Los Conteos Rápidos del INE y la Topología del Símplex Unitario p-dimensional

Presenta:
M. en C. Arnoldo Daniel Miranda Fournier

Asesor:
Dr. Gabriel Nuñez Antonio

Laboratorio de Ciencia de Datos
Una travesía por la Ciencia de Datos para matemáticos y no matemáticos

17 de Septiembre de 2020

Contenido

Objetivo General

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

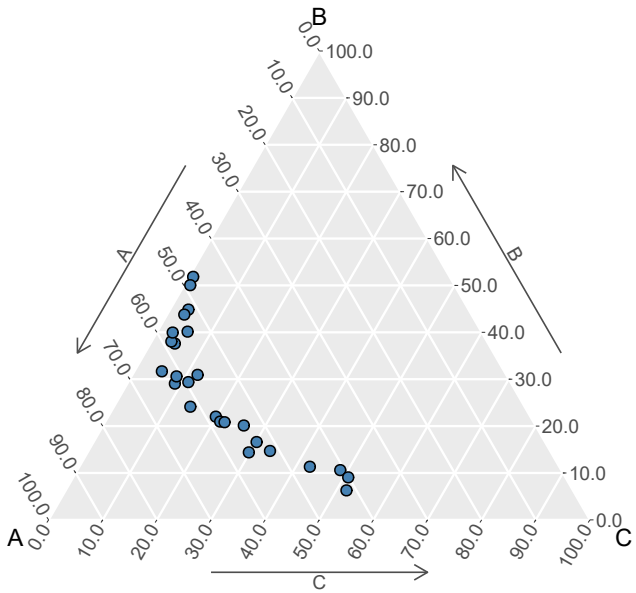
Introducción

¿Qué son los Datos Composicionales?

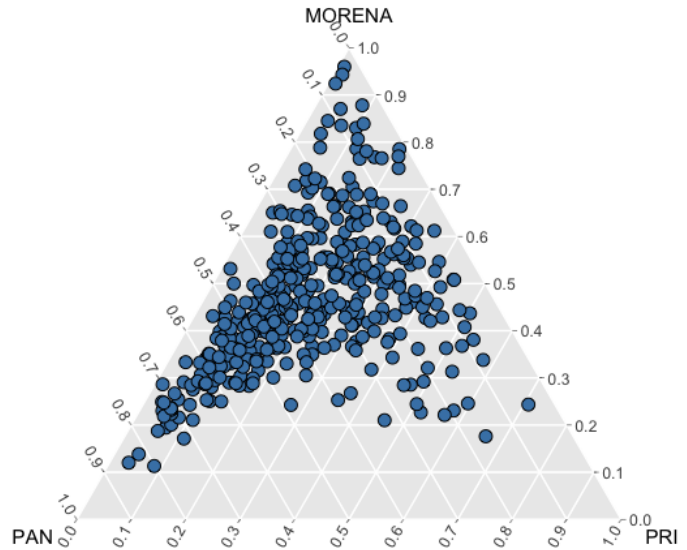
- Composición Geoquímica de Rocas
- Composición Geoquímica de Sedimentos
- Composición Proteica de la Leche
- Procesos Electorales

Ejemplos

- Composición Geoquímica de Rocas



- Datos Electorales (Puebla 2019)



- Geología, Política, Economía, Medicina, etc.

Su naturaleza y problemática

El símplex como espacio muestral

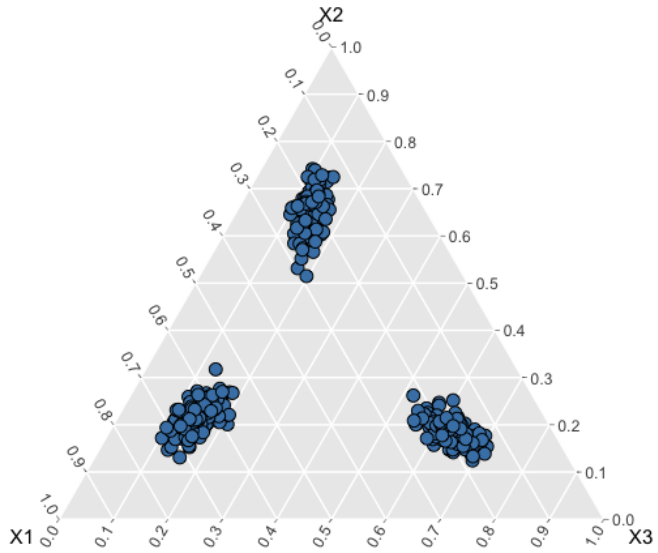
$$\mathbb{S}^p = \{\boldsymbol{x} \in \mathbb{R}^p : x_i \geq 0, \sum_{i=1}^{p} x_i = 1\}$$

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

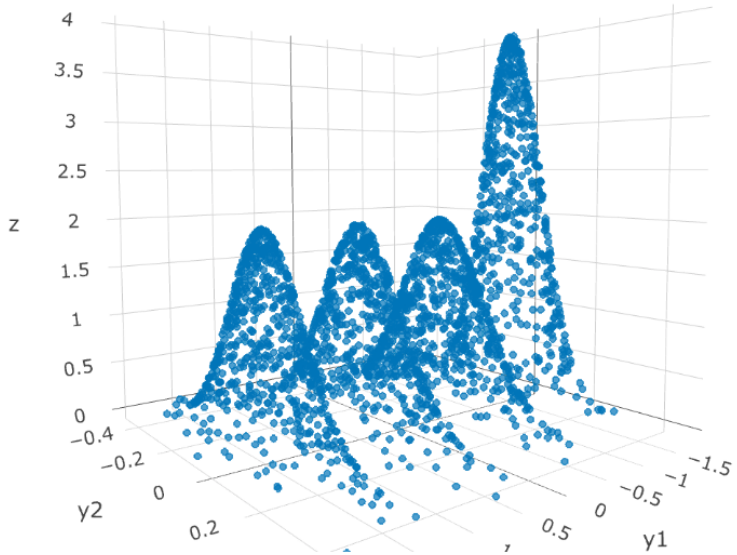
- Aitchison (1982) estudia las características y operaciones de los datos composicionales.

Objetivo específico

- Patrones más complejos

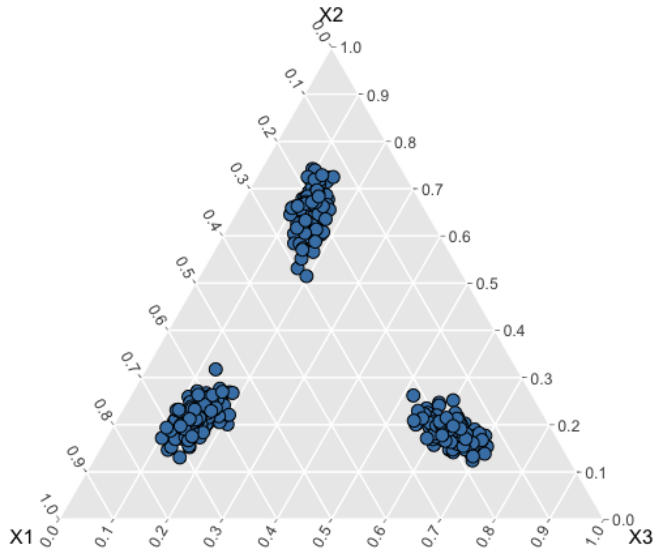


- Mezcla de Distribuciones Normales Multivariadas

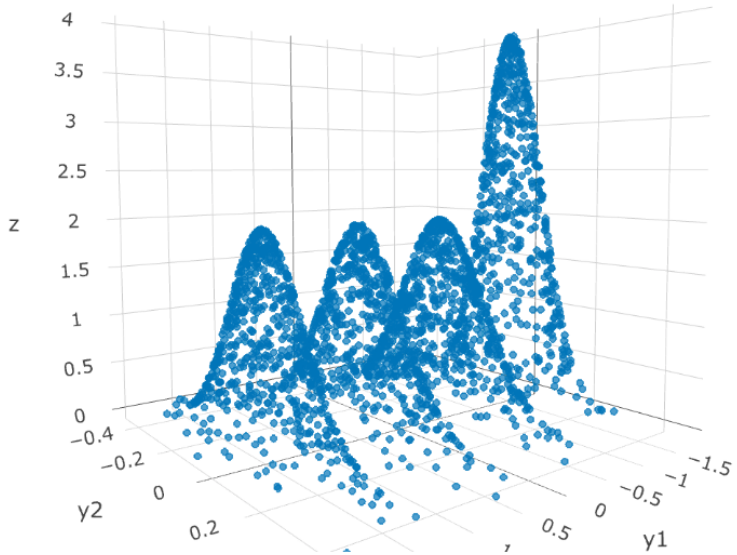


Objetivo específico

- Patrones más complejos



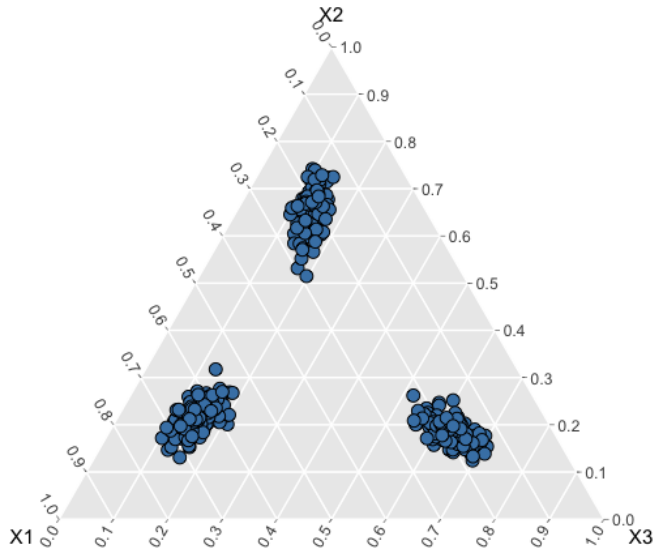
- Mezcla de Distribuciones Normales Multivariadas



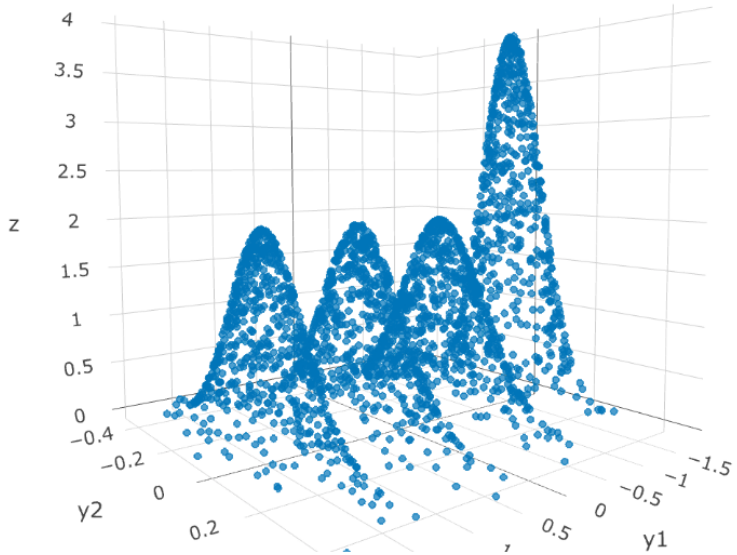
- Mediante la conocida transformación log-cociente ilr, definida por Egozcue et al. (2003).

Objetivo específico

- Patrones más complejos



- Mezcla de Distribuciones Normales Multivariadas



- Mediante la conocida transformación log-cociente ilr, definida por Egozcue et al. (2003).
- Los modelos de mezclas ofrecen una gran versatilidad en la descripción de datos, su flexibilidad permite aproximar problemas desde un enfoque no paramétrico, son ampliamente utilizados en Estadística.

Un modelo no-paramétrico

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

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

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

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

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):

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

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):

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

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.

Metodología

- Inferencia desde el enfoque Bayesiano:
  - Procesos de Dirichlet.
  - Modelo de Mezclas de Procesos Dirichlet.

Metodología

- Inferencia desde el enfoque Bayesiano:
  - Procesos de Dirichlet.
  - Modelo de Mezclas de Procesos Dirichlet.
- 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.

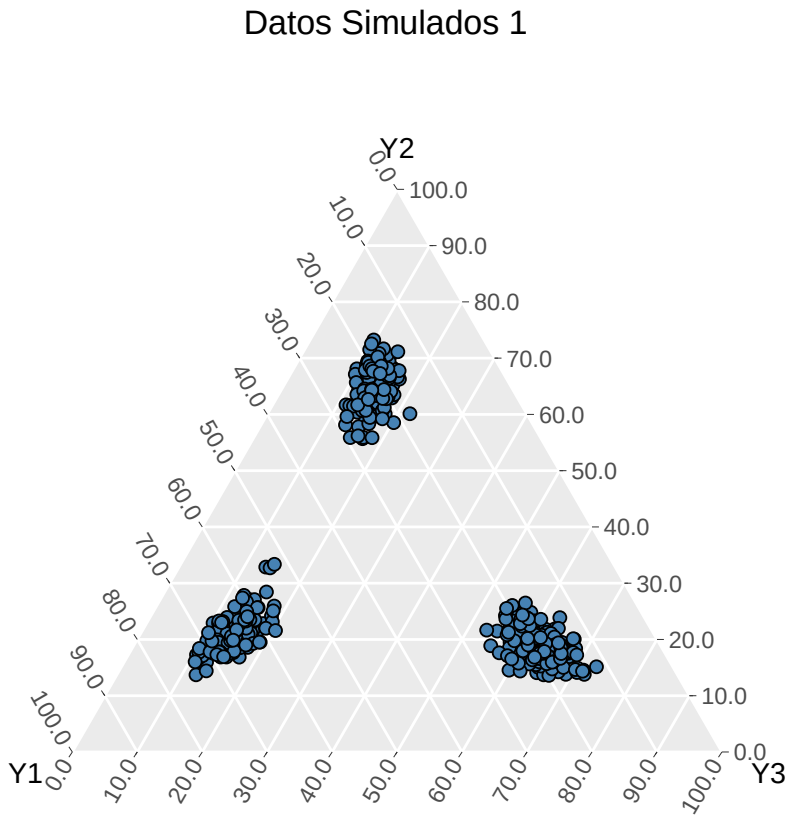
Metodología

- Inferencia desde el enfoque Bayesiano:
  - Procesos de Dirichlet.
  - Modelo de Mezclas de Procesos Dirichlet.
- 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 $\mathbf{R}$.

Metodología

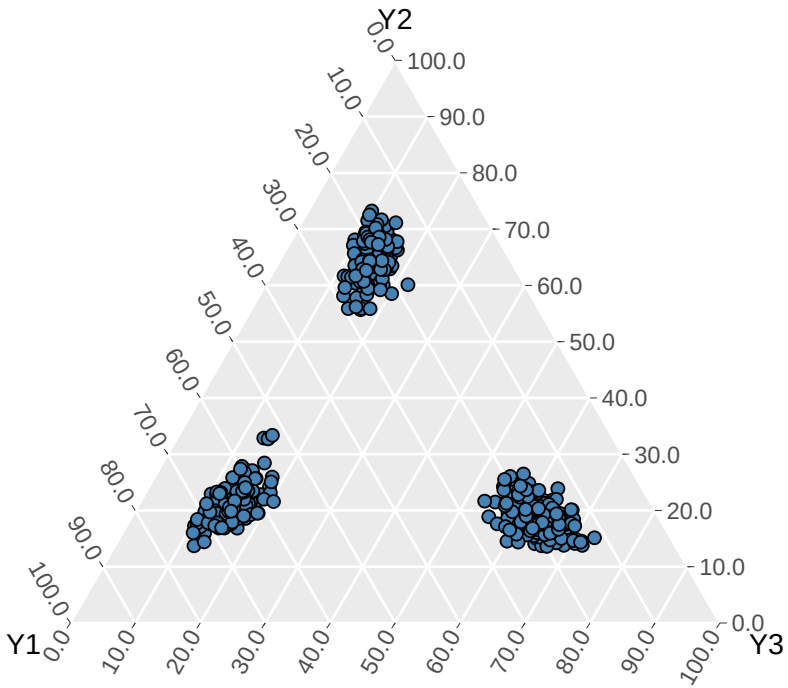
- Inferencia desde el enfoque Bayesiano:
  - Procesos de Dirichlet.
  - Modelo de Mezclas de Procesos Dirichlet.
- Se utilizó el algoritmo dado por Walker (2007):
  - Gibbs Sampling.
  - Slice Sampling.
  - Variables latentes.
- Se complemento con la propuesta realizada por Kalli et al. (2011):
  - Slice 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



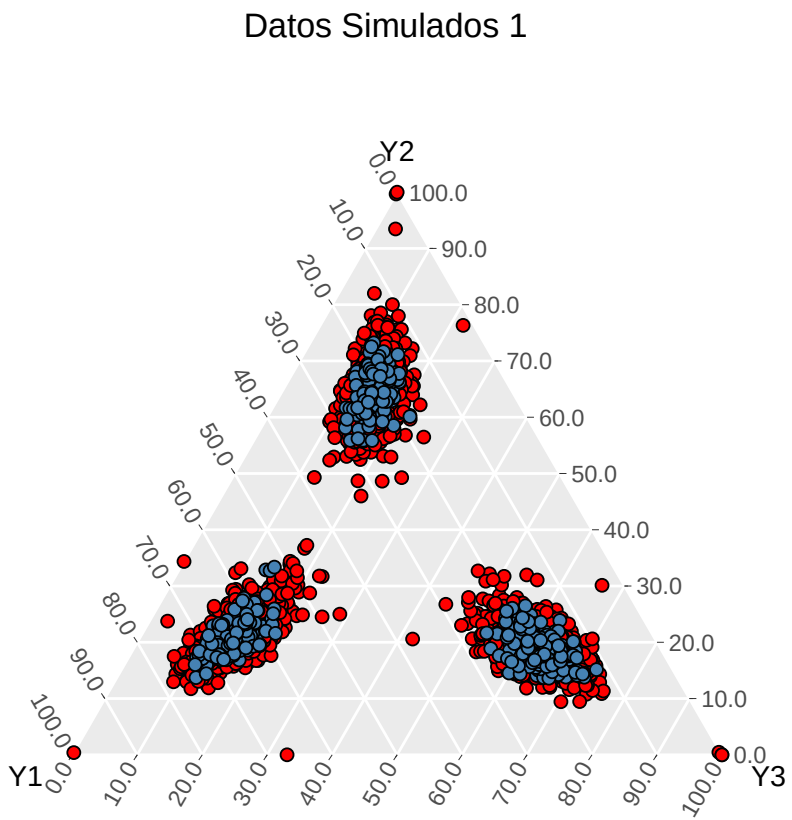
Datos Simulados

Datos Simulados 1



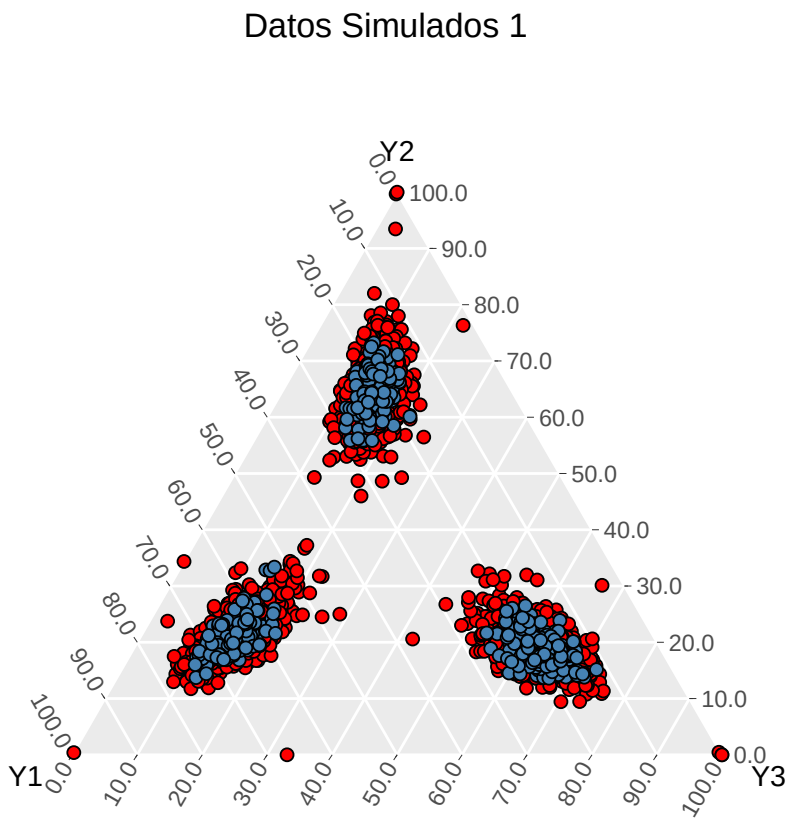
Intervalos Poblacionales Marginales		
Componente	Inferior	Superior
Y1	14.68 %	70.38 %
Y2	14.73 %	70.39 %
Y3	10.51 %	68.41 %

Datos Simulados



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

Datos Simulados

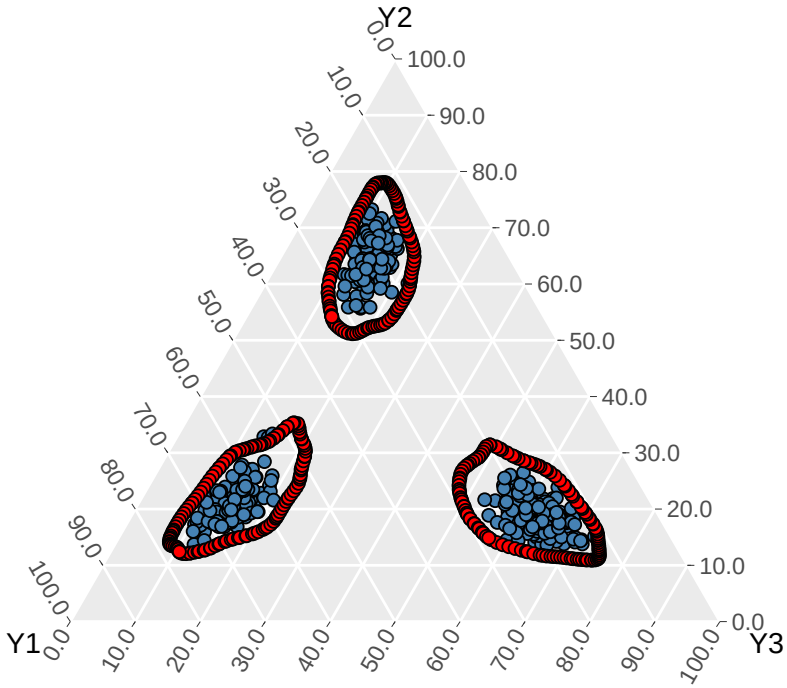


Intervalos Poblacionales Marginales		
Componente	Inferior	Superior
Y1	14.68 %	70.38 %
Y2	14.73 %	70.39 %
Y3	10.51 %	68.41 %

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

Datos Simulados

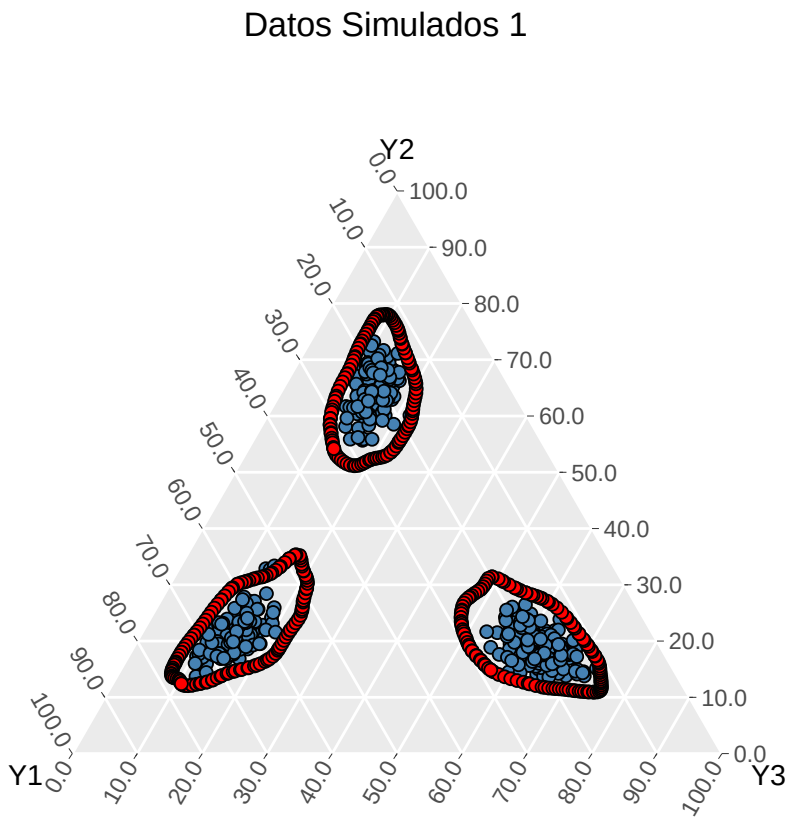
Datos Simulados 1



Intervalos Poblacionales Marginales		
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 %

Datos Simulados



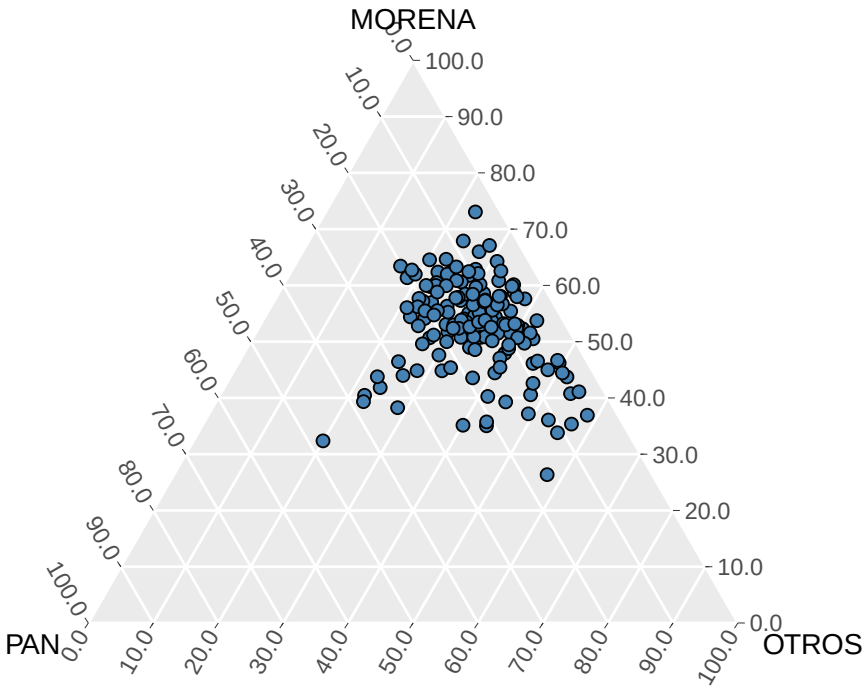
Intervalos Poblacionales Marginales		
Componente	Inferior	Superior
Y1	14.68 %	70.38 %
Y2	14.73 %	70.39 %
Y3	10.51 %	68.41 %

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.

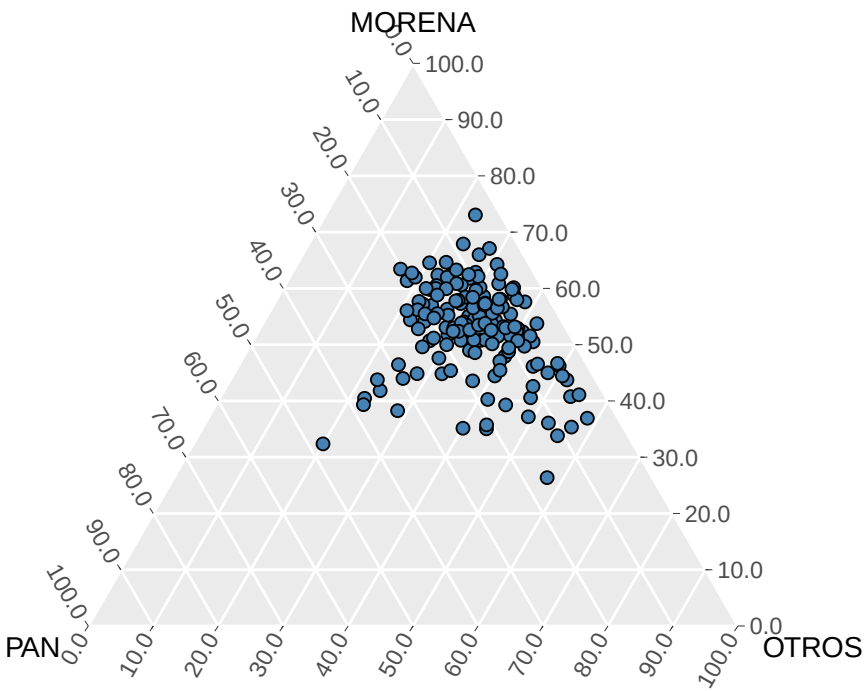
Datos Reales

Gubernatura del estado de Morelos 2018



Datos Reales

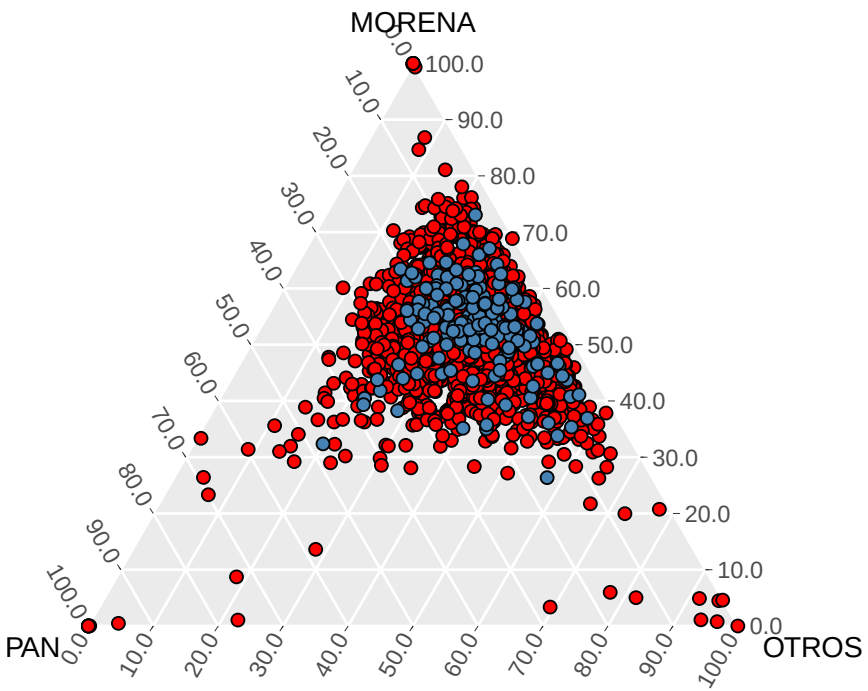
Gubernatura del estado de Morelos 2018



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

Datos Reales

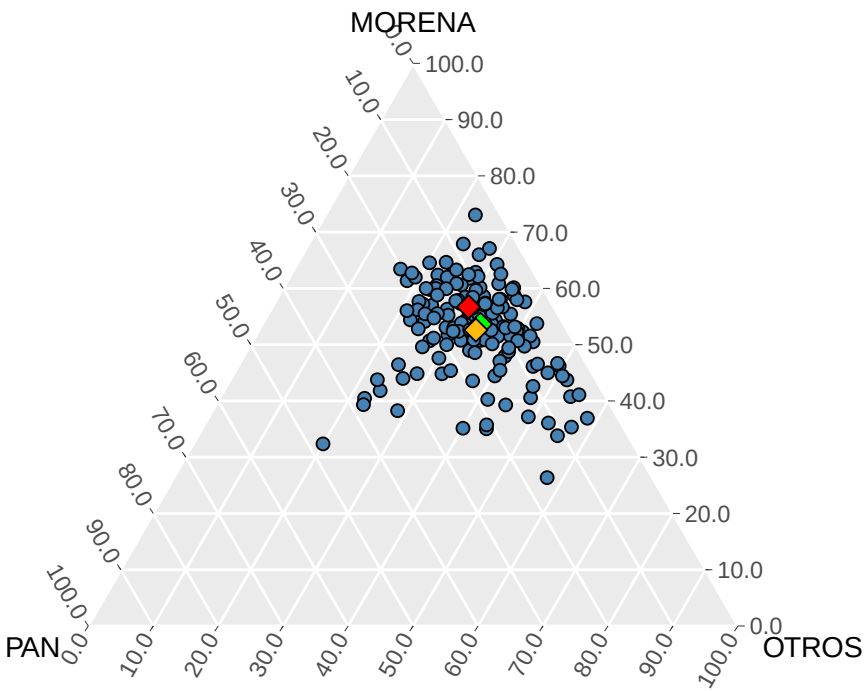
Gubernatura del estado de Morelos 2018



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

Datos Reales

Gubernatura del estado de Morelos 2018

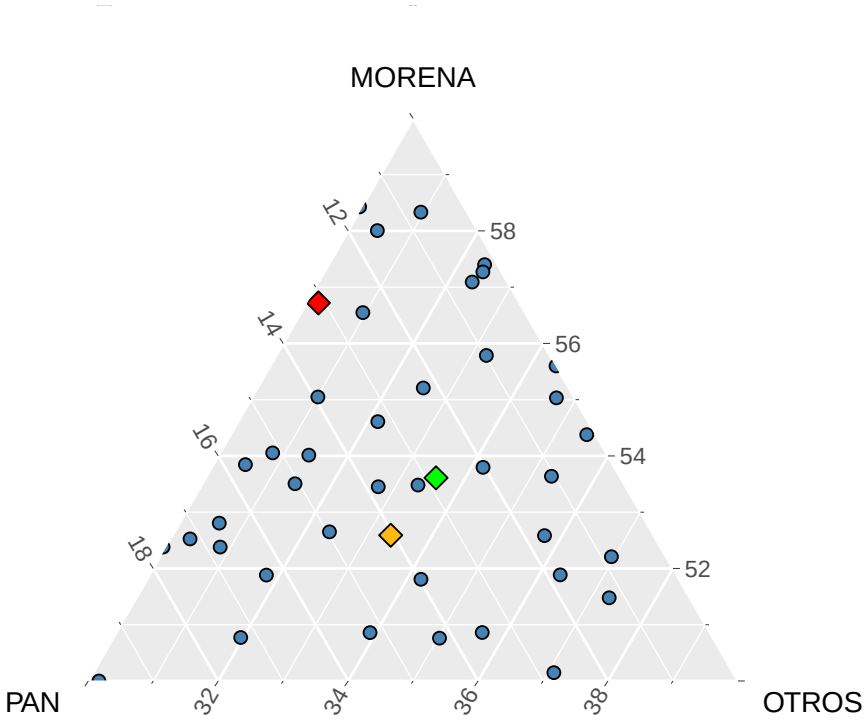


Cómputo Distrital ♦		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

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

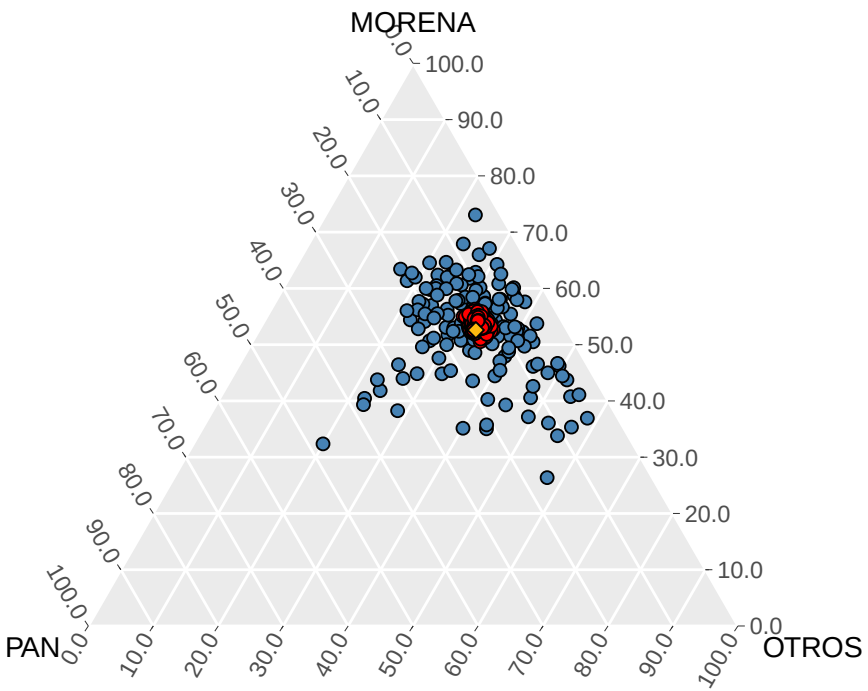


Cómputo Distrital ♦		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

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

Datos Reales

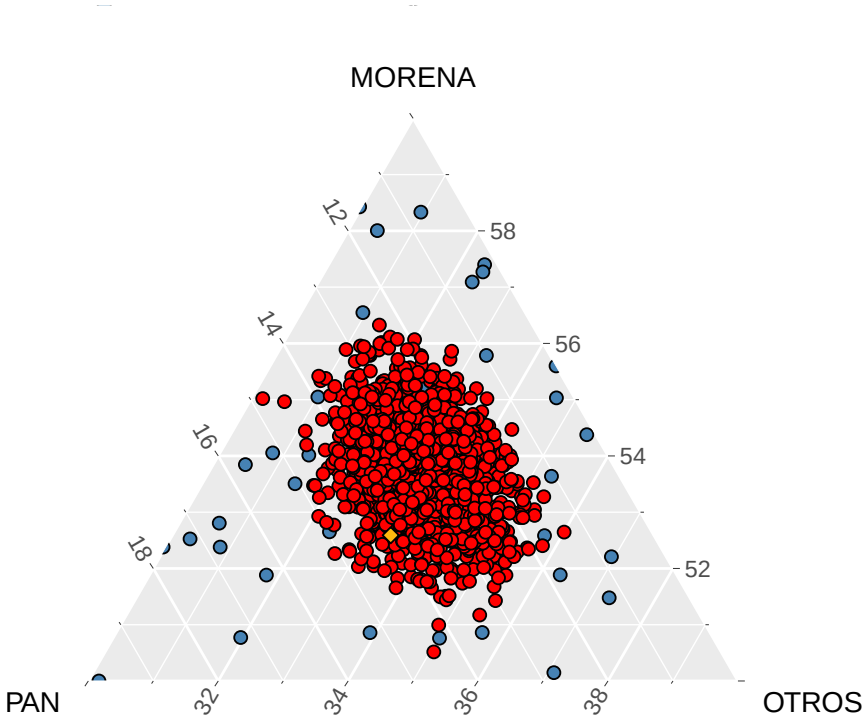
Gubernatura del estado de Morelos 2018



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

Datos Reales

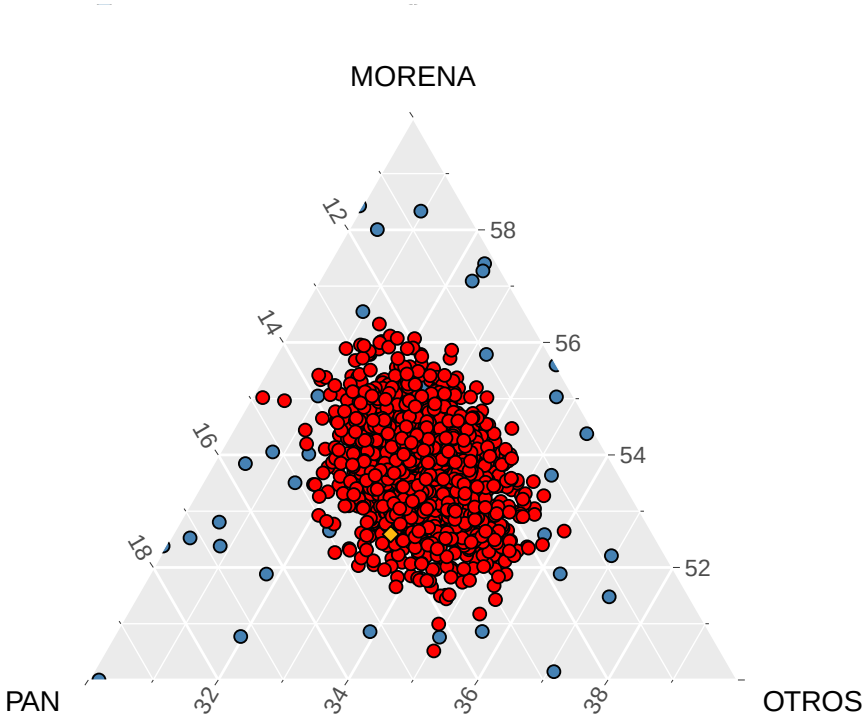
Gubernatura del estado de Morelos 2018



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

Datos Reales

Gubernatura del estado de Morelos 2018

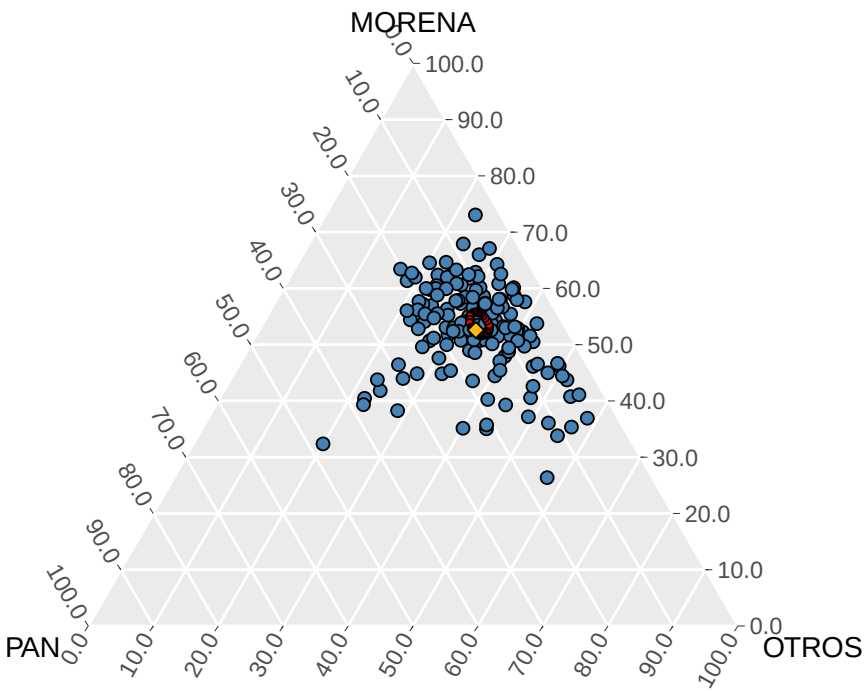


Cómputo Distrital ♦		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

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

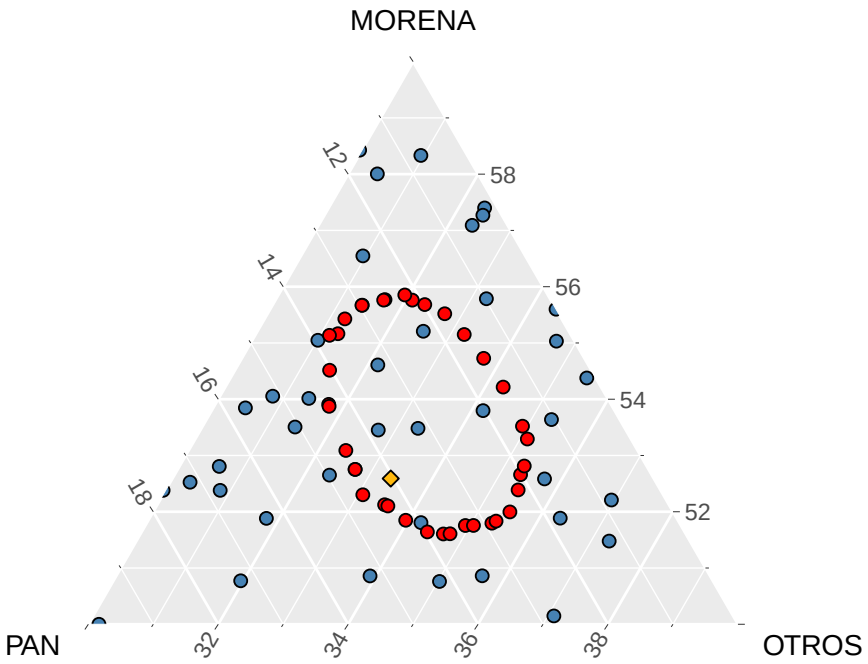


Cómputo Distrital ♦		
Coalición		
MORENA	PAN	OTROS
52.59 %	14.05 %	33.36 %

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

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	Tamaño
MORENA	52.20 %	55.31 %	3.11 %
PAN	11.80 %	14.22 %	2.43 %
OTROS	31.70 %	34.96 %	3.26 %

Intervalos del INE			
Coalición	Inferior	Superior	Tamaño
MORENA	51.00 %	53.80 %	2.8 %
PAN	13.40 %	16.10 %	2.7 %

Comparando con los resultados del INE nuestros intervalos exceden en 0.31 % a la componente **MORENA** y disminuye en 0.27 % a la componente **PAN**.

Comparación entre Modelos

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

Modelo de Mezclas de Normales Medianas

Intervalos de Probabilidad Marginales			
Coalición	Inferior	Superior	Tamaño
MORENA	52.20 %	55.31 %	3.11 %
PAN	11.80 %	14.22 %	2.43 %
OTROS	31.70 %	34.96 %	3.26 %

Modelo de Mezclas de Gammas Medias

Intervalos de Probabilidad Marginales			
Coalición	Inferior	Superior	Tamaño
MORENA	51.99 %	54.57 %	2.58 %
PAN	12.25 %	14.46 %	2.21 %
OTROS	31.88 %	34.91 %	3.03 %

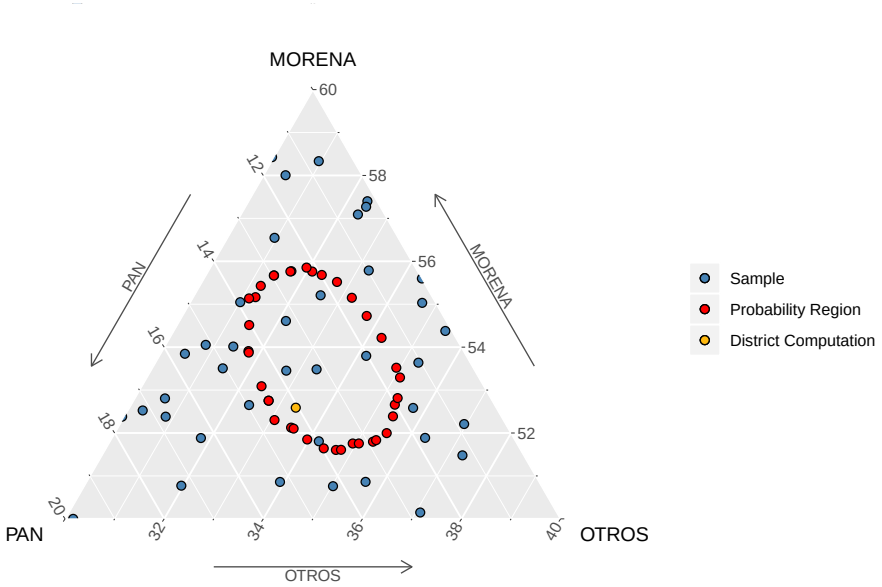
Intervalos del INE			
Coalición	Inferior	Superior	Tamaño
MORENA	51.00 %	53.80 %	2.8 %
PAN	13.40 %	16.10 %	2.7 %

Notas: Comparando con los resultados del INE el Modelo de Mezclas de Normales con la **Mediana** excede en 0.31 % a la componente **MORENA** y disminuye en 0.27 % a la componente **PAN**.
Comparando con los resultados del INE el Modelo de Mezclas de Gammas con la **Media** disminuye en 0.22 % a la componente **MORENA** y disminuye en 0.49 % a la componente **PAN**.
Por lo tanto, para el estado de**Morelos** aparentemente es más adecuado el Modelo de Mezclas de Gammas.

Comparación entre Modelos

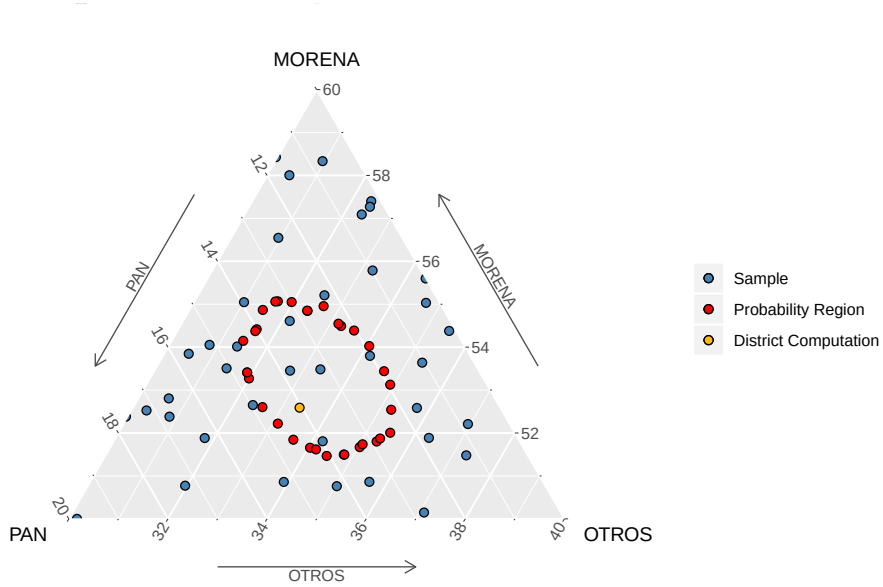
Modelo de Mezclas de Normales Medianas

Gubernatura Morelos 2018



Modelo de Mezclas de Gammas Medias

Gubernatura Morelos 2018



Temas abarcados

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

Conclusiones

- 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

- 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.

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

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

Mi correo:
danielfournier270790@gmail.com

¡Gracias!