tag{19}$$

Algo que vale la pena mencionar es que a pesar de que el método LDA presenta supuestos más restrictivos, este método requiere menos parámetros a estimar, por lo tanto, se presentan menores fallas de clasificación errónea y así mismo se puede evitar el sobreajuste de los datos en comparación con el método de QDA. Por último vale la pena resaltar que el uso de estos dos métodos depende en sí del tipo de datos con los cuales se este trabajando.

La implementación computacional se ve más adelante en la sección 5 adicionalmente para mayores detalles acerca de esta metodología remitirse a (Pawlowsky & Buccianti 2011).

# 3. Objetivos

## 3.1. Objetivo General

Presentar un resumen actualizado referente al análisis de datos composicionales, ilustrando el manejo de estas herramientas en R con los resultados de las votaciones presidenciales en la primera vuelta del año 2014 en Colombia así como con los resultados del plebiscito por la paz del año 2016.

## 3.2. Objetivos específicos

- Construir un compendio de herramientas para el análisis de datos composicionales.
- Aplicar las herramientas existentes sobre datos composicionales a un conjunto de datos reales.
- Ilustrar el manejo de los datos composicionales en R.

# 4. Metodología

1. Para el desarrollo de este estudio y como base importante para ejecución del trabajo, se llevó a cabo una revisión de la bibliografía existente relativa a la teoría y las técnicas que actualmente se utilizan para llevar a cabo la investigación estadística en base de datos composicionales. Como producto de ello se elaboró un compendio resumido 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.

2. Una vez realizada la revisión de la documentación existente alrededor de los desarrollos teóricos, se avanzó en el examen de las herramientas computacionales que hasta la actualidad han sido desarrolladas y utilizadas para el análisis de datos composicionales para establecer así su aplicabilidad en un campo de interés particular como lo es el análisis de los resultados electorales en Colombia. Específicamente se trabajó con el paquete "compositions" el cual proporciona funciones para el análisis coherente de los mismos siguiendo los lineamientos propuestos por Aitchison y Pawlowsky-Glahn.

3. Con los aportes y lo que significó la comprensión sobre estas metodologías a partir de la revisión teórica y computacional, se lograron establecer las condiciones que se requieren para proceder a ejemplificar estos conocimientos utilizando bases de datos reales, con el fin de evidenciar cual es la forma adecuada a la hora de analizar este tipo de datos y la aplicabilidad en una área particular.

4. A partir del conocimiento logrado, tanto en términos teóricos y prácticos, fue posible elaborar y plantear una serie de ideas, sugerencias y propuestas pueden ser desarrolladas a partir de investigaciones de diversa naturaleza para desarrollar en un futuro cercano.

5. 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

Este ejercicio de análisis composicional de datos electorales, fue planteado de manera específica para explicar el resultado obtenido en el plebiscito por la paz y, su desarrollo, fue soportado en una serie de datos sobre las tendencias y resultados estadísticos de procesos electorales similares aunque no relacionadas directamente pero que, considerados y utilizados de manera conjunta en el marco de la teoría de análisis composicional, pueden llegar a explicar los resultados que se obtuvieron en el proceso electoral ya mencionado.

Específicamente las variables que se tuvieron en cuenta a lo largo de todo el ejercicio fueron las siguientes:

- Cifras del plebiscito por la paz
- Cifras de la primera vuelta de las elecciones presidenciales 2014
- Cifras de la segunda vuelta de las elecciones presidenciales 2014
- Cifras de proyección de población al 2014
- Cifras del potencial de votantes del 2014
- Cifras del índice de necesidades básicas insatisfechas

La información utilizada proviene de dos fuentes distintas: las cifras de los tres procesos electorales y del potencial de votantes provienen de la registraduría nacional del estado civil de Colombia, y las otras variables del DANE.

## 5.1. Análisis descriptivos

Para los desarrollos computacionales que se observan en esta sección se implementó la librería "compositions".

**Cierre de cada composición:**

```
[1] 0.1103690 0.1277162 0.3177005 0.5167653 0.2170613 0.1108218
```

```
[1] 0.2820990 0.1820280 0.3520000 0.5077063 0.3817844 0.1797776
```

```
[1] 0.16921758 0.08623902 0.11145997 0.21447602 0.09524245 0.07833054
```

Este primer resultado es una breve mirada a lo que resulta de realizar el proceso de escalamiento de cada una de las tres variables de tipo composicional del análisis, proceso que como ya se dijo anteriormente se debe tener claro antes de realizar cualquier procedimiento.

### 5.1.1. Gráficos descriptivos: Diagramas ternarios

**Composición de la primera vuelta:**

La utilización de los gráficos como método descriptivo en el campo de análisis estadístico resulta ser un recurso muy importante, ya que con el soporte de este instrumento se puede lograr una visión clara del comportamiento de los datos con los que se trabaja antes de implementar otros métodos más elaborados.

Este primer tipo de gráficos está conformados por una parte fija, la cual es para cada caso la media geométrica de todos los candidatos excluyendo los que ya estén siendo mostrados en cada uno de los gráficos. (La parte fija se denota como $*$)

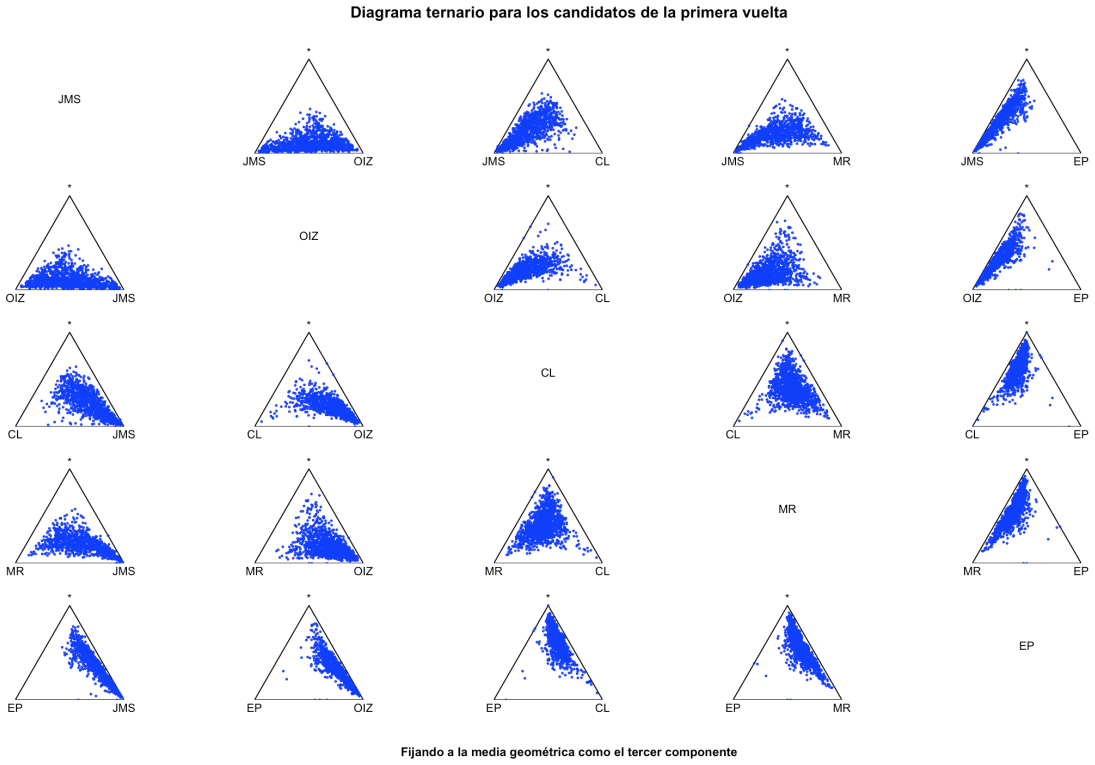



Figura 1: Visualización del diagrama ternario para la primera vuelta de las elecciones presidenciales con la media geométrica como el componente fijo.

En el gráfico anterior se pueden evidenciar varios comportamientos que constituyen la expresión visual de lo que representa el contenido del contexto de los datos:

- A partir del gráfico anterior se evidencia que en aquellos gráficos ternarios donde no participa Juan Manuel Santos (JMS) y/o Oscar Iván Zuluaga (OIZ), los puntos que denotan a los municipios de Colombia tienen a ubicarse cercanos al promedio geométrico, esto se debe a que cuando los candidatos no se ven explicitos en el gráfico es porque están incluidos en este último.
- Por otro lado, en los gráficos donde aparece alguno de los dos candidatos (JMS o OIZ) los puntos tienden a ubicarse cercanos a ellos indicando la preferencia de los municipios a inclinarse por alguno de ellos el día de las elecciones.

Una variación del gráfico anterior se muestra a continuación, en donde ahora se toma como parte fija uno de los candidatos, en este caso Juan Manuel Santos:

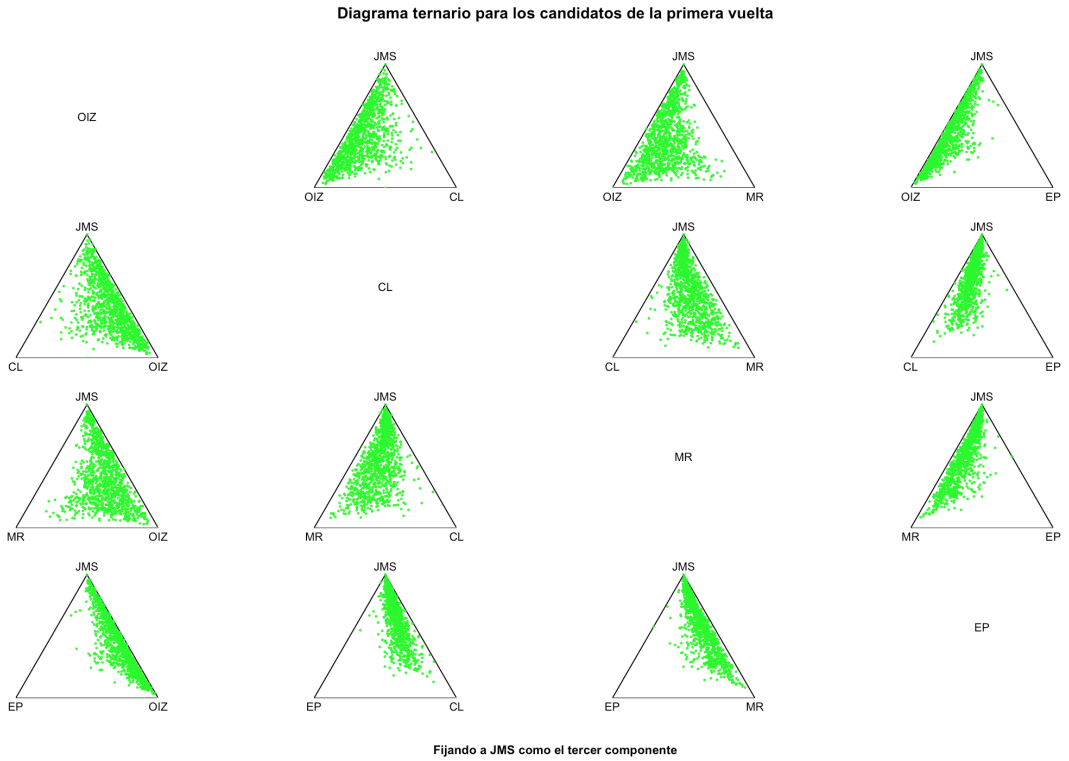


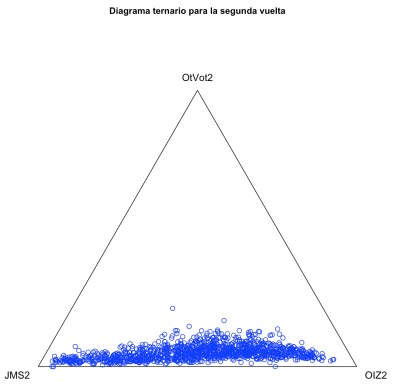

Figura 2: Visualización del diagrama ternario para la primera vuelta de las elecciones presidenciales con el candidato JMS como el componente fijo.

Del anterior gráfico se pueden mencionar varios aspectos:

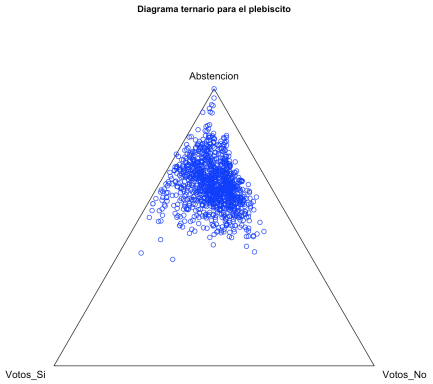
- En los gráficos en los que OIZ no está siendo considerado, se evidencia un comportamiento que sin importar cuales sean los candidatos que se estén considerando la tendencia de la ubicación de los puntos se concentra claramente hacia el vértice en el cual se encuentra JMS, esto se debe a que la cantidad de votos que recibieron estos candidatos no resulta significativa.
- Por el contrario en los casos en los cuales OIZ si está incluido, el comportamiento de los puntos se distribuye entre los vértices que contienen respectivamente a OIZ y JMS, debido a la poca diferencia que hubo entre estos dos candidatos, los cuales se sabe fueron los que mayor número de votos tuvieron.

Los resultados anteriores son acordes con la realidad política que se vivió en estas elecciones, de las cuales se sabe que los candidatos que tuvieron mayor número de votos a favor fueron Juan Manuel Santos y Oscar Iván Zuluaga.

Los gráficos para la composición de la segunda vuelta y para el plebiscito se muestran a continuación:




(a) Diagrama ternario para la segunda vuelta presidencial.




(b) Diagrama ternario para el plebiscito.

Figura 3: Diagramas ternarios correspondientes a la segunda vuelta presidencial y al plebiscito.

- El panel de la izquierda el cual hace referencia a la segunda vuelta lo que se puede evidenciar es que los dos vertices que representan al candidato Juan Manuel Santos y Oscar Iván Zuluaga son los que concentran la mayor parte de la información, es decir, que como bien se sabe del contexto de estas elecciones, aunque el candidato ganador fue JMS, los votos estuvieron bastantes divididos.
- El panel de la derecha el cual hace referencia al plebiscito pone en evidencia que la mayor parte de la información se encuentra ubicada hacia el vértice que representa la abstención en el proceso electoral, lo cual es coherente con el hecho de que el porcentaje de abstención que se registró en esa jornada de votaciones fue de más del 60 %.

## 5.2. Modelo para los datos

En el espacio del simplex la ecuación del modelo se plantea así:

$$Y_i = a \oplus X_{1i} \odot B_1 \oplus X_{2i} \odot B_2 \oplus x_{3i} \odot b_3 \oplus x_{4i} \odot b_4 \oplus x_{5i} \odot b_5 \oplus \varepsilon_i \tag{20}$$

- En donde **a** y **b**: Son constantes composicionales.
- **Y** es una composición como variable respuesta.
- **x** es una composición como variable explicativa.
- **X** es una variable real explicativa.
- **B** es una matriz cuadrada que representa una transformación lineal entre composiciones en el espacio ilr, como cuando en el espacio real se tienen matrices para representar asignaciones lineales de vectores.

Teniendo en cuenta la información que se tuvó disponible, el modelo que se ajustó considera en su planteamiento lo siguiente:

| Descripción de la variable | Tipo | Composicional |
|---|---|---|
| Composición referente al plebiscito. (Si, No, Abstención) | Respuesta | Si |
| Composición referente a la primera vuelta de elecciones presidenciales. (Juan Manuel Santos, Oscar Ivan Zuluaga, Clara Lopez, Martha Lucía Ramírez, Enrique Peñalosa) | Explicativa | Si |
| Composición referente a la segunda vuelta de elecciones presidenciales. (Juan Manuel Santos, Oscar Ivan Zuluaga, Otros votos) | Explicativa | Si |
| Proyección de población al 2014 | Explicativa | No |
| Potencial de votantes del año 2014 | Explicativa | No |
| Índice de necesidades básicas insatisfechas al 2008 | Expliccativa | No |

Tabla 4: Explicación de variables

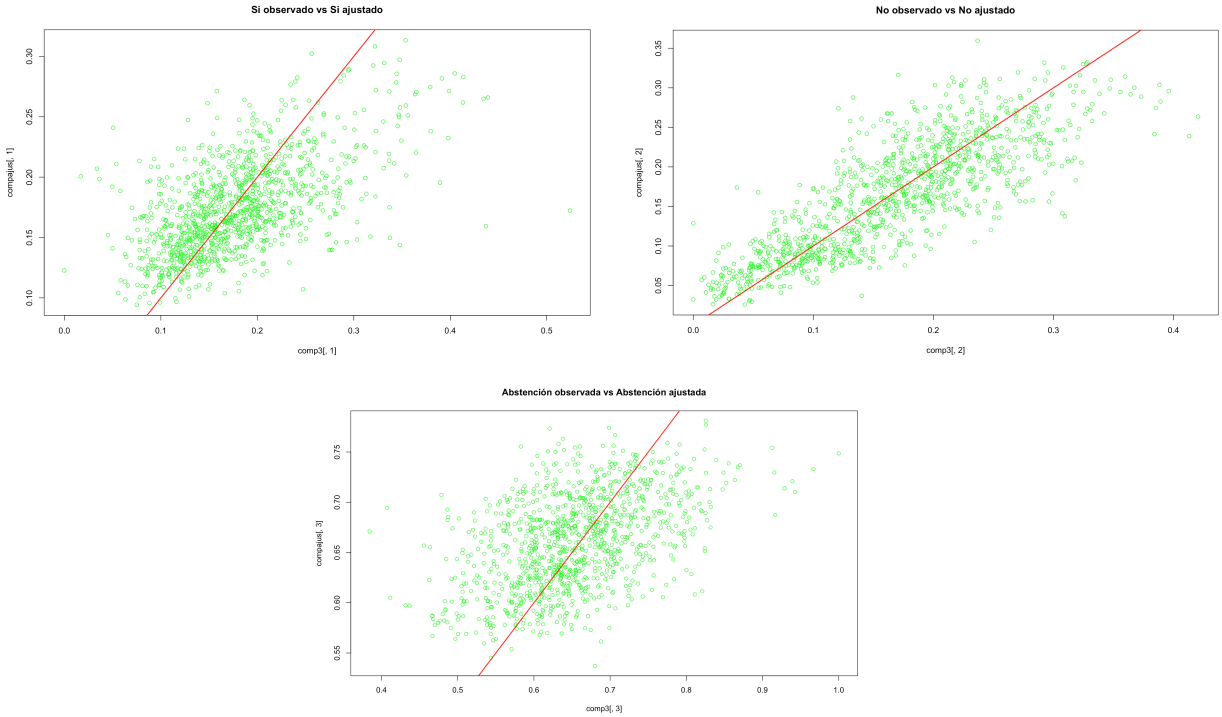
Al aplicar la transformación ilr sobre el modelo 20 se obtiene la siguiente ecuación, propia de un modelo especificado en el espacio $\mathbb{R}^2$:

$$ilr(Y_i) = ilr(a) + ilr(X_{1i})B_1 + ilr(X_{2i})B_2 + x_{3i}b_3 + x_{4i}b_4 + x_{5i}b_5 + \varepsilon_i \tag{21}$$

- $X_{1i}$: Composición referente a la primera vuelta presidencial.
- $X_{2i}$: Composición referente a la segunda vuelta presidencial.
- $Y_i$: Composición referente al plebiscito.
- $x_{3i}$: Proyección de población al 2014.
- $x_{4i}$: Potencial electoral del año 2014.
- $x_{5i}$ Índice de necesidades básicas insatisfechas al 2008.

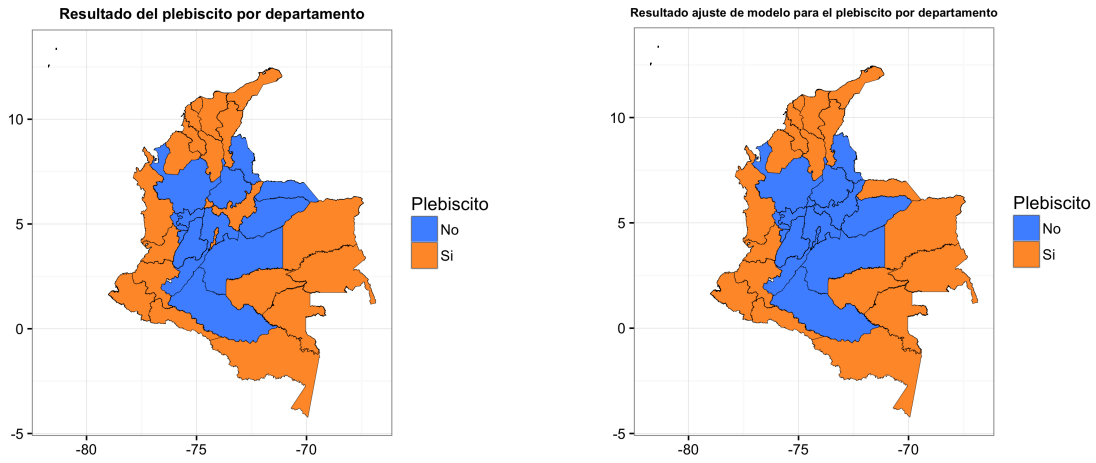
Para la aplicación del modelo, a los elementos que sean de tipo composicional se les incluye dentro de la notación del modelo, solo que agregándo el comando ilr(x), ya que de esa manera es posible ajustar un modelo lineal convencional a cada coordenada ilr.

De cualquier manera, un aspecto que hay que tener en cuenta para efectos de análisis es que debido a las transformaciones que se le realizan a los datos, al final resulta difícil interpretar los coeficientes del modelo como suele hacerse con los métodos clásicos. Teniendo en cuenta la dificultad que se tiene para interpretar los coeficientes del modelo, 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.




Validación ajuste del modelo vs Datos observados

Teniendo en cuenta que la linea roja presente en cada gráfico ayuda a definir que tan bueno está siendo el ajuste del modelo a los datos, se puede decir entonces que al menos en términos visuales a grandes rasgos el modelo presenta un buen ajuste; para verificar esto y debido a lo mencionado en cuanto a la complejidad de la interpretación de los coeficientes del modelo se puede ver a continuación una comparación geográfica entre dos mapas de Colombia a nivel departamental, el primero de ellos muestra los resultados originales del plebiscito y el segundo los resultados que se logran mediante la aplicación del modelo.




(a) Mapa departamental con los resultados originales. (b) Mapa departamental con lo ajustado por el modelo.

Figura 5: Comparación geográfica a nivel departamental: Resultado real del plebiscito vs Ajuste del modelo.

En virtud de este resultado es posible reiterar lo ya mencionado anteriormente, a partir de la interpretación de los gráficos de ajuste del modelo, en el sentido que como se aprecia en los mapas, con el modelo se logra predecir el resultado ganador en cada uno de los departamentos exceptuando a Bogotá, Boyacá y Aráuca, entidades para los cuales el modelo indica que en Bogotá y Boyacá gana el NO cuando en realidad ganó el SI y para el caso de Aráuca en que establece que gana el SI cuando en realidad ganó el NO; lo anterior puede estar explicado a que estos municipios tuvieron un cambio de "parecer" en cuanto al perfil electoral al que apoyan.

Adicionalmente es posible plantear una conclusión de carácter general en el sentido que, con el ejercicio realizado a partir de las variables utilizadas es posible afirmar que los resultados reales del proceso electoral a nivel nacional coinciden con los que se logran establecer mediante la aplicación del modelo como se muestra a continuación:

| | RESULTADOS A NIVEL NACIONAL | |
|---|---|---|
| DATOS | SI | NO |
| DATOS REALES | 49,76% | 50,24% |
| MODELO | 49,59% | 50,41% |

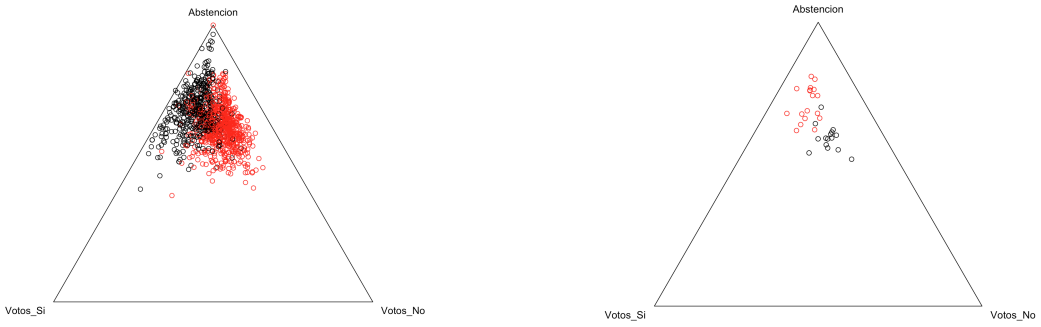
Tabla 5: Comparación porcentajes reales y estimados por el modelo a nivel nacional para el plebiscito

Como conclusión del ejercicio es posible señalar que el modelo utilizado capta y logra reproducir de forma coherente lo que fueron los resultados del plebiscito por la paz, lo cual permite afirmar que las variables que se usaron fueron adecuadas ya que están inmersas en la situación social y política del país en ese momento.

## 5.3. Análisis de cluster

Para realizar el análisis de cluster en este tipo de datos, al igual que al momento de plantear un modelo se debe usar una tranformación para poder hacer uso de las metodologías clásicas, las variables tenidas en cuenta para el desarrollo de este ejercicio fueron unicamente la primera y segunda vuelta referentes a las elecciones presidenciales , el resultado gráfico se muestra a continuación:

**Cluster a nivel municipal y departamental**



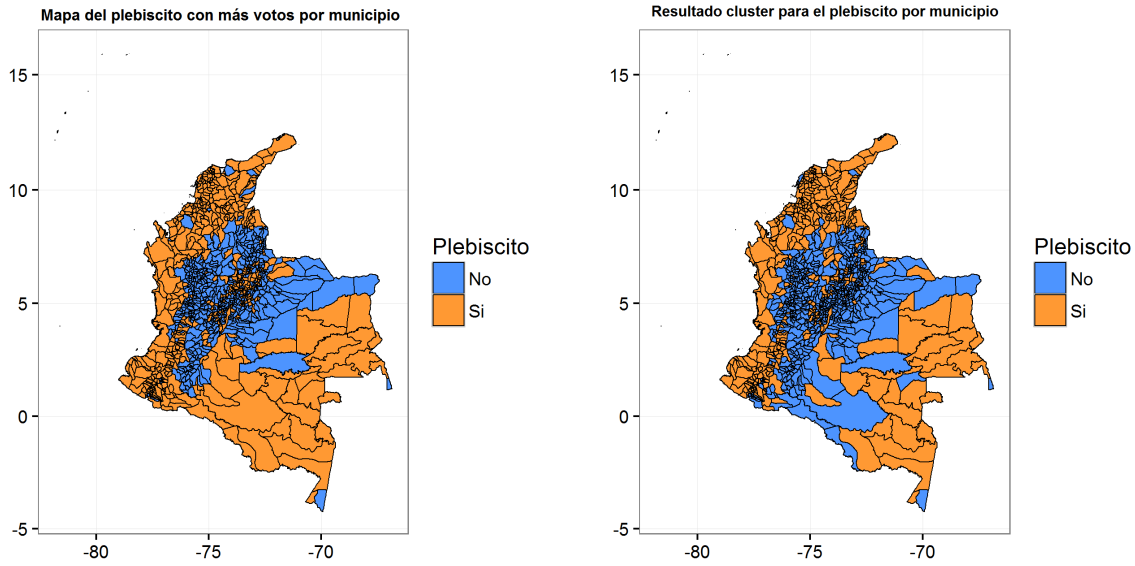

(a) Cluster a nivel municipal.

(b) Cluster a nivel departamental.

Figura 6: Diagramas ternarios que permiten visualizar el ajuste del cluster tanto a nivel municipal como departamental.

- Del resultado anterior se aprecia que los dos grupos creados en por el cluster, pueden ser asociados a los municipios que apoyaron el SI y el NO respectivamente. Cabe aclarar que esta información asociada al plebiscito no fue utilizada para la división en los dos conglomerados, solamente para la construcción del gráfico. Adicionalmente se realizó el mismo ejercicio pero a nivel departamental.

- Como en el caso municipal se aprecia una clara división de dos grupos, los cuales hacen referencia a los que apoyaron el Si y el No respectivamente en el plebiscito; con el fin de tener un resultado que brinde una mayor interpretabilidad se muestra a continuación una comparación geográfica como la vista anteriormente para el caso del modelo.




(a) Mapa municipal con los resultados originales. (b) Mapa municipal con lo ajustado por el cluster.

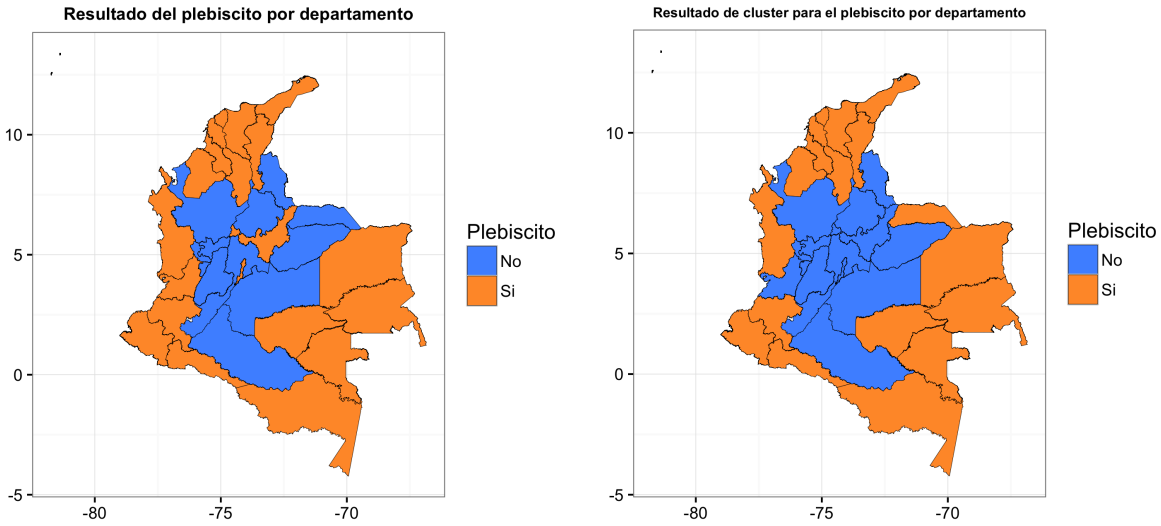
Figura 7: Comparación geográfica a nivel municipal: Resultado real del plebiscito vs Ajuste del cluster.

Aunque concluir algo del resultado municipal resulta un tanto difícil dado la cantidad de municipios al interior del país, se puede apreciar lo siguiente:

- En las dos mapas se hace una muy buena clasificación de los municipios en los cuales ganó el SI.

- En el interior del país se hace una buena clasficación de municipios donde ganó el NO.

- Aunque en regiones como la Orinoquía y Amazonía el análisis cluster clasifica como partidarios del NO a algunos municipios que no lo fueron el día del plebiscito, esta clasificación va más acorde a la situación real del país, ya que el NO fue el vencedor el día del plebiscito.




(a) Mapa departamental con los resultados originales. (b) Mapa departamental con lo ajustado por el cluster.

Figura 8: Comparación geográfica a nivel departamental: Resultado real del plebiscito vs Ajuste del cluster.

Este resultado del estudio visto a nivel departamental resulta ser más claro que cuando se trabaja a nivel municipal. En el caso departamental los cifras observadas son prácticamente las mismas obtenidas mediante el ajuste del modelo salvo con una diferencia en el departamento del Valle en el cual ahora el cluster indica que gana el NO cuando en realidad en este departamento ganó el SI.

Por último para efectos de dejar en claro la utilidad y eficacia de la aplicación de este método, se presenta a continuación una tabla de contingencia, tanto para el caso municipal como para el departamental, la cual permite ver más claramente cuál es el número de aciertos y errores, usando, en ambos casos, el método del cluster.

| | | CLUSTER DEPARTAMENTAL | |
|---|---|---|---|
| | | SI | NO |
| OBSERVADO | SI | 17 | 3 |
| | NO | 1 | 12 |

Porcentaje de aciertos: 88%

| | | CLUSTER MUNICIPAL | |
|---|---|---|---|
| | | SI | NO |
| OBSERVADO | SI | 406 | 171 |
| | NO | 37 | 508 |

Porcentaje de aciertos: 81%

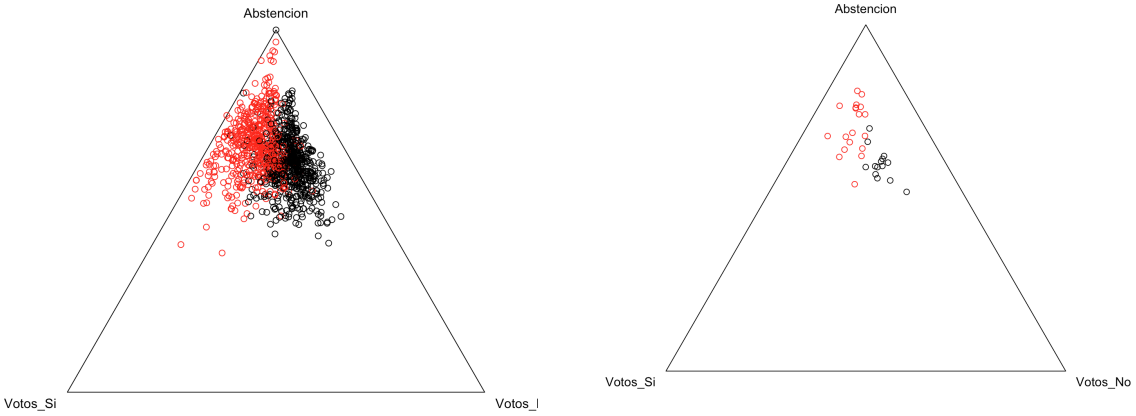
Tabla 6: Tabla de contingencia aciertos y errores con el método de cluster

Como conclusión de este ejercicio es posible afirmar que los dos ejercicios tanto a nivel departamental como a nivel municipal tienen resultados muy buenos ya que en ambos casos el porcentaje de aciertos se encuentra por encima del 80 %.

## 5.4. Análisis discriminante

Para realizar el análisis discriminante en este tipo de datos, al igual que con los métodos anteriores se debe usar la tranformación ilr para poder hacer uso de las metodologías clásicas.

Para el desarrollo de este ejercicio se hizo uso de las 6 variables disponibles, el resultado gráfico se muestra a continuación:




(a) Discriminante a nivel municipal

(b) Discriminante a nivel departamental

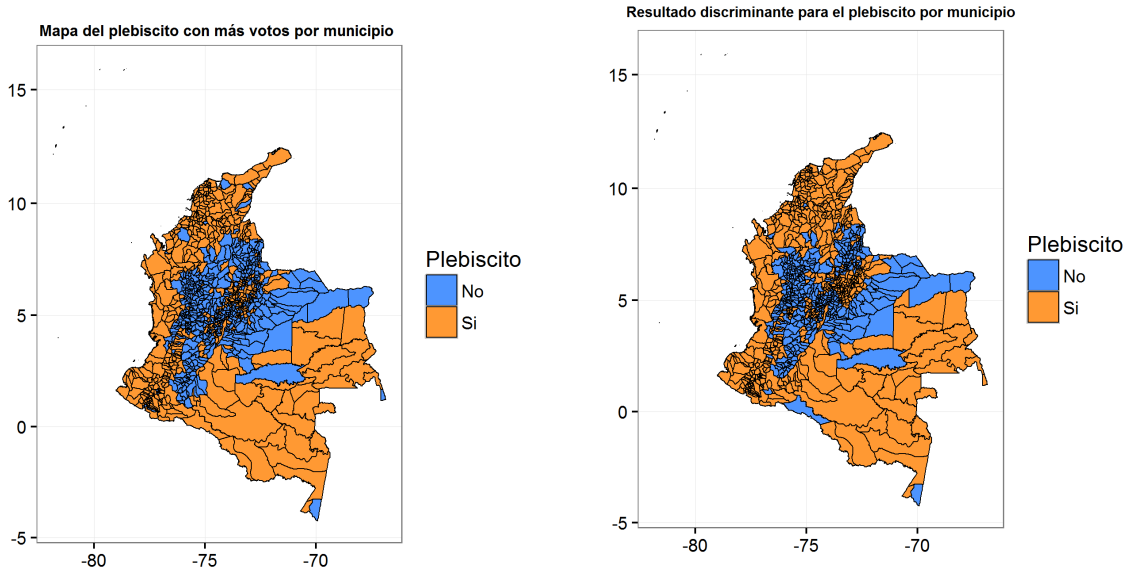
Figura 9: Diagramas ternarios que permiten visualizar el ajuste del discriminante tanto a nivel municipal como departamental.

- Del resultado anterior se aprecia que los dos grupos creados se asocian a los municipios y departamentos que apoyaron el SI y el NO respectivamente.

- Si se compara el resultado municipal obtenido por el cluster con el de este método, a simple vista se ve una mayor similitud de este último con las tendencias reales del plebiscito para cada uno de los municipios.

- Adicionalmente, es importante resaltar que la información asociada al plebiscito SI fue utilizada al momento de llevar a cabo este análisis; con el fin de tener un resultado que brinde una mayor interpretabilidad se muestra a continuación una comparación geográfica como la vista anteriormente para el caso del cluster.

Del procedimiento de análisis discriminante se obtienen los siguientes mapas tanto a nivel departamental como a nivel municipal:

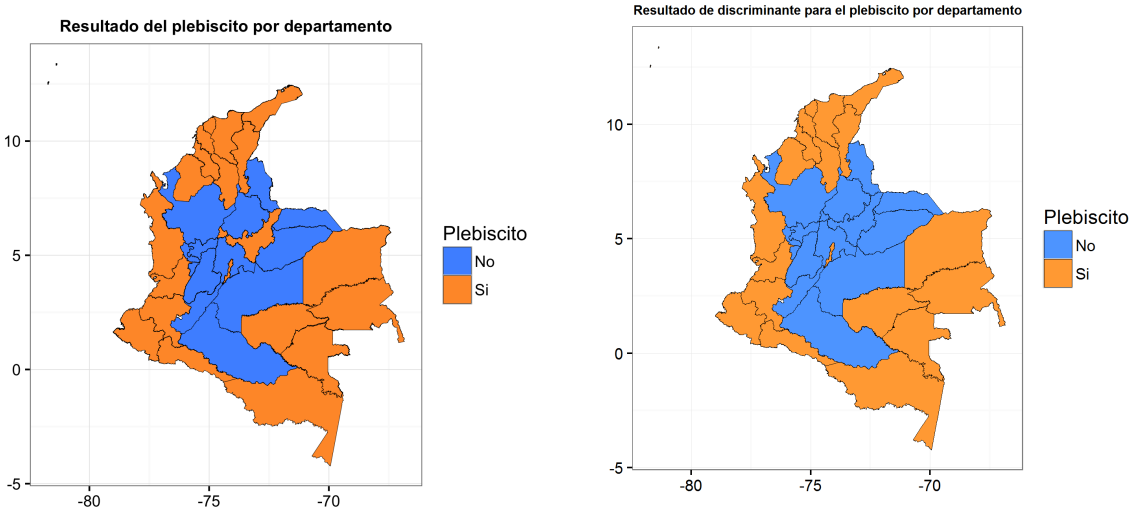



(a) Mapa municipal con los resultados originales.

(b) Mapa municipal con lo ajustado por el análisis discriminante.

Figura 10: Comparación geográfica a nivel municipal: Resultado real del plebiscito vs Ajuste del análisis discriminante.

Aunque obtener conclusiones en el caso de los resultados municipales, resulta un tanto difícil, dado la cantidad de municipios en que está dividido el país, es incuestionable señalar que, a grandes rasgos, al igual que el cluster, este método, tiende a establecer que hay más municipios que apoyaron el NO de los que en realidad lo hicieron. Para un análisis gráfico un poco más claro y entendible se puede ver a continuación la comparación de los mapas pero en este caso a nivel departamental:




(a) Mapa departamental con los resultados originales.

(b) Mapa departamental con lo ajustado por el análisis discriminante.

Figura 11: Comparación geográfica a nivel departamental: Resultado real del plebiscito vs Ajuste del análisis discriminante.

Este resultado del estudio visto a nivel departamental resulta ser más claro que cuando se trabaja a nivel municipal. En el caso departamental los cifras observadas son prácticamente iguales los resultados reales del ejercicio electoral, el único departamento que este método no es capaz de predecir es Boyacá en el cual indica que gana el NO cuando en realidad en este departamento ganó el SI.

Por último se muestra a continuación una tabla de contingencia tanto a nivel municipal como departamental, la cual permite ver un poco más claramente cual es el número de aciertos y errores en ambos casos usanto el método del análisis discriminante.

| | | DISCRIMINANTE DEPARTAMENTAL | |
|---|---|---|---|
| | | SI | NO |
| OBSERVADO | SI | 19 | 1 |
| | NO | 0 | 13 |

Porcentaje de aciertos: 96,97%

| | | DISCRIMINANTE MUNICIPAL | |
|---|---|---|---|
| | | SI | NO |
| OBSERVADO | SI | 466 | 111 |
| | NO | 48 | 497 |

Porcentaje de aciertos: 85,83%

Tabla 7: Tabla de contingencia aciertos y errores con el método discriminante

Como conclusión de este ejercicio es posible afirmar que al igual que en el caso del cluster los dos ejercicios tanto a nivel departamental como a nivel municipal tienen resultados muy buenos, ya que en ambos casos el porcentaje de aciertos se encuentra por encima del 80 %, y aunque con este método el resultado a nivel departamental resulta mucho mejor, no se debe olvidar que este ejercicio a diferencia del cluster si hace uso de los resultados observados en el plebiscito.

# 6. Discusión

## 6.1. Conclusiones

Como aporte de los resultados del ejercicio de análisis composicional en el campo de variables electores, a continuación se presentan un conjunto de conclusiones, limitaciones y propuestas de trabajo que pueden realizarse sobre esta temática. Se trata de tres aspectos importantes en el campo de la estadística y que son en todo caso el fruto del ejercicio de aplicaciones metodologías y análisis que se ha desarrollado.

1. En primera instancia cabe señalar que como resultado del trabajo se pone en evidencia lo importantes y relevantes que son las técnicas de análisis DaCo y de lo útiles que pueden llegar a ser si se aplican de manera adecuada. De esa manera queda claro que en esta área de la estadística se requieren más estudios y profundización de su análisis con el fin de lograr comprenderla en su totalidad y con ello hacer aportes importantes en diversas áreas de aplicación de estas metodologías.

2. Los resultados de los tres ejercicios aplicados en este caso (el modelo, el cluster y el análisis discriminante), permiten resaltar las grandes bondades y capacidades de las técnicas para DaCo, ya que si se comparan las predicciones que en su momento aportaron distintas firmas encuestadoras en periodos previos al ejercicio electoral con los resultados obtenidos en este trabajo se sabe que estos últimos hubiesen dado resultados más acordes con lo que se observó en la realidad.

3. Aunque el objetivo de este ejercicio no era hallar el mejor ajuste a los resultados del proceso electoral, cabe señalar que, con las variables que se utilizaron como base para el análisis, se logra una una buena aproximación a los datos reales; lo cual motiva a que se realicen con mayor profundidad estudios más amplios y completos a partir de los cuales se pueda llegar a encontrar y dar uso a variables que ayuden a mejorar los resultados obtenidos.

4. De otro lado es importante indicar que los resultados del ejercicio desarrollado, en esta oportunidad, tanto en lo que tiene que ver con los aspectos teóricos como a nivel práctico responden adecuadamente a los objetivos y lineamientos propuestos en este trabajo.

## 6.2. Limitaciones

1. Como con cualquier proyecto de investigación se tienen limitaciones, y para este ejercicio en particular cabe señalar el hecho de que al momento de ajustar un modelo a los datos, los coeficientes del mismo pierden su interpretabilidad debido a la serie de transformaciones que se le hacen a los datos, lo anterior lleva a que la interpretación de los mismos no sea recomendable debido a que se puede incurrir en conclusiones erróneas.

2. Una segunda limitación tiene que ver con el hecho de que, cuando se implementa el cluster o el análisis discriminante, el resultado indica únicamente cual hubiese sido el resultado ganador para cada uno de los entes territoriales considerados, en este caso, municipios o departamentos, lo cual imposibilita conocer el porcentaje o el total de votos que hubiese tenido cada uno de los posibles resultados; lo anterior es de vital importancia en la medida que al no disponer de este dato, no se le está dando la importancia que debe tener la densidad poblacional de cada región, la cual finalmente es la que define el resultado ganador a nivel nacional.

## 6.3. Trabajos futuros

1. A partir de las conclusiones que se expusieron y las limitaciones que subyacen en el trabajo, una primera propuesta sería la necesidad de ahondar en la búsqueda de otras variables que puedan aportar elementos para mejorar los resultados que se obtuvieron con este ejercicio. Cabria considerar

por ejemplo, variables de tipo demográfico que describan la composición de la población, para cada uno de los entes territoriales considerados. (municipios − departamentos)

2. Cabría sin duda la necesidad de implementar otras técnicas que ya han sido desarrolladas para la aplicación en DaCo como bien podría ser la geoestadística; la implementación de metodologías adicionales permitiría hacer aún más comparaciones entre las diferentes metodologías.

3. Podrían además llevarse a cabo ejercicios en este caso referidos a los procesos electorales al interior de las grandes ciudades como Bogotá y Medellín usando la información referente a las mismas variables empleadas en este proyecto pero, a nivel de las localidades y de las comunas, cuyos comportamientos electorales pueden ser muy diferentes dado las diferencias socioeconómicas en los territorios de una y otra ciudad.

4. Cuando se genera un nuevo proceso electoral, es conveniente enriquecer los pronósticos que hacen las firmas encuestadoras con información auxiliar, para este fin es necesario desarrollar procedimientos de estimación en áreas pequeñas para datos composicionales.

# Referencias

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, The Blackburn Press.

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

Pawlowsky, V. & Buccianti, A. (2011), *Compositional Data Anlalysis. Theory and Applications*, Wiley.

van den Boogaart, K. & Delgado, R. (2013), *Analyzing Compositional Data with R*, Springer.

van den Boogaart, K. G., Tolosana, R. & Bren, M. (2014), *compositions: Compositional Data Analysis*. R package version 1.40-1.
*https://CRAN.R-project.org/package=compositions

# 7. Anexo: Código en R

```
library(car)
library(graphics)
library(compositions)
library(xlsx)

comp1 <- acomp(datos[5:9]) # Composicion de la primera vuelta - sin otros votos M1
comp2 <- acomp(datos[18:20]) # Composicion de la segunda vuelta - con otros votos M2
comp3 <- acomp(datos[15:17]) # Composicion del plebiscito - si, no y abstenci\'on.

head(comp1)
head(comp2)
head(comp3)

###################### GRAFICOS DESCRIPTIVOS ##############################
###########################################################################

plot(comp1, cex=0.3,col="blue",plotMissings=FALSE,main="Diagrama ternario para los candidatos de la
vuelta contra la media geomÃ©trica")
plot(comp1, cex=0.3,col="green",plotMissings=FALSE,margin="JMS",main="Diagrama ternario para los can
de la primera vuelta contra Juan Manuel Santos   ")
plot(comp2, cex=0.3,col="green",plotMissings=FALSE,main="Diagrama ternario para los candidatos
de la segunda vuelta")
plot(comp3,cex=0.3,col="green",plotMissings=FALSE,main="Diagrama ternario para el plebiscito")

############################### MODELO ####################################
###########################################################################

primvuel <- acomp(datos[5:9]) # A dataset with usefull regressor#primera vuelta
segvuel <- acomp(datos[18:20])#segunda vuelta
proy.pob <- datos$Proyeccion # The (here categorial) regressor
nbi <- datos$NBI

pot.vot <- datos$Potencial_Votantes
pleb <- acomp(datos[15:17])



mylm <- lm(ilr(pleb)~ilr(primvuel)+ilr(segvuel)+proy.pob+nbi+pot.vot)

compajus <- ilrInv(mylm$fitted.values, orig = pleb)

#GRAFICOS AJUSTE DEL MODELO

plot(comp3[,1],compajus[,1],main="Si observado vs Si ajustado",col="green")
abline(a=0,b=1,col=2,lwd=2)

plot(comp3[,2],compajus[,2],main="No observado vs No ajustado",col="green")
abline(a=0,b=1,col=2,lwd=2)

plot(comp3[,3],compajus[,3],main="Abstenci\'on observada vs Abstenci\'on ajustada",col="green")
```

```
abline(a=0,b=1,col=2,lwd=2)

############################### CLUSTER ###################################
###########################################################################

#MUNICIPAL

clus4 <- hclust(dist(comp1,method="euclidean"))
means4 <- acomp(t(sapply(split(comp1,factor(cutree(clus4,2))),mean)))

basex <- data.frame(ilr(comp1),ilr(comp2))

km5 <- kmeans(basex,centers=2)
plot(comp3,col=km5$cluster,plotMissings=FALSE)

#DEPARTAMENTAL

comp1dep <- acomp(cluster[2:6])
comp2dep <- acomp(cluster[13:15])
comp3dep <- acomp(cluster[10:12])
clusf <- hclust(dist(comp3dep,method="euclidean"))
meansf <- acomp(t(sapply(split(comp3dep,factor(cutree(clusf,2))),mean)))


baseclus <- data.frame(ilr(comp1dep),ilr(comp2dep))
kmf <- kmeans(baseclus,centers=2)
plot(comp3dep,col=kmf$cluster)

############################ DISCRIMINANTE ################################
###########################################################################


#DEPARTAMENTAL
comp1dep <- acomp(cluster[2:6])
comp2dep <- acomp(cluster[13:15])
comp3dep <- acomp(cluster[10:12])
base3x <- data.frame(comp1dep,comp2dep,cluster$Potencial_Votantes,cluster$Proyeccion,cluster$NBI)
SI <- cluster$GANOSI

res3 = lda( x=data.frame(ilr(base3x)), grouping=SI)# a nivel departamental

disDEP<- predict(res3)$class

V=ilrBase(base3x)

plot(comp3dep,col=disDEP,main="Diagrama ternario para el plebiscito")

#MUNICIPAL

base4x <- data.frame(comp1,comp2,datos$Potencial_Votantes,datos$Proyeccion,datos$NBI)
SI1 <- datos$GANOSI
```

```
res4 = lda( x=data.frame(ilr(base4x)), grouping=SI1)# a nivel municipal
#res4

disMUN<- predict(res4)$class

V1=ilrBase(base4x)

plot(comp3,col=disMUN,main="Diagrama ternario para el plebiscito", plotMissings=FALSE)
```