Modelación de datos composicionales vía mezclas de distribuciones normales multivariadas

Presenta:
Arnoldo Daniel Miranda Fournier

Asesor:
Gabriel Nuñez Antonio

Maestría en Ciencias Matemáticas Aplicadas e Industriales
UAM-Iztapalapa

15 de Julio de 2020

Contenido

Objetivo General

El objetivo de esta tesis es proponer una metodología para contribuir al análisis, descripción y modelación de datos composicionales desde un enfoque bayesiano.

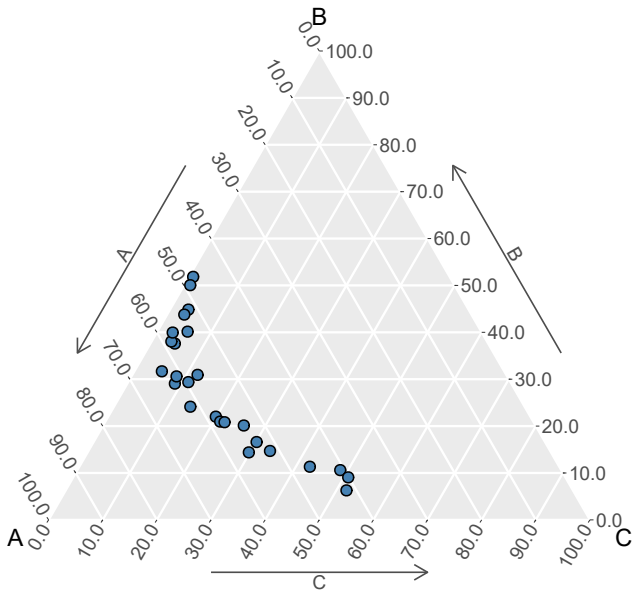
Introducción

¿Qué son los Datos Composicionales?

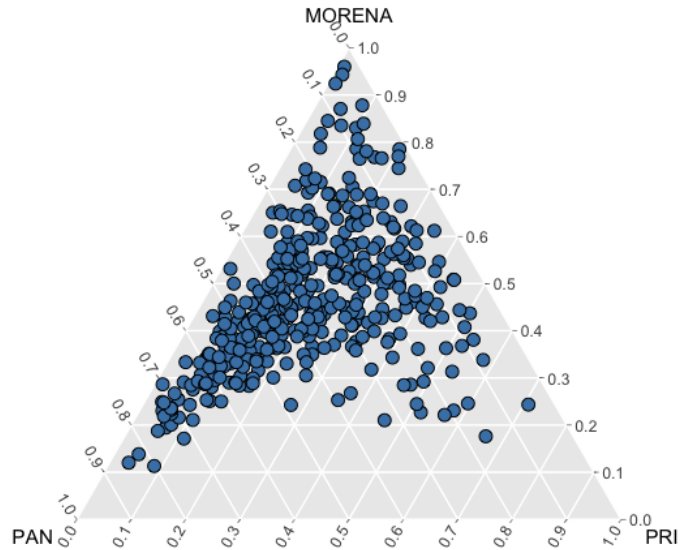
- Composición Geoquímica de Rocas
- Composición Geoquímica de Sedimentos
- Composición Proteica de la Leche
- Procesos Electorales

Ejemplos

- Composición Geoquímica de Rocas



- Datos Electorales (Puebla 2019)



- Geología, Política, Economía, Medicina, etc.

Su naturaleza y problemática

El símplex como espacio muestral

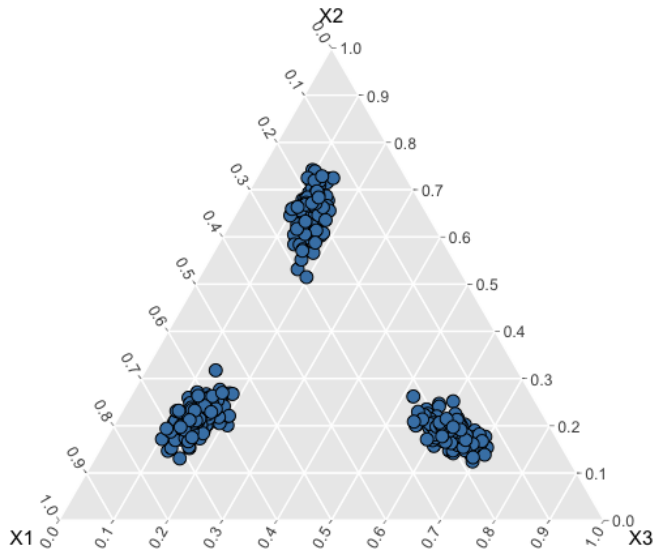
$$\mathbb{S}^p = \{\boldsymbol{x} \in \mathbb{R}^p : x_i \geq 0, \sum_{i=1}^{p} x_i = 1\}$$

$\mathbb{S}^2 \subset \mathbb{R}^2$	$\mathbb{S}^3 \subset \mathbb{R}^3$	$\mathbb{S}^4 \subset \mathbb{R}^4$
Segmento	Diagrama Ternario	Tetraedro
		

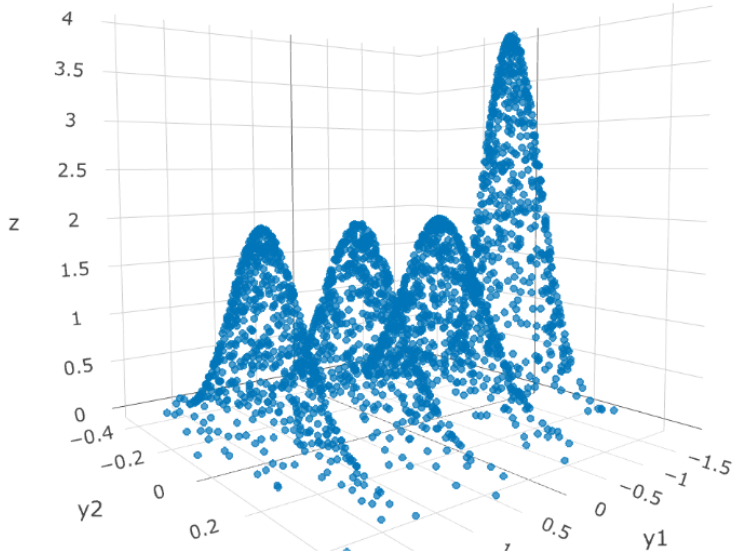
- Aitchison (1982) estudia las características y operaciones de los datos composicionales.

Objetivo específico

- Patrones más complejos



- Mezcla de Distribuciones Normales Multivariadas



- Mediante la conocida transformación log-cociente ilr, definida por Egozcue et al. (2003).

Modelos de Mezclas

- Los modelos de mezclas ofrecen una gran versatilidad en la descripción de datos, son ampliamente utilizados en Estadística.
- Su flexibilidad permite aproximar problemas desde un enfoque no paramétrico.
- Algunas aplicaciones de estos modelos es la estimación de la densidad y el análisis de conglomerados.
- Lo (1984) es el primero en contextualizar de manera general los modelos de mezclas.
- Escobar and West (1995) propusieron la primer implementación computacional factible.

Un modelo no-paramétrico

Sea $\boldsymbol{y} \in \mathbb{S}^{p+1}$. Entonces mediante la transformación *ilr*:

$$ilr(\boldsymbol{y}) = \boldsymbol{x}, \qquad \boldsymbol{x} \in \mathbb{R}^p$$

El modelo no-parámetrico propuesto para modelar a $\boldsymbol{x}$ es:

$$\boldsymbol{x} \sim f(\boldsymbol{x}|\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\lambda})$$

Donde

$$f\left(\boldsymbol{x} \mid \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\lambda}\right) = \sum_{i=1}^{\infty} \omega_i N_p\left(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\lambda}_i\right) \quad \text{donde} \quad \sum_{i=1}^{\infty} \omega_i = 1, \tag{1}$$

Con N_p una densidad *Normal Multivariada* con media $\boldsymbol{\mu}_i \in \mathbb{R}^p$ y matriz de precisión $\boldsymbol{\lambda}_i$ simétrica y positiva definida, Bernardo and Smith (2009):

$$N_p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) = |\boldsymbol{\lambda}|^{\frac{1}{2}} \left(2\pi\right)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)^t \boldsymbol{\lambda}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right\}$$

Los parámetros del modelo son las colecciones infinitas de $\{\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\lambda}\}$.

El enfoque Bayesiano de la Estadística

- $X \sim f(x \mid \theta)$: Modelo de Probabilidad.
- $\pi(\theta)$: Distribución Inicial.
- Muestra aleatoria $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$ del modelo $f(x \mid \theta)$
  Verosimilitud: $f(\boldsymbol{x} \mid \theta) = f(x_1, x_2, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$
- Teorema de Bayes...Distribución Final:

$$\begin{aligned}
\pi(\theta \mid \boldsymbol{x}) &= \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{f(\boldsymbol{x})} \\
&= \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int_{\Theta} f(\boldsymbol{x} \mid \theta)\pi(\theta)\mathrm{d}\theta} \\
&= \frac{\prod_{i=1}^{n} f(x_i \mid \theta)\pi(\theta)}{\int_{\Theta} \prod_{i=1}^{n} f(x_i \mid \theta)\pi(\theta)\mathrm{d}\theta}
\end{aligned}$$

- El enfoque Bayesiano permite transformar la estructura compleja de un modelo de mezclas en un conjunto de estructuras simples.

Proceso Dirichlet

- $P \sim D(c, P_0)$
- Las f.d.p aleatorias generadas son discretas (c.s)
- Una aplicación particular es como la distribución inicial en modelos de mezclas infinitas.
- Representación Stick-Breaking:

$$P = \sum_{j=1}^{\infty} \omega_j \delta_{\boldsymbol{\phi}_j} \quad \text{donde} \quad \sum_{i=1}^{\infty} \omega_i = 1, \tag{2}$$

donde $\delta_{\boldsymbol{\phi}_j}$ denota la medida con una masa puntual de 1 en $\boldsymbol{\phi}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\lambda}_j\}$, los pesos ω_j y $\boldsymbol{\phi}_j$ son localizaciones aleatorias distribuidas de acuerdo a una medida de probabilidad P_0.

Modelo de Mezclas de Procesos Dirichlet (MDP)

- Se basa en la idea de construir funciones de distribución absolutamente continuas.
- Así, el modelo propuesto se puede escribir como:

$$f\left(\boldsymbol{x} \mid P\right) = \int N_p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\lambda}) \mathrm{d}P(\boldsymbol{\mu}, \boldsymbol{\lambda}) \tag{3}$$

- Entonces a partir de la ecuación (2) para P, el modelo MDP se puede representar como:

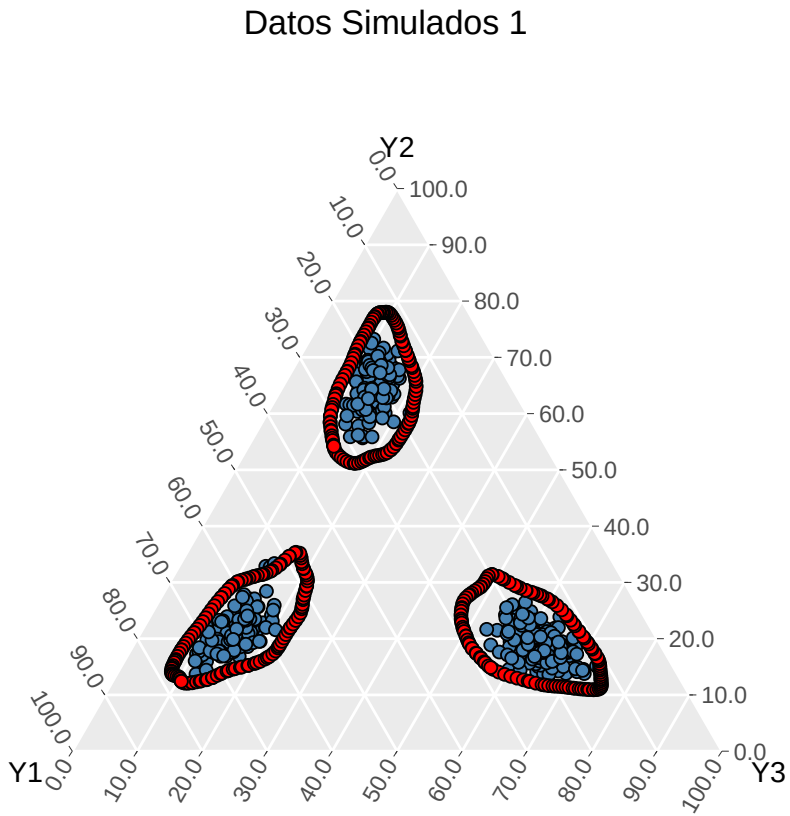
$$f\left(\boldsymbol{x} \mid \boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\lambda}\right) = \sum_{i=1}^{\infty} \omega_i N_p\left(\boldsymbol{x} \mid \boldsymbol{\mu}_i, \boldsymbol{\lambda}_i\right)$$

- Así, realizar inferencias sobre las colecciones infinitas de parámetros del modelo (1) se reduce a realizar inferencias en el modelo de MDP.

Muestreo del modelo de MDP

- Se utilizó el algoritmo dado por Walker (2007):
  - Gibbs Sampling.
  - Slice Sampling.
  - Variables latentes.
- Se complemento con la propuesta realizada por Kalli et al. (2011):
  - Slice Sampling más general.
- La combinación y adaptación de ambos algoritmos se implementaron en el lenguaje estadísitico $\mathbf{R}$.
- El algoritmo regresa una muestra de la distribución predictiva final en $\mathbb{R}^p$ que es mapeada de regreso al Símplex a través de la transformación ilr^{-1}.

Datos Simulados



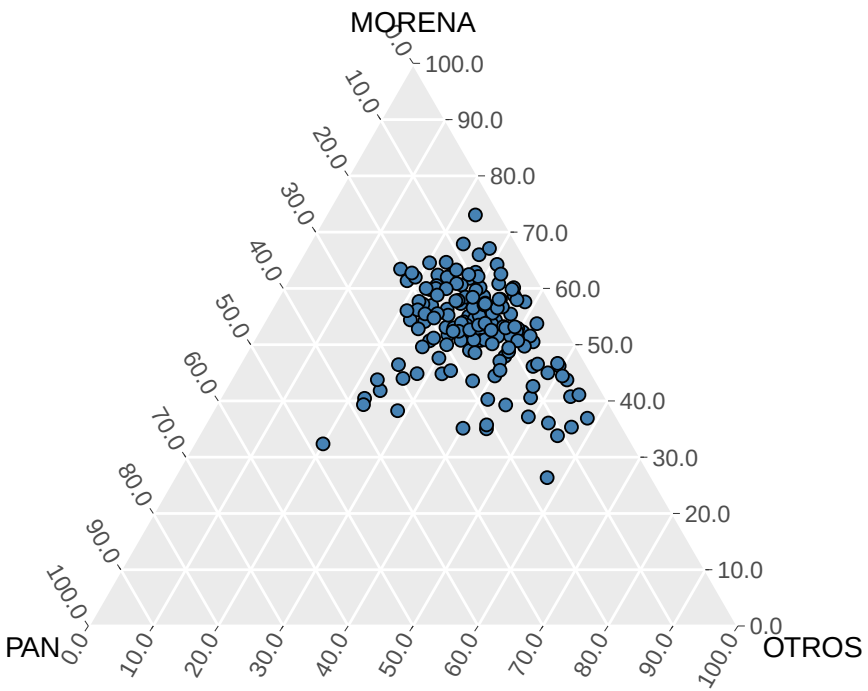
Intervalos Poblacionales		
Componente	Inferior	Superior
Y1	14.68 %	70.38 %
Y2	14.73 %	70.39 %
Y3	10.51 %	68.41 %

Intervalos de Probabilidad Marginales		
Componente	Inferior	Superior
Y1	13.42 %	72.70 %
Y2	13.71 %	72.45 %
Y3	09.34 %	69.74 %

- Se propuso establecer una especifiación inicial no informativa en el Símplex.

Datos Reales

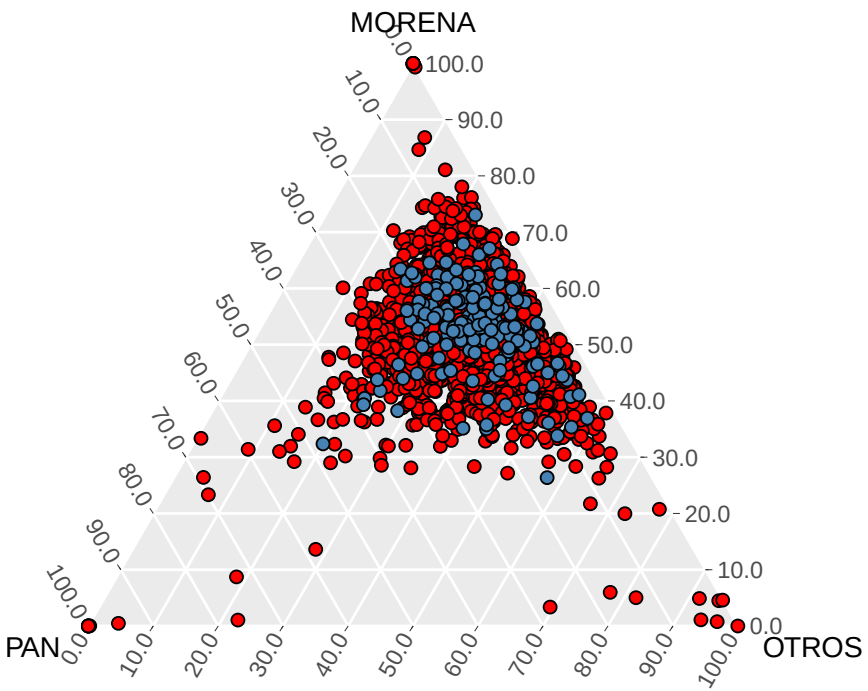
Gubernatura del estado de Morelos 2018



Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Datos Reales

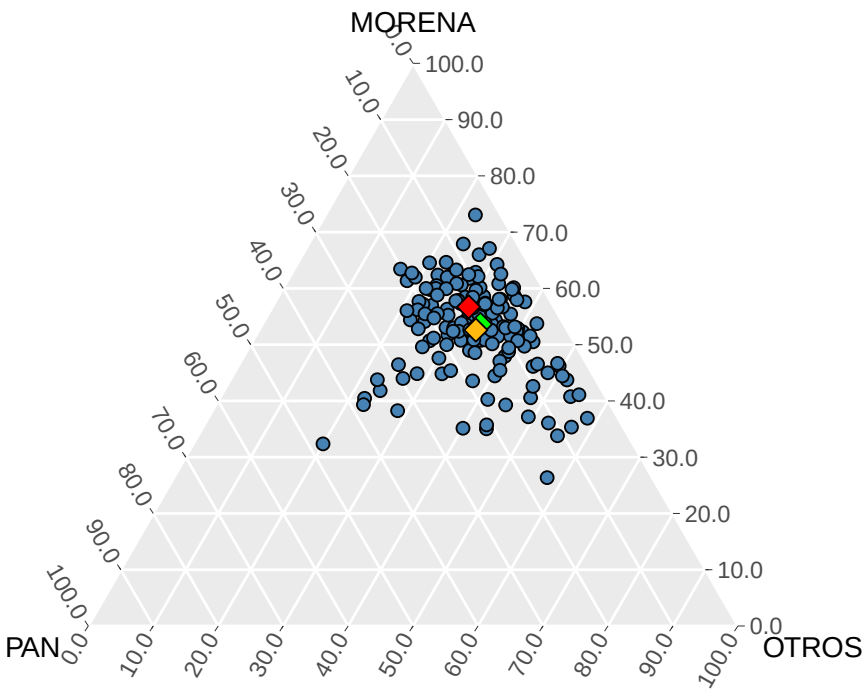
Gubernatura del estado de Morelos 2018



Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Datos Reales

Gubernatura del estado de Morelos 2018

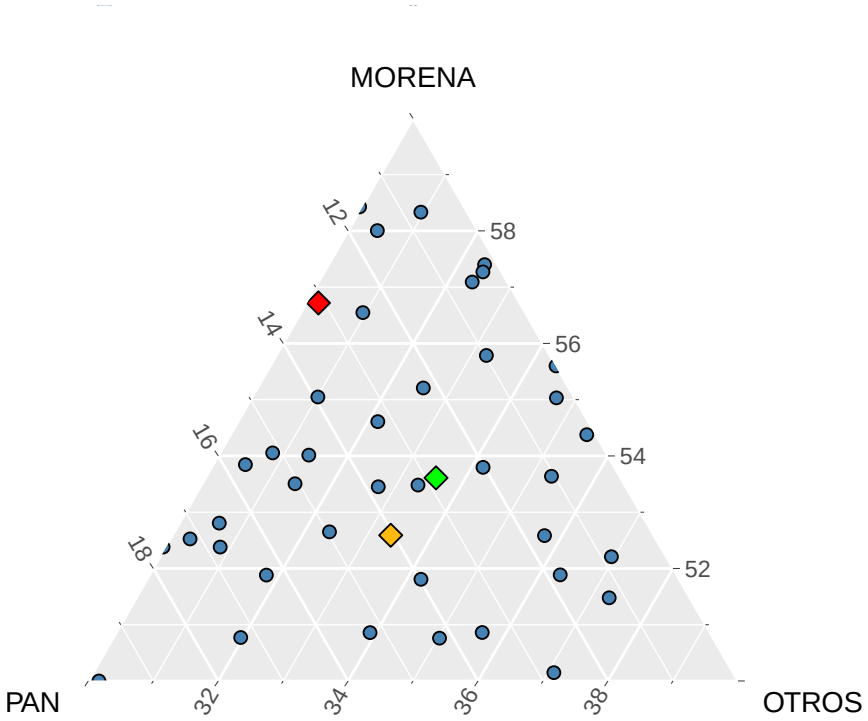


Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Estimadores Puntuales			
Estimador	Coalición		
	MORENA	PAN	OTROS
Media	56.72 %	13.10 %	30.18 %
Mediana	53.61 %	12.84 %	33.55 %

Datos Reales

Gubernatura del estado de Morelos 2018

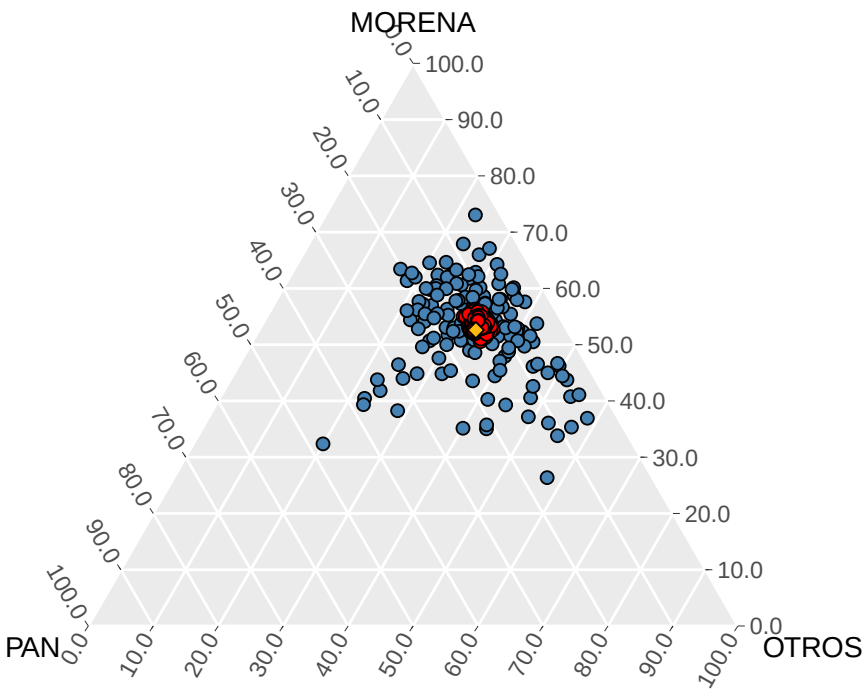


Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Estimadores Puntuales			
Estimador	Coalición		
	MORENA	PAN	OTROS
Media	56.72 %	13.10 %	30.18 %
Mediana	53.61 %	12.84 %	33.55 %

Datos Reales

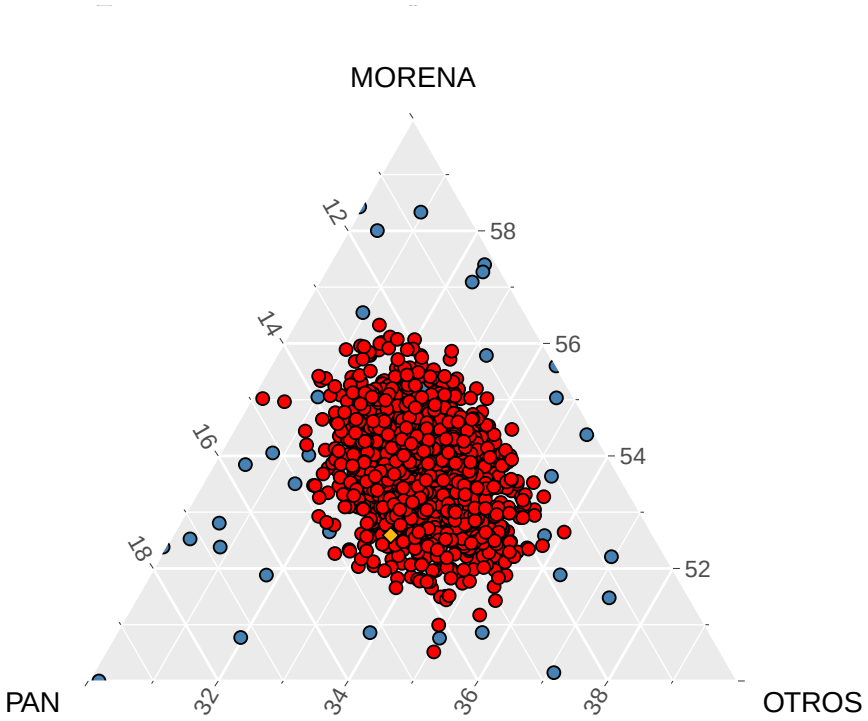
Gubernatura del estado de Morelos 2018



Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Datos Reales

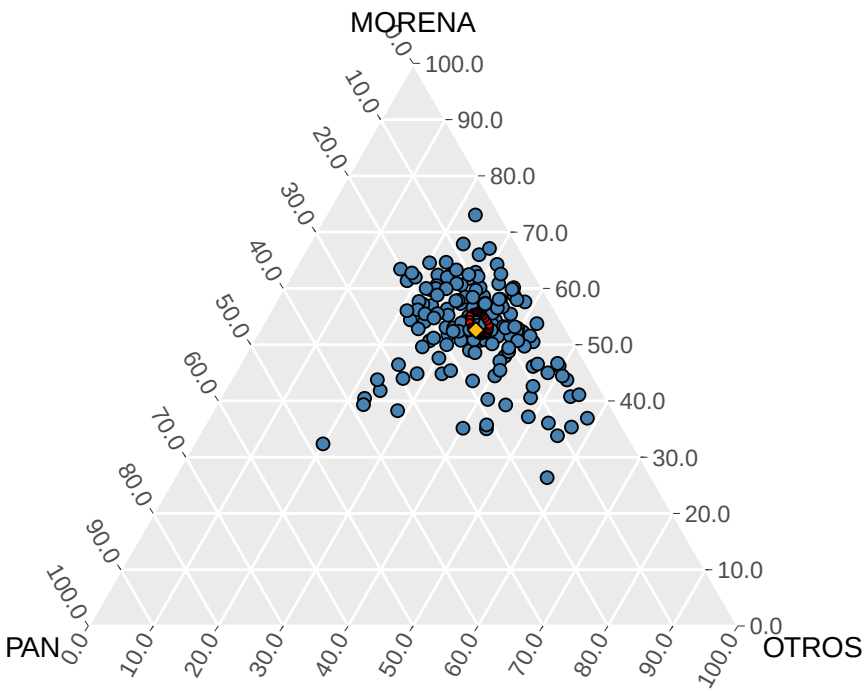
Gubernatura del estado de Morelos 2018



Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Datos Reales

Gubernatura del estado de Morelos 2018

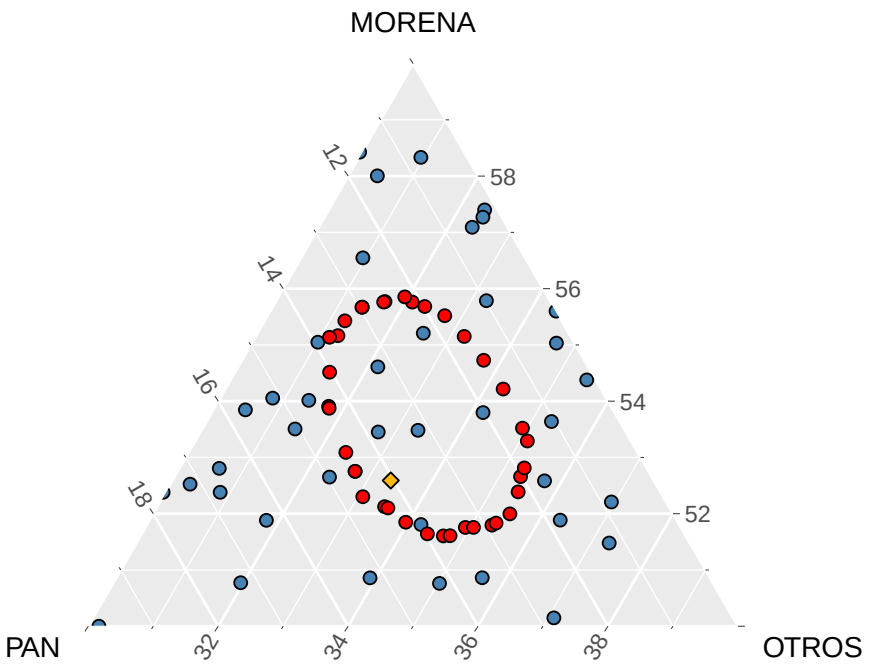


Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Intervalos de Probabilidad Marginales		
Coalición	Inferior	Superior
MORENA	52.20 %	55.31 %
PAN	11.80 %	14.22 %
OTROS	31.70 %	34.96 %

Datos Reales

Gubernatura del estado de Morelos 2018



Cómputo Distrital		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

Intervalos de Probabilidad Marginales		
Coalición	Inferior	Superior
MORENA	52.20 %	55.31 %
PAN	11.80 %	14.22 %
OTROS	31.70 %	34.96 %

Intervalos del INE		
Coalición	Inferior	Superior
MORENA	51.00 %	53.80 %
PAN	13.40 %	16.10 %

Temas abarcados

- Datos Composicionales
- Estadistica Bayesiana
- Modelos de mezclas
- Procesos Dirchlet
- Métodos de simulación
- Programación

Conclusiones y Trabajo por realizar

- Se propone un modelo que toma en cuenta la naturaleza de los datos (datos composicionales).
- Con esta metodología se esta en condiciones de ofrecer estimadores puntuales que cumplan con la restricción de que la suma de sus componentes sea uno.
- De igual manera se pueden ofrecer intervalos de probabilidad marginales, para cada componente involucrada en el vector composicional.
- La variabilidad de la proporción se mide al correr, cada vez el programa completo, obteniendo la mediana.

Conclusiones y Trabajo por realizar

- El modelo no-parámetrico (basado en mezclas infinitas de distribuciones normales multivariadas) fue viable gracias al uso de la transformación log-cociente ilr.
- Se podrían considerar modelos de mezclas con distribuciones definidas en ortantes positivos, con el propósito de omitir cualquier tipo de transformación log-cociente.
- El modelo puede mejorarse optimizando los algoritmos utilizados para la inferencia, por ejemplo, se podrían paralelizar algunos procesos.
- Aunque quedan cosas por hacer, esta propuesta de tesis sienta las bases para el análisis bayesiano de datos composicionales.

Referencias

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

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

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, 21(1):93–105.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357.

Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54.

¡Gracias!